openstax™

Statistics

HIGH SCHOOL

Statistics

SENIOR CONTRIBUTING AUTHORS

Barbara Illowsky, De Anza College

Susan Dean, De Anza College

openstax™

OpenStax
Rice University
6100 Main Street MS-375
Houston, Texas 77005

To learn more about OpenStax, visit https://openstax.org.
Individual print copies and bulk orders can be purchased through our website.

©2020 Texas Education Agency (TEA). This instructional material was initially created through a Texas Education Agency (TEA) initiative to provide high-quality open-source instructional materials to districts free of charge. Changes were made to the original material, including updates to art, structure, and other content updates. The original material is available at: https://www.texasgateway.org/book/tea-statistics. Textbook content is licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/) (CC BY 4.0).

For questions regarding this licensing, please contact support@openstax.org.

Trademarks
The OpenStax name, OpenStax logo, OpenStax book covers, OpenStax CNX name, OpenStax CNX logo, OpenStax Tutor name, Openstax Tutor logo, Connexions name, Connexions logo, Rice University name, and Rice University logo are not subject to the license and may not be reproduced without the prior and express written consent of Rice University.

HARDCOVER BOOK ISBN-13	**978-1-975076-53-5**
B&W PAPERBACK BOOK ISBN-13	**978-1-975076-52-8**
DIGITAL VERSION ISBN-13	**978-1-951693-22-0**
ORIGINAL PUBLICATION YEAR	**2020**

10 9 8 7 6 5 4 3 2 1

OpenStax

OpenStax provides free, peer-reviewed, openly licensed textbooks for introductory college and Advanced Placement® courses and low-cost, personalized courseware that helps students learn. A nonprofit ed tech initiative based at Rice University, we're committed to helping students access the tools they need to complete their courses and meet their educational goals.

Rice University

OpenStax, OpenStax CNX, and OpenStax Tutor are initiatives of Rice University. As a leading research university with a distinctive commitment to undergraduate education, Rice University aspires to path-breaking research, unsurpassed teaching, and contributions to the betterment of our world. It seeks to fulfill this mission by cultivating a diverse community of learning and discovery that produces leaders across the spectrum of human endeavor.

Philanthropic Support

OpenStax is grateful for our generous philanthropic partners, who support our vision to improve educational opportunities for all learners.

Laura and John Arnold Foundation

Arthur and Carlyse Ciocca Charitable Foundation

Ann and John Doerr

Bill & Melinda Gates Foundation

Girard Foundation

Google Inc.

The William and Flora Hewlett Foundation

Rusty and John Jaggers

The Calvin K. Kazanjian Economics Foundation

Charles Koch Foundation

Leon Lowenstein Foundation, Inc.

The Maxfield Foundation

Burt and Deedee McMurtry

Michelson 20MM Foundation

National Science Foundation

The Open Society Foundations

Jumee Yhu and David E. Park III

Brian D. Patterson USA-International Foundation

The Bill and Stephanie Sick Fund

Robin and Sandy Stuart Foundation

The Stuart Family Foundation

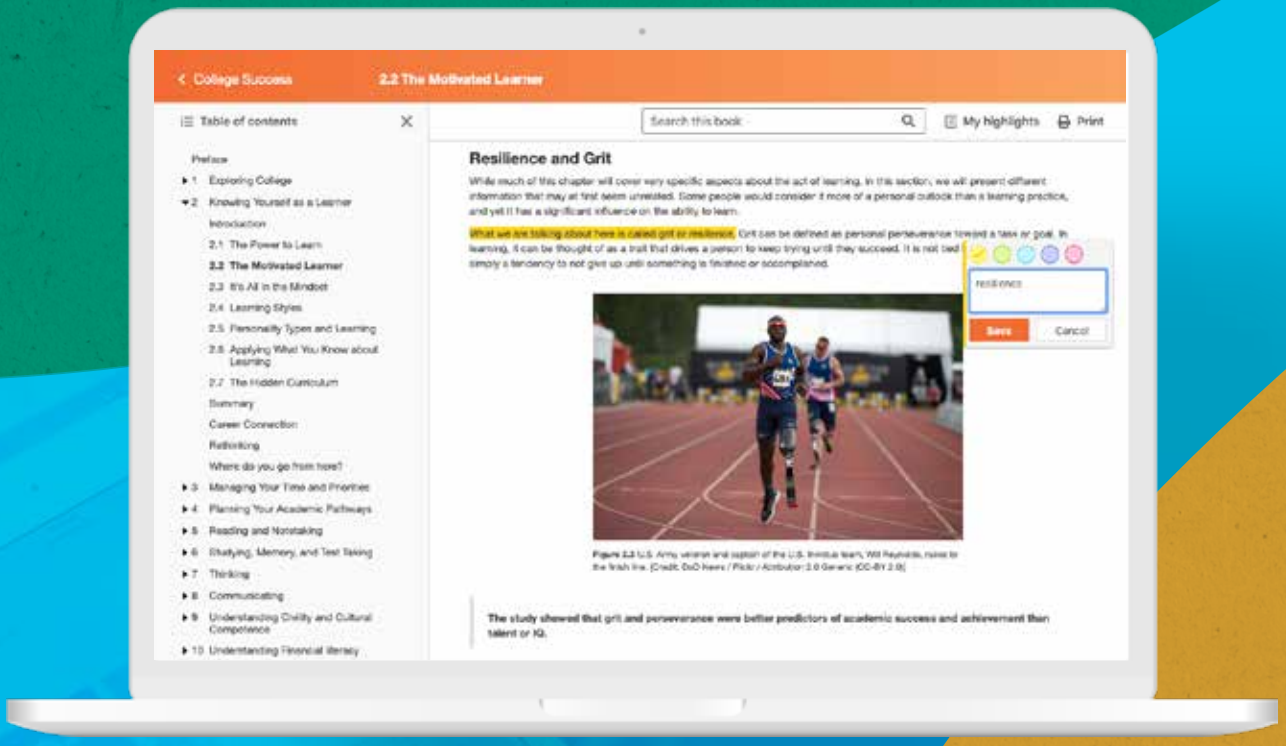
Tammy and Guillermo Treviño

openstax™

Study where you want, what you want, when you want.

When you access your book in our web view, you can use our new online **highlighting and note-taking** features to create your own study guides.

Our books are free and flexible, forever.
Get started at **openstax.org/details/books/statistics**



Access. The future of education.
openstax.org

RICE

Contents

Preface 1

1 Sampling and Data 5

Introduction 5
1.1 Definitions of Statistics, Probability, and Key Terms 5
1.2 Data, Sampling, and Variation in Data and Sampling 10
1.3 Frequency, Frequency Tables, and Levels of Measurement 27
1.4 Experimental Design and Ethics 35
1.5 Data Collection Experiment 39
1.6 Sampling Experiment 41
Key Terms 44
Chapter Review 45
Practice 46
Homework 52
Bringing It Together: Homework 59
References 60
Solutions 61

2 Descriptive Statistics 67

Introduction 67
2.1 Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs 68
2.2 Histograms, Frequency Polygons, and Time Series Graphs 77
2.3 Measures of the Location of the Data 87
2.4 Box Plots 95
2.5 Measures of the Center of the Data 100
2.6 Skewness and the Mean, Median, and Mode 105
2.7 Measures of the Spread of the Data 108
2.8 Descriptive Statistics 120
Key Terms 122
Chapter Review 122
Formula Review 123
Practice 124
Homework 138
Bringing It Together: Homework 150
References 156
Solutions 157

3 Probability Topics 169

Introduction 169
3.1 Terminology 170
3.2 Independent and Mutually Exclusive Events 175
3.3 Two Basic Rules of Probability 183
3.4 Contingency Tables 188
3.5 Tree and Venn Diagrams 194
3.6 Probability Topics 205
Key Terms 208

Chapter Review 208
Formula Review 209
Practice 209
Bringing It Together: Practice 213
Homework 214
Bringing It Together: Homework 222
References 225
Solutions 226

4 Discrete Random Variables 233

Introduction 233
4.1 Probability Distribution Function (PDF) for a Discrete Random Variable 234
4.2 Mean or Expected Value and Standard Deviation 237
4.3 Binomial Distribution (Optional) 243
4.4 Geometric Distribution (Optional) 250
4.5 Hypergeometric Distribution (Optional) 254
4.6 Poisson Distribution (Optional) 257
4.7 Discrete Distribution (Playing Card Experiment) 262
4.8 Discrete Distribution (Lucky Dice Experiment) 266
Key Terms 270
Chapter Review 271
Formula Review 272
Practice 273
Homework 279
References 291
Solutions 292

5 Continuous Random Variables 301

Introduction 301
5.1 Continuous Probability Functions 303
5.2 The Uniform Distribution 306
5.3 The Exponential Distribution (Optional) 314
5.4 Continuous Distribution 325
Key Terms 328
Chapter Review 328
Formula Review 329
Practice 330
Homework 337
References 342
Solutions 342

6 The Normal Distribution 349

Introduction 349
6.1 The Standard Normal Distribution 350
6.2 Using the Normal Distribution 355
6.3 Normal Distribution—Lap Times 363
6.4 Normal Distribution—Pinkie Length 365

Key Terms 367
Chapter Review 367
Formula Review 367
Practice 367
Homework 371
References 376
Solutions 377

7 The Central Limit Theorem 381

Introduction 381
7.1 The Central Limit Theorem for Sample Means (Averages) 382
7.2 The Central Limit Theorem for Sums (Optional) 387
7.3 Using the Central Limit Theorem 390
7.4 Central Limit Theorem (Pocket Change) 398
7.5 Central Limit Theorem (Cookie Recipes) 400
Key Terms 404
Chapter Review 404
Formula Review 405
Practice 405
Homework 408
References 417
Solutions 417

8 Confidence Intervals 423

Introduction 423
8.1 A Single Population Mean Using the Normal Distribution 425
8.2 A Single Population Mean Using the Student's t-Distribution 435
8.3 A Population Proportion 439
8.4 Confidence Interval (Home Costs) 445
8.5 Confidence Interval (Place of Birth) 447
8.6 Confidence Interval (Women's Heights) 448
Key Terms 451
Chapter Review 451
Formula Review 452
Practice 453
Homework 459
References 468
Solutions 469

9 Hypothesis Testing with One Sample 481

Introduction 481
9.1 Null and Alternative Hypotheses 482
9.2 Outcomes and the Type I and Type II Errors 484
9.3 Distribution Needed for Hypothesis Testing 486
9.4 Rare Events, the Sample, and the Decision and Conclusion 487
9.5 Additional Information and Full Hypothesis Test Examples 489
9.6 Hypothesis Testing of a Single Mean and Single Proportion 504

Key Terms 507
Chapter Review 507
Formula Review 508
Practice 509
Homework 512
References 522
Solutions 523

10 Hypothesis Testing with Two Samples 533

Introduction 533
10.1 Two Population Means with Unknown Standard Deviations 534
10.2 Two Population Means with Known Standard Deviations 541
10.3 Comparing Two Independent Population Proportions 544
10.4 Matched or Paired Samples (Optional) 549
10.5 Hypothesis Testing for Two Means and Two Proportions 554
Key Terms 557
Chapter Review 557
Formula Review 557
Practice 558
Homework 563
Bringing It Together: Homework 573
References 575
Solutions 576

11 The Chi-Square Distribution 585

Introduction 585
11.1 Facts About the Chi-Square Distribution 586
11.2 Goodness-of-Fit Test 587
11.3 Test of Independence 596
11.4 Test for Homogeneity 600
11.5 Comparison of the Chi-Square Tests 605
11.6 Test of a Single Variance 605
11.7 Lab 1: Chi-Square Goodness-of-Fit 607
11.8 Lab 2: Chi-Square Test of Independence 610
Key Terms 612
Chapter Review 612
Formula Review 612
Practice 613
Homework 619
Bringing It Together: Homework 632
References 632
Solutions 632

12 Linear Regression and Correlation 641

Introduction 641
12.1 Linear Equations 641
12.2 The Regression Equation 644

12.3 Testing the Significance of the Correlation Coefficient (Optional) 653
12.4 Prediction (Optional) 657
12.5 Outliers 658
12.6 Regression (Distance from School) (Optional) 667
12.7 Regression (Textbook Cost) (Optional) 669
12.8 Regression (Fuel Efficiency) (Optional) 670
Key Terms 673
Chapter Review 673
Formula Review 674
Practice 674
Homework 680
Bringing It Together: Homework 689
References 694
Solutions 694

13 **F Distribution and One-way Anova 703**

Introduction 703
13.1 One-Way ANOVA 704
13.2 The F Distribution and the F Ratio 704
13.3 Facts About the F Distribution 709
13.4 Test of Two Variances 716
13.5 Lab: One-Way ANOVA 718
Key Terms 720
Chapter Review 720
Formula Review 721
Practice 721
Homework 725
References 737
Solutions 737

A **Appendix A Review Exercises (Ch 3–13) 743**

B **Appendix B Practice Tests (1–4) and Final Exams 769**

C **Data Sets 823**

D **Group and Partner Projects 827**

E **Solution Sheets 833**

F **Mathematical Phrases, Symbols, and Formulas 837**

G **Notes for the TI-83, 83+, 84, 84+ Calculators 845**

H Tables 855

Index 857

Preface

Welcome to *Statistics*, an OpenStax resource. This textbook was written to increase teacher and student access to high-quality learning materials, maintaining the highest standards of academic rigor at little to no cost.

About OpenStax

OpenStax is a nonprofit based at Rice University, and it's our mission to improve student access to education. Our first openly licensed college textbook was published in 2012, and our library has since scaled to over 35 books used by hundreds of thousands of students for college and AP® courses. OpenStax Tutor and Rover, our low-cost personalized learning tools, are being used in college and high school courses throughout the country. Through our partnerships with philanthropic foundations and our alliance with other educational resource organizations, OpenStax is breaking down the most common barriers to learning and empowering students and instructors to succeed.

About OpenStax Resources

Customization

Statistics is licensed under a Creative Commons Attribution 4.0 International (CC BY) license, which means that you can distribute, remix, and build upon the content, as long as you provide attribution to OpenStax and its content contributors.

Because our books are openly licensed, you are free to use the entire book or pick and choose the sections that are most relevant to the needs of your students. Feel free to remix the content by assigning your students certain chapters and sections in your syllabus, in the order that you prefer. You can even provide a direct link in your syllabus or student assignment system to the sections in the web view of your book.

Instructors also have the option of creating a customized version of their OpenStax book. The custom version can be made available to students in low-cost print or digital form through their campus bookstore. Visit the Instructor Resources section of your book page on openstax.org for more information.

Art Attribution in *Statistics*

In *Statistics*, most art contains attribution to its title, creator or rights holder, host platform, and license within the caption. For art that is openly licensed, anyone may reuse the art as long as they provide the same attribution to its original source. Some art has been provided through permissions and should only be used with the attribution or limitations provided in the credit.

Errata

All OpenStax textbooks undergo a rigorous review process. However, like any professional-grade textbook, errors sometimes occur. The good part is, since our books are web-based, we can make updates periodically. If you have a correction to suggest, submit it through our errata reporting tool. We will review your suggestion and make necessary changes.

Format

You can access this textbook for free in web view or PDF through openstax.org, and for a low cost in print.

About *Statistics*

This instructional material was initially created through a Texas Education Agency (TEA) initiative to provide high-quality open-source instructional materials to districts free of charge. Funds were allocated by the 84th Texas Legislature (2015) for the creation of state-developed, open-source instructional materials with the request that advanced secondary courses supporting the study of science, technology, engineering, and mathematics should be prioritized.

Statistics covers the scope and sequence requirements of a typical one-year statistics course. The text provides comprehensive coverage of statistical concepts, including quantitative examples, collaborative activities, and practical applications. *Statistics* was designed to meet and exceed the requirements of the relevant Texas Essential Knowledge and Skills (TEKS) (http://ritter.tea.state.tx.us/rules/tac/chapter111/ch111c.html#111.47), while allowing significant flexibility for instructors.

Qualified and experienced Texas faculty were involved throughout the development process, and the textbooks were reviewed extensively to ensure effectiveness and usability in each course. Reviewers considered each resource's clarity, accuracy, student support, assessment rigor and appropriateness, alignment to TEKS, and overall quality. Their invaluable suggestions provided the basis for continually improved material and helped to certify that the books are ready for use. The writers and reviewers also considered common course issues, effective teaching strategies, and student engagement to provide instructors and students with useful, supportive content and drive effective learning experiences.

Coverage and Scope

Statistics presents the appropriate statistical concepts and skills in a logical and engaging progression that should be familiar to faculty.

Chapter 1: Sampling and Data
Chapter 2: Descriptive Statistics
Chapter 3: Probability Topics
Chapter 4: Discrete Random Variables
Chapter 5: Continuous Random Variables
Chapter 6: The Normal Distribution
Chapter 7: The Central Limit Theorem
Chapter 8: Confidence Intervals
Chapter 9: Hypothesis Testing with One Sample
Chapter 10: Hypothesis Testing with Two Samples
Chapter 11: The Chi-Square Distribution
Chapter 12: Linear Regression and Correlation
Chapter 13: F Distribution and One-Way ANOVA

Flexibility

Like any OpenStax content, this textbook can be modified as needed for use by the instructor depending on the needs of the students in the course. Each set of materials created by OpenStax is organized into units and chapters and can be used like a traditional textbook as the entire syllabus for each course. The materials can also be accessed in smaller chunks for more focused use with a single student or an entire class. Instructors are welcome to download and assign the PDF version of the textbook through a learning management system or can use their LMS to link students to specific chapters and sections of the book relevant to the concept being studied. The entire textbook will be available during the fall of 2020 in an editable Google document, and until then instructors are welcome to copy and paste content from the textbook to modify as needed prior to instruction.

Student-Centered Focus

Statistics was developed with detailed and practical guidance from experienced high school teachers and curriculum experts. Their contributions helped create a resource that provides easy-to-follow explanations with ample opportunities for enrichment and practice. In addition to clear and grade-level appropriate main text coverage, the following features are meant to enhance student understanding of statistics concepts:

- *Examples* are placed strategically throughout the text to show students the step-by-step process of interpreting and solving statistical problems. To keep the text relevant for students, the examples are drawn from a broad spectrum of practical topics, including examples from academic life and learning, health and medicine, retail and business, and sports and entertainment.
- *Try It* practice problems immediately follow many examples and give students the opportunity to practice as they read the text. Like the Examples, the Try It problems are usually based on practical and familiar topics.
- *Collaborative Exercises* provide an in-class scenario for students to work together and learn from each other as they explore course concepts.
- *Calculator Guidance* shows students step-by-step instructions for input using the TI-83, 83+, 84, and 84+ calculators and helps them consider how to use these tools in their studies. The Technology Icon indicates where the use of a TI calculator or computer software is recommended.
- *Practice, Homework, and Bringing It Together* problems give the students problems at various degrees of difficulty while including real-world scenarios to engage students.

Statistics Labs

These innovative activities were developed by Barbara Illowsky and Susan Dean (both of De Anza College) and allow students to design, implement, and interpret statistical analyses. They are drawn from actual experiments and data-gathering processes and offer a unique hands-on and collaborative experience. Statistics Labs appear at the end of each chapter and begin with student learning outcomes, general estimates for time on task, and global implementation notes. Students are then provided with step-by-step guidance, including sample data tables and calculation prompts. This detailed assistance will help the students successfully apply statistics concepts and lay the groundwork for future collaborative or individual work.

Answers to Questions in the Book

Answers to Examples are provided directly below the question. Try It solutions are not provided. Students can find answers to odd-numbered Practice, Homework, and Bringing it Together questions at the end of each chapter, and all answers to Review Problems and Practice Test questions in Appendix A and B. Answers to even-numbered Practice,

Homework, and Bringing it Together questions are provided only to instructors in the Instructor Answer Guide via the Instructor Resources page.

Additional Resources

Student and Instructor Resources

We've compiled additional resources for both students and instructors, including Getting Started Guides, PowerPoint slides, and an instructor answer guide. Instructor resources require a verified instructor account, which you can apply for when you log in or create your account on OpenStax.org. Take advantage of these resources to supplement your OpenStax book.

Partner Resources

OpenStax Partners are our allies in the mission to make high-quality learning materials affordable and accessible to students and instructors everywhere. Their tools integrate seamlessly with our OpenStax titles at a low cost. To access the partner resources for your text, visit your book page on OpenStax.org.

About the Authors

Senior Contributing Authors

Barbara Illowsky, De Anza College
Susan Dean, De Anza College

Contributing Authors

Daniel Birmajer, Nazareth College
Bryan Blount, Kentucky Wesleyan College
Sheri Boyd, Rollins College
Matthew Einsohn, Prescott College
James Helmreich, Marist College
Lynette Kenyon, Collin County Community College
Sheldon Lee, Viterbo University
Jeff Taub, Maine Maritime Academy

Reviewers of Prior Editions

Laurel Chiappetta, University of Pittsburgh
Lenore Desilets, De Anza College
Lisa Markus, De Anza College
Mary Teegarden, San Diego Mesa College
Carol Olmstead, De Anza College
Carol Weideman, St. Petersburg College
Charles Ashbacher, Upper Iowa University, Cedar Rapids
Charles Klein, De Anza College
Cheryl Wartman, University of Prince Edward Island
David French, Tidewater Community College
Dennis Walsh, Middle Tennessee State University
Diane Mathios, De Anza College
John Thomas, College of Lake County
Jing Chang, College of Saint Mary
Sara Lenhart, Christopher Newport University
Sarah Boslaugh, Kennesaw State University
Abdulhamid Sukar, Cameron University
Abraham Biggs, Broward Community College
Adam Pennell, Greensboro College
Alexander Kolovos
Ann Flanigan, Kapiolani Community College
Robert McDevitt, Germanna Community College
Roberta Bloom, De Anza College
Rupinder Sekhon, De Anza College
Sudipta Roy, Kankakee Community College
Cindy Moss, Skyline College
Ernest Bonat, Portland Community College
Kathy Plum, De Anza College

Andrew Wiesner, Pennsylvania State University
Jonathan Oaks, Macomb Community College
Michael Greenwich, College of Southern Nevada
Miriam Masullo, SUNY Purchase
Mo Geraghty, De Anza College
Larry Green, Lake Tahoe Community College
Nydia Nelson, St. Petersburg College
Philip J. Verrecchia, York College of Pennsylvania
Robert Henderson, Stephen F. Austin State University
Benjamin Ngwudike, Jackson State University
Mel Jacobsen, Snow College
Birgit Aquilonius, West Valley College
Jim Lucas, De Anza College
David Bosworth, Hutchinson Community College
Frank Snow, De Anza College
George Bratton, University of Central Arkansas
Inna Grushko, De Anza College
Janice Hector, De Anza College
Javier Rueda, De Anza College
Lisa Rosenberg, Elon University
Mark Mills, Central College
Mary Jo Kane, De Anza College
Travis Short, St. Petersburg College
Valier Hauber, De Anza College
Vladimir Logvenenko, De Anza College
Wendy Lightheart, Lane Community College
Yvonne Sandoval, Pima Community College

Editorial Review Board

Linda Gann (6-12 Mathematics Coordinator, Boerne ISD) taught mathematics and statistics for over twenty-five years at Northside ISD, and currently serves as the 6-12 Mathematics Coordinator in Boerne ISD. She was awarded the Presidential Award for Excellence in Teaching, the Radio Shack National Teacher Award, the HEB Teaching Excellence Award (State Finalist), and the AP Siemens Award. For many years, Linda worked for the College Board as a consultant for AP Calculus AB, BC, and Statistics, and as a reader for AP Statistics. She has also served as the co-chair for the College and Career Readiness Standards for Mathematics for all three writing phases. Her educational background consists of a B.S. in Mathematics from Illinois State University and an M.S. in Mathematics from the University of Texas, San Antonio. Additionally, she is nearing completion of her Ph.D. in Interdisciplinary of Learning and Teaching from UTSA. She presently serves as president of the Alamo District Council of Teachers of Mathematics and scholarship chair for the Priest Holmes Foundation.

Wendy Martinez (Cedar Park High School) has been a teacher since 1994. She currently teaches PreAP Geometry and on-level Statistics at Cedar Park High School in Leander ISD. She has taught at Rouse High School, Lake Travis High School, and Pflugerville Middle School.

Alexander Teich (Rice University Graduate Student, Master's Degree in Applied Mathematics) has teaching experience back to 2004 and has taught math classes at Spring Woods High School, in Cambridge, Massachusetts, and Philadelphia, Pennsylvania. He formerly sponsored the Spring Wood Chess Club, and has a wealth of varied practical experience outside the classroom.

Amanda Yowell (Pleasant Grove High School) earned a Bachelors of Science in Business Administration and Finance from the University of Arkansas and worked in Financial Management. She teaches Mathematics classes at Pleasant Grove High School in Texarkana, TX. In her free time, she enjoys spending time with her husband and their three children.

1 SAMPLING AND DATA

Figure 1.1 We encounter statistics in our daily lives more often than we probably realize and from many different sources, like the news. (David Sim)

Chapter Objectives

By the end of this chapter, the student should be able to do the following:

- Recognize and differentiate between key terms
- Apply various types of sampling methods to data collection
- Create and interpret frequency tables

Introduction

You are probably asking yourself the question, "When and where will I use statistics?" If you read any newspaper, watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a television news program, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or *fact*. Statistical methods can help you make the *best educated guess*.

Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques for analyzing the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Included in this chapter are the basic ideas and words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and what *good* data can be distinguished from *bad*.

1.1 Definitions of Statistics, Probability, and Key Terms

The science of **statistics** deals with the collection, analysis, interpretation, and presentation of **data**. We see and use data in our everyday lives.

COLLABORATIVE EXERCISE

In your classroom, try this exercise. Have class members write down the average time—in hours, to the nearest half-hour—they sleep per night. Your instructor will record the data. Then create a simple graph, called a dot plot, of the data. A dot plot consists of a number line and dots, or points, positioned above the number line. For example, consider the following data:

5, 5.5, 6, 6, 6, 6.5, 6.5, 6.5, 6.5, 7, 7, 8, 8, 9.

The dot plot for this data would be as follows:

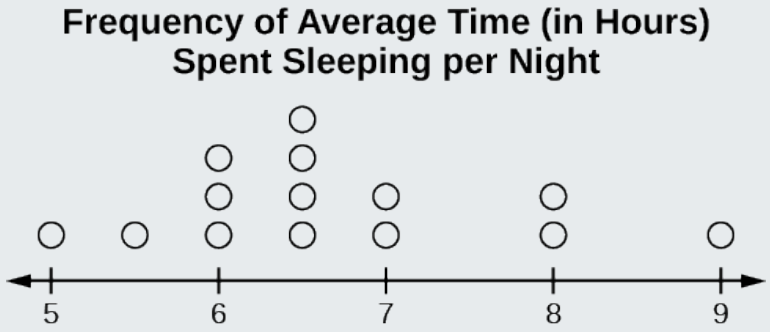


Figure 1.2

Does your dot plot look the same as or different from the example? Why? If you did the same example in an English class with the same number of students, do you think the results would be the same? Why or why not?

Where do your data appear to cluster? How might you interpret the clustering?

The questions above ask you to analyze and interpret your data. With this example, you have begun your study of statistics.

In this course, you will learn how to organize and summarize data. Organizing and summarizing data is called **descriptive statistics**. Two ways to summarize data are by graphing and by using numbers, for example, finding an average. After you have studied probability and probability distributions, you will use formal methods for drawing conclusions from *good* data. The formal methods are called **inferential statistics**. Statistical inference uses probability to determine how confident we can be that our conclusions are correct.

Effective interpretation of data, or inference, is based on good procedures for producing data and thoughtful examination of the data. You will encounter what will seem to be too many mathematical formulas for interpreting data. The goal of statistics is not to perform numerous calculations using the formulas, but to gain an understanding of your data. The calculations can be done using a calculator or a computer. The understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life.

Statistical Models

Statistics, like all other branches of mathematics, uses **mathematical models** to describe phenomena that occur in the real world. Some mathematical models are deterministic. These models can be used when one value is precisely determined from another value. Examples of deterministic models are the quadratic equations that describe the acceleration of a car from rest or the differential equations that describe the transfer of heat from a stove to a pot. These models are quite accurate and can be used to answer questions and make predictions with a high degree of precision. Space agencies, for example, use deterministic models to predict the exact amount of thrust that a rocket needs to break away from Earth's gravity and achieve orbit.

However, life is not always precise. While scientists can predict to the minute the time that the sun will rise, they cannot say precisely where a hurricane will make landfall. **Statistical models** can be used to predict life's more uncertain situations. These special forms of mathematical models or functions are based on the idea that one value affects another value. Some statistical models are mathematical functions that are more precise—one set of values can predict or determine another set of values. Or some statistical models are mathematical functions in which a set of values do not precisely determine other values. Statistical models are very useful because they can describe the probability or likelihood of an event occurring and provide alternative outcomes if the event does not occur. For example, weather forecasts are examples of statistical models. Meteorologists cannot predict tomorrow's weather with certainty. However, they often use statistical models to tell you how likely it is to rain at any given time, and you can prepare yourself based on this probability.

Probability

Probability is a mathematical tool used to study randomness. It deals with the chance of an event occurring. For example, if you toss a **fair** coin four times, the outcomes may not be two heads and two tails. However, if you toss the same coin 4,000 times, the outcomes will be close to half heads and half tails. The expected theoretical probability of

heads in any one toss is $\frac{1}{2}$ or .5. Even though the outcomes of a few repetitions are uncertain, there is a regular pattern of outcomes when there are many repetitions. After reading about the English statistician Karl **Pearson** who tossed a coin 24,000 times with a result of 12,012 heads, one of the authors tossed a coin 2,000 times. The results were 996 heads. The fraction $\frac{996}{2{,}000}$ is equal to .498 which is very close to .5, the expected probability.

The theory of probability began with the study of games of chance such as poker. Predictions take the form of probabilities. To predict the likelihood of an earthquake, of rain, or whether you will get an *A* in this course, we use probabilities. Doctors use probability to determine the chance of a vaccination causing the disease the vaccination is supposed to prevent. A stockbroker uses probability to determine the rate of return on a client's investments.

Key Terms

In statistics, we generally want to study a **population**. You can think of a population as a collection of persons, things, or objects under study. To study the population, we select a **sample**. The idea of **sampling** is to select a portion, or subset, of the larger population and study that portion—the sample—to gain information about the population. Data are the result of sampling from a population.

Because it takes a lot of time and money to examine an entire population, sampling is a very practical technique. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In presidential elections, opinion poll samples of 1,000–2,000 people are taken. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if a 16-ounce can contains 16 ounces of carbonated drink.

From the sample data, we can calculate a statistic. A **statistic** is a number that represents a property of the sample. For example, if we consider one math class as a sample of the population of *all* math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. Since we do not have the data for all math classes, that statistic is our best estimate of the average for the entire population of math classes. If we happen to have data for *all* math classes, we can find the population parameter. A **parameter** is a numerical characteristic of the whole population that can be estimated by a statistic. Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. In order to have an accurate sample, it must contain the characteristics of the population in order to be a **representative sample**. We are interested in both the sample statistic and the population parameter in inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

A **variable**, usually notated by capital letters such as *X* and *Y*, is a characteristic or measurement that can be determined for each member of a population. Variables may describe values like weight in pounds or favorite subject in school. **Numerical variables** take on values with equal units such as weight in pounds and time in hours. **Categorical variables** place the person or thing into a category. If we let *X* equal the number of points earned by one math student at the end of a term, then *X* is a numerical variable. If we let *Y* be a person's party affiliation, then some examples of *Y* include Republican, Democrat, and Independent. *Y* is a categorical variable. We could do some math with values of *X*—calculate the average number of points earned, for example—but it makes no sense to do math with values of *Y*—calculating an average party affiliation makes no sense.

Data are the actual values of the variable. They may be numbers or they may be words. **Datum** is a single value.

Two words that come up often in statistics are **mean** and **proportion**. If you were to take three exams in your math classes and obtain scores of 86, 75, and 92, you would calculate your mean score by adding the three exam scores and dividing by three. Your mean score would be 84.3 to one decimal place. If, in your math class, there are 40 students and 22 are males and 18 females, then the proportion of men students is $\frac{22}{40}$ and the proportion of women students is $\frac{18}{40}$. Mean and proportion are discussed in more detail in later chapters.

NOTE

The words *mean* and *average* are often used interchangeably. In this book, we use the term *arithmetic mean* for mean.

EXAMPLE 1.1

Problem

Determine what the population, sample, parameter, statistic, variable, and data referred to in the following study.

We want to know the mean amount of extracurricular activities in which high school students participate. We randomly surveyed 100 high school students. Three of those students were in 2, 5, and 7 extracurricular activities, respectively.

Solution

The **population** is all high school students.

The **sample** is the 100 high school students interviewed.

The **parameter** is the mean amount of extracurricular activities in which all high school students participate.

The **statistic** is the mean amount of extracurricular activities in which the sample of high school students participate.

The **variable** could be the amount of extracurricular activities by one high school student. Let X = the amount of extracurricular activities by one high school student.

The **data** are the number of extracurricular activities in which the high school students participate. Examples of the data are 2, 5, 7.

TRY IT 1.1

Find an article online or in a newspaper or magazine that refers to a statistical study or poll. Identify what each of the key terms—population, sample, parameter, statistic, variable, and data—refers to in the study mentioned in the article. Does the article use the key terms correctly?

EXAMPLE 1.2

Problem

Determine what the key terms refer to in the following study.

A study was conducted at a local high school to analyze the average cumulative GPAs of students who graduated last year. Fill in the letter of the phrase that best describes each of the items below.

1. Population ____ 2. Statistic ____ 3. Parameter ____ 4. Sample ____ 5. Variable ____ 6. Data ____

a) all students who attended the high school last year
b) the cumulative GPA of one student who graduated from the high school last year
c) 3.65, 2.80, 1.50, 3.90
d) a group of students who graduated from the high school last year, randomly selected
e) the average cumulative GPA of students who graduated from the high school last year
f) all students who graduated from the high school last year
g) the average cumulative GPA of students in the study who graduated from the high school last year

Solution

1. f; 2. g; 3. e; 4. d; 5. b; 6. c

EXAMPLE 1.3

Problem

Determine what the population, sample, parameter, statistic, variable, and data referred to in the following study.

As part of a study designed to test the safety of automobiles, the National Transportation Safety Board collected and reviewed data about the effects of an automobile crash on test dummies (The Data and Story Library, n.d.). Here is the

criterion they used.

Speed at which Cars Crashed	Location of *Driver* (i.e., dummies)
35 miles/hour	Front seat

Table 1.1

Cars with dummies in the front seats were crashed into a wall at a speed of 35 miles per hour. We want to know the proportion of dummies in the driver's seat that would have had head injuries, if they had been actual drivers. We start with a simple random sample of 75 cars.

Solution

The **population** is all cars containing dummies in the front seat.

The **sample** is the 75 cars, selected by a simple random sample.

The **parameter** is the proportion of driver dummies—if they had been real people—who would have suffered head injuries in the population.

The **statistic** is proportion of driver dummies—if they had been real people—who would have suffered head injuries in the sample.

The **variable** X = whether driver dummies—if they had been real people—would have suffered head injuries.

The **data** are either: yes, had head injury, or no, did not.

EXAMPLE 1.4

Problem

Determine what the population, sample, parameter, statistic, variable, and data referred to in the following study.

An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been involved in a malpractice lawsuit.

Solution

The **population** is all medical doctors listed in the professional directory.

The **parameter** is the proportion of medical doctors who have been involved in one or more malpractice suits in the population.

The **sample** is the 500 doctors selected at random from the professional directory.

The **statistic** is the proportion of medical doctors who have been involved in one or more malpractice suits in the sample.

The **variable** X records whether a doctor has or has not been involved in a malpractice suit.

The **data** are either: yes, was involved in one or more malpractice lawsuits; or no, was not.

COLLABORATIVE EXERCISE

Do the following exercise collaboratively with up to four people per group. Find a population, a sample, the parameter, the statistic, a variable, and data for the following study: You want to determine the average—mean—number of glasses of milk college students drink per day. Suppose yesterday, in your English class, you asked five students how many glasses of milk they drank the day before. The answers were 1, 0, 1, 3, and 4 glasses of milk.

1.2 Data, Sampling, and Variation in Data and Sampling

Data may come from a population or from a sample. Lowercase letters like x or y generally are used to represent data values. Most data can be put into the following categories:

- Qualitative
- Quantitative

Qualitative data are the result of categorizing or describing attributes of a population. Qualitative data are also often called **categorical data**. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative data. Qualitative data are generally described by words or letters. For instance, hair color might be black, dark brown, light brown, blonde, gray, or red. Blood type might be AB+, O–, or B+. Researchers often prefer to use quantitative data over qualitative data because it lends itself more easily to mathematical analysis. For example, it does not make sense to find an average hair color or blood type.

Quantitative data are always numbers. Quantitative data are the result of **counting** or **measuring** attributes of a population. Amount of money, pulse rate, weight, number of people living in your town, and number of students who take statistics are examples of quantitative data. Quantitative data may be either **discrete** or **continuous**.

All data that are the result of counting are called **quantitative discrete data**. These data take on only certain numerical values. If you count the number of phone calls you receive for each day of the week, you might get values such as zero, one, two, or three.

Data that are not only made up of counting numbers, but that may include fractions, decimals, or irrational numbers, are called **quantitative continuous data**. Continuous data are often the results of measurements like lengths, weights, or times. A list of the lengths in minutes for all the phone calls that you make in a week, with numbers like 2.4, 7.5, or 11.0, would be quantitative continuous data.

EXAMPLE 1.5

Data Sample of Quantitative Discrete Data

The data are the number of books students carry in their backpacks. You sample five students. Two students carry three books, one student carries four books, one student carries two books, and one student carries one book. The numbers of books, 3, 4, 2, and 1, are the quantitative discrete data.

TRY IT 1.5

The data are the number of machines in a gym. You sample five gyms. One gym has 12 machines, one gym has 15 machines, one gym has 10 machines, one gym has 22 machines, and the other gym has 20 machines. What type of data is this?

EXAMPLE 1.6

Data Sample of Quantitative Continuous Data

The data are the weights of backpacks with books in them. You sample the same five students. The weights, in pounds, of their backpacks are 6.2, 7, 6.8, 9.1, 4.3. Notice that backpacks carrying three books can have different weights. Weights are quantitative continuous data.

TRY IT 1.6

The data are the areas of lawns in square feet. You sample five houses. The areas of the lawns are 144 sq. ft., 160 sq. ft., 190 sq. ft., 180 sq. ft., and 210 sq. ft. What type of data is this?

EXAMPLE 1.7

You go to the supermarket and purchase three cans of soup (19 ounces tomato bisque, 14.1 ounces lentil, and 19 ounces Italian wedding), two packages of nuts (walnuts and peanuts), four different kinds of vegetable (broccoli, cauliflower, spinach, and carrots), and two desserts (16 ounces pistachio ice cream and 32 ounces chocolate chip cookies).

Problem

Name data sets that are quantitative discrete, quantitative continuous, and qualitative.

Solution

A possible solution

- One example of a quantitative discrete data set would be three cans of soup, two packages of nuts, four kinds of vegetables, and two desserts because you count them.
- The weights of the soups (19 ounces, 14.1 ounces, 19 ounces) are quantitative continuous data because you measure weights as precisely as possible.
- Types of soups, nuts, vegetables, and desserts are qualitative data because they are categorical.

Try to identify additional data sets in this example.

EXAMPLE 1.8

The data are the colors of backpacks. Again, you sample the same five students. One student has a red backpack, two students have black backpacks, one student has a green backpack, and one student has a gray backpack. The colors red, black, black, green, and gray are qualitative data.

TRY IT 1.8

The data are the colors of houses. You sample five houses. The colors of the houses are white, yellow, white, red, and white. What type of data is this?

Note

You may collect data as numbers and report it categorically. For example, the quiz scores for each student are recorded throughout the term. At the end of the term, the quiz scores are reported as A, B, C, D, or F.

EXAMPLE 1.9

Problem

Work collaboratively to determine the correct data type: quantitative or qualitative. Indicate whether quantitative data are continuous or discrete. Hint: Data that are discrete often start with the words *the number of*.

- the number of pairs of shoes you own
- the type of car you drive
- the distance from your home to the nearest grocery store
- the number of classes you take per school year
- the type of calculator you use
- weights of sumo wrestlers
- number of correct answers on a quiz
- IQ scores (This may cause some discussion.)

Solution

Items a, d, and g are quantitative discrete; items c, f, and h are quantitative continuous; items b and e are qualitative or

categorical.

TRY IT 1.9

Determine the correct data type, quantitative or qualitative, for the number of cars in a parking lot. Indicate whether quantitative data are continuous or discrete.

EXAMPLE 1.10

Problem

A statistics professor collects information about the classification of her students as freshmen, sophomores, juniors, or seniors. The data she collects are summarized in the pie chart Figure 1.3. What type of data does this graph show?

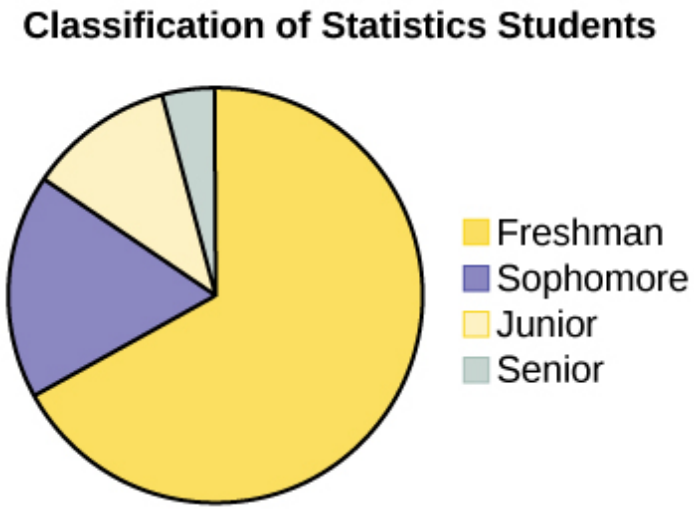


Figure 1.3

Solution

This pie chart shows the students in each year, which is **qualitative or categorical data**.

TRY IT 1.10

A large school district keeps data of the scores students earn on an end of the year standardized exam. The data he collects are summarized in the histogram. The class boundaries are 50 to less than 60, 60 to less than 70, 70 to less than 80, 80 to less than 90, and 90 to less than 100.

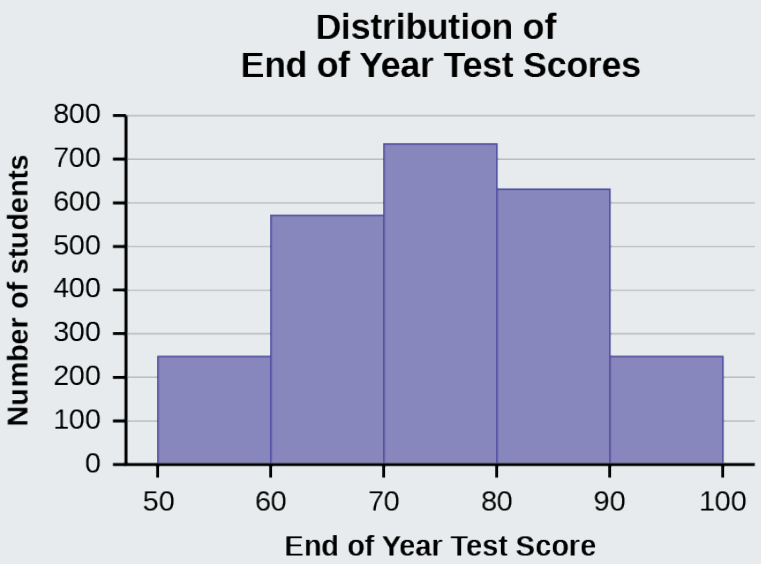


Figure 1.4

Qualitative Data Discussion

Below are tables comparing the number of part-time and full-time students at De Anza College and Foothill College enrolled for the spring 2010 quarter. The tables display counts, frequencies, and percentages or proportions, relative frequencies. For instance, to calculate the percentage of part time students at De Anza College, divide 9,200/22,496 to get .4089. Round to the nearest thousandth—third decimal place and then multiply by 100 to get the percentage, which is 40.9 percent.

So, the percent columns make comparing the same categories in the colleges easier. Displaying percentages along with the numbers is often helpful, but it is particularly important when comparing sets of data that do not have the same totals, such as the total enrollments for both colleges in this example. Notice how much larger the percentage for part-time students at Foothill College is compared to De Anza College.

De Anza College				Foothill College		
	Number	Percent			Number	Percent
Full-time	9,200	40.90%		Full-time	4,059	28.60%
Part-time	13,296	59.10%		Part-time	10,124	71.40%
Total	22,496	100%		Total	14,183	100%

Table 1.2 Fall Term 2007 (Census day)

Tables are a good way of organizing and displaying data. But graphs can be even more helpful in understanding the data.

Two graphs that are used to display qualitative data are pie charts and bar graphs.

In a **pie chart**, categories of data are shown by wedges in a circle that represent the percent of individuals/items in each category. We use pie charts when we want to show parts of a whole.

In a **bar graph**, the length of the bar for each category represents the number or percent of individuals in each category. Bars may be vertical or horizontal. We use bar graphs when we want to compare categories or show changes over time.

A **Pareto chart** consists of bars that are sorted into order by category size (largest to smallest).

Look at Figure 1.5 and Figure 1.6 and determine which graph (pie or bar) you think displays the comparisons better.

It is a good idea to look at a variety of graphs to see which is the most helpful in displaying the data. We might make different choices of what we think is the *best* graph depending on the data and the context. Our choice also depends on what we are using the data for.

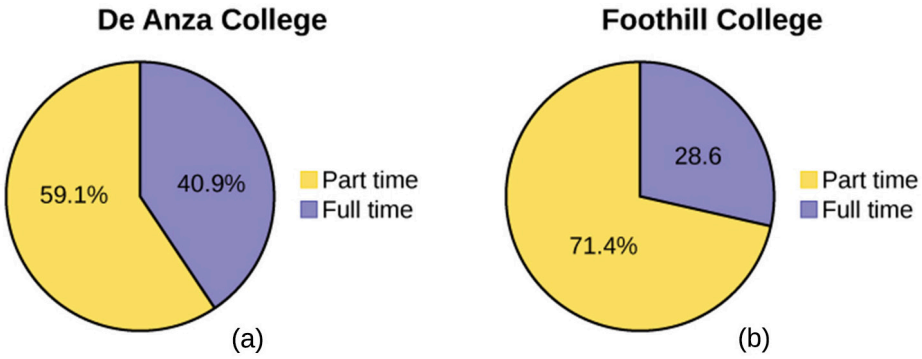


Figure 1.5

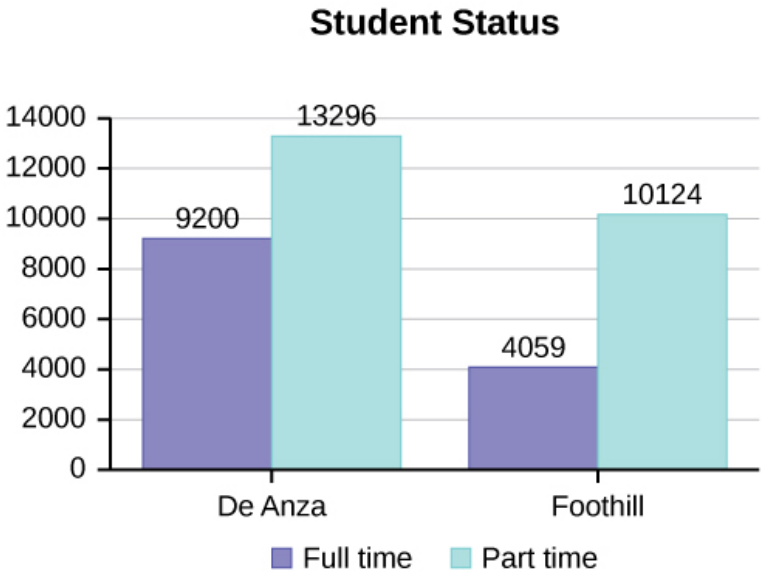


Figure 1.6

Percentages That Add to More (or Less) Than 100 Percent

Sometimes percentages add up to be more than 100 percent (or less than 100 percent). In the graph, the percentages add to more than 100 percent because students can be in more than one category. A bar graph is appropriate to compare the relative size of the categories. A pie chart cannot be used. It also could not be used if the percentages added to less than 100 percent.

Characteristic/Category	Percent
Students studying technical subjects	40.9%
Students studying non-technical subjects	48.6%
Students who intend to transfer to a four-year educational institutional	61.0%
TOTAL	150.5%

Table 1.3 De Anza College Year 2010

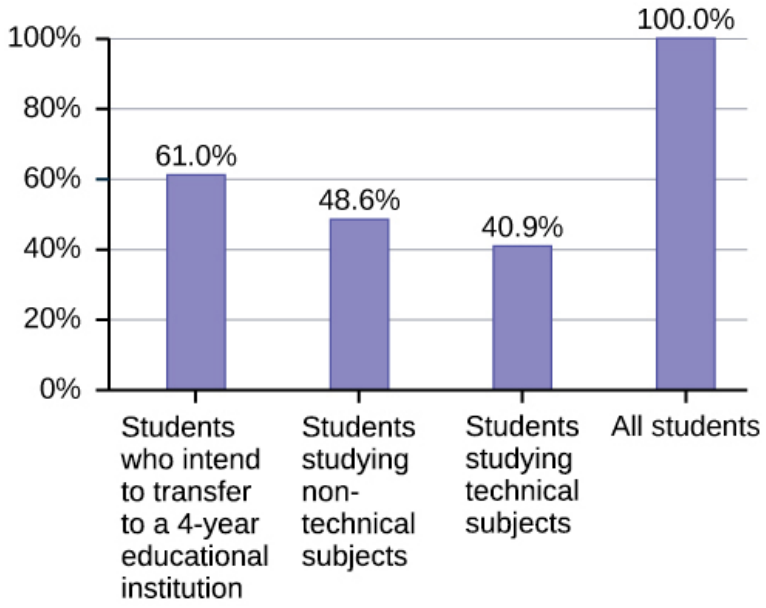


Figure 1.7

Omitting Categories/Missing Data

The table displays Ethnicity of Students but is missing the *Other/Unknown* category. This category contains people who did not feel they fit into any of the ethnicity categories or declined to respond. Notice that the frequencies do not add up to the total number of students. In this situation, create a bar graph and not a pie chart.

	Frequency	Percent
Asian	8,794	36.1%
Black	1,412	5.8%
Filipino	1,298	5.3%
Hispanic	4,180	17.1%
Native American	146	.6%
Pacific Islander	236	1.0%
White	5,978	24.5%
TOTAL	22,044 out of 24,382	90.4% out of 100%

Table 1.4 Ethnicity of Students at De Anza College Fall Term 2007 (Census Day)

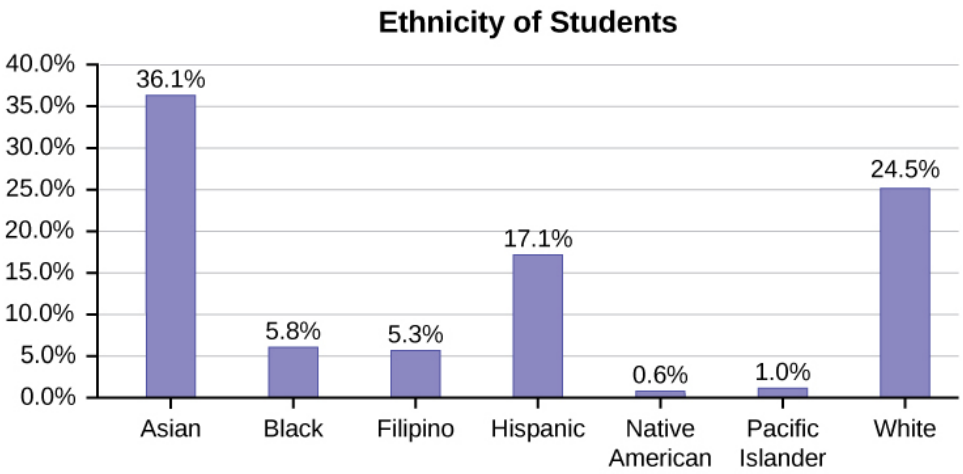


Figure 1.8

The following graph is the same as the previous graph but the *Other/Unknown* percent (9.6 percent) has been included. The *Other/Unknown* category is large compared to some of the other categories (Native American, .6 percent, Pacific Islander 1.0 percent). This is important to know when we think about what the data are telling us.

This particular bar graph in Figure 1.9 can be difficult to understand visually. The graph in Figure 1.10 is a Pareto chart. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.

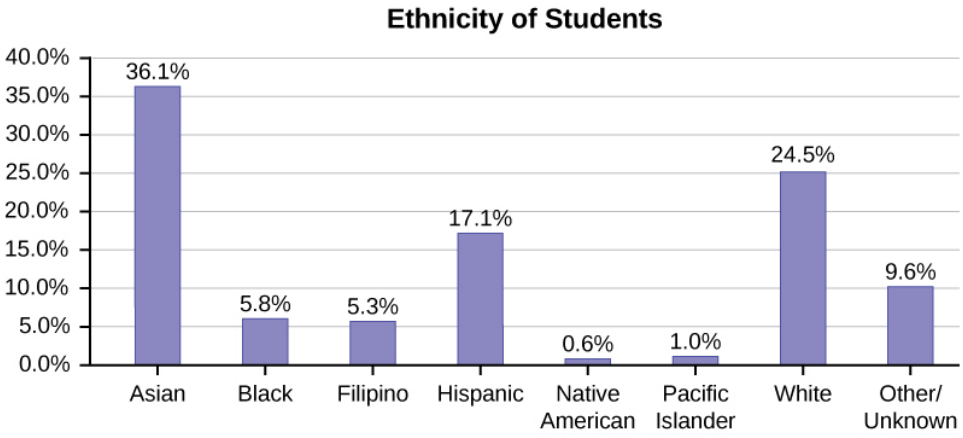


Figure 1.9 Bar Graph with Other/Unknown Category

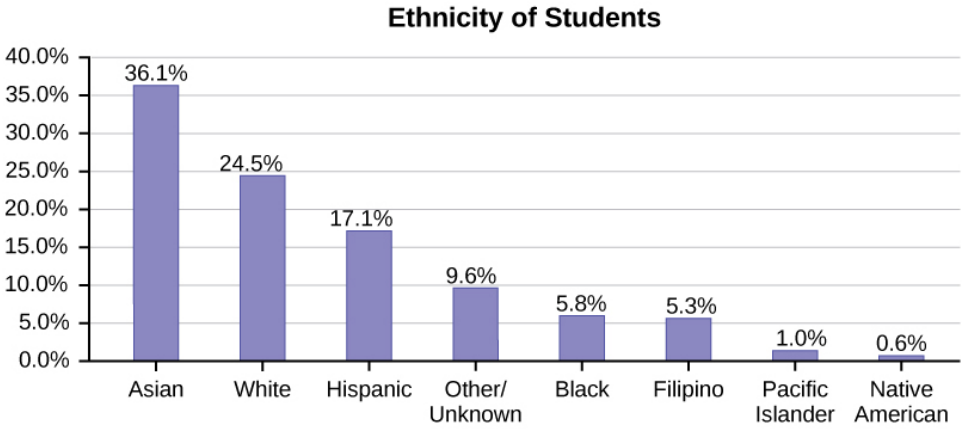


Figure 1.10 Pareto Chart With Bars Sorted by Size

Pie Charts: No Missing Data

The following pie charts have the *Other/Unknown* category included since the percentages must add to 100 percent. The chart in Figure 1.11b is organized by the size of each wedge, which makes it a more visually informative graph than the

unsorted, alphabetical graph in Figure 1.11a.

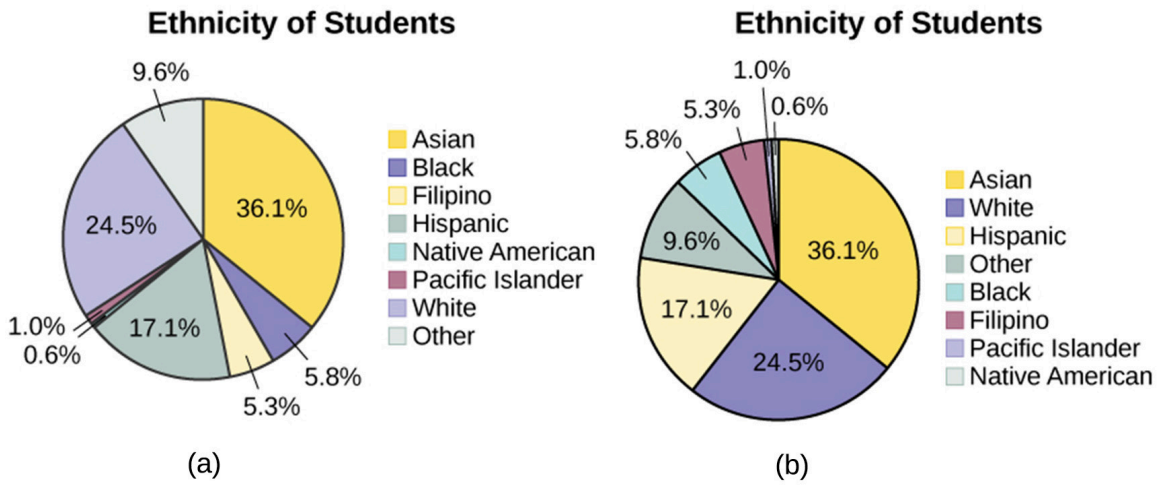


Figure 1.11

Marginal Distributions in Two-Way Tables

Below is a *two-way table*, also called a contingency table, showing the favorite sports for 50 adults: 20 women and 30 men.

	Football	Basketball	Tennis	Total
Men	20	8	2	30
Women	5	7	8	20
Total	25	15	10	50

Table 1.5

This is a two-way table because it displays information about two categorical variables, in this case, gender and sports. Data of this type (two variable data) are referred to as *bivariate data*. Because the data represent a count, or tally, of choices, it is a two-way *frequency* table. The entries in the total row and the total column represent marginal frequencies or *marginal distributions*. Note—The term marginal distributions gets its name from the fact that the distributions are found in the margins of frequency distribution tables. Marginal distributions may be given as a fraction or decimal: For example, the total for men could be given as .6 or 3/5 since $30/50 = .6 = 3/5$. Marginal distributions require bivariate data and only focus on *one* of the variables represented in the table. In other words, the reason 20 is a marginal frequency in this two-way table is because it represents the margin or portion of the total population that is women (20/50). The reason 25 is a marginal frequency is because it represents the portion of those sampled who favor football (25/50). Note: The values that make up the body of the table (e.g., 20, 8, 2) are called *joint frequencies*.

Conditional Distributions in Two-Way Tables

The distinction between a marginal distribution and a *conditional distribution* is that the focus is on only a particular subset of the population (not the entire population). For example, in the table, if we focused only on the subpopulation of women who prefer football, then we could calculate the conditional distributions as shown in the two-way table below.

	Football	Basketball	Tennis	Total
Men	20	8	2	30
Women	5	7	8	20
Total	25	15	10	50

Table 1.6

To find the first sub-population of women who prefer football, read the value at the intersection of the Women row and Football column which is 5. Then, divide this by the total population of football players which is 25. So, the subpopulation of football players who are women is 5/25 which is .2.

Similarly, to find the subpopulation of women who play football, use the value of 5 which is the number of women who play football. Then, divide this by the total population of women which is 20. So, the subpopulation of women who play football is 5/20 which is .25.

Presenting Data

After deciding which graph best represents your data, you may need to present your statistical data to a class or other group in an oral report or multimedia presentation. When giving an oral presentation, you must be prepared to explain exactly how you collected or calculated the data, as well as why you chose the categories, scales, and types of graphs that you are showing. Although you may have made numerous graphs of your data, be sure to use only those that actually demonstrate the stated intentions of your statistical study. While preparing your presentation, be sure that all colors, text, and scales are visible to the entire audience. Finally, make sure to allow time for your audience to ask questions and be prepared to answer them.

EXAMPLE 1.11

Problem

Suppose the guidance counselors at De Anza and Foothill need to make an oral presentation of the student data presented in Figures 1.5 and 1.6. Under what context should they choose to display the pie graph? When might they choose the bar graph? For each graph, explain which features they should point out and the potential display problems that might exist.

Solution

The guidance counselors should use the pie graph if the desired information is the percentage of each school's enrollment. They should use the bar graph if knowing the exact numbers of students and the relative sizes of each category at each school are important points to be made. For the pie graph, they should point out which color represents part-time students and which represents full-time students. They should also be sure that the numbers and colors are visible when displayed. For the bar graph, they should point out the scale and the total numbers for each category, and they should be sure that the numbers, colors, and scale marks are all displayed clearly.

TRY IT 1.11

Suppose you were asked to give an oral presentation of the data graphed in the pie chart in Figure 1.11(b). What features would you point out on the graph? What potential display problems with the graph should you check before giving your presentation?

Sampling

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we use a sample of the population. **A sample should have the same characteristics as the population it is representing.** Most statisticians use various methods of random sampling in an attempt to achieve this goal. This section will describe a few of the most common methods. There are several different methods of **random sampling**. In each form of random

sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons. The easiest method to describe is called a **simple random sample**. In a simple random sample, each group has the same chance of being selected. In other words, each sample of the same size has an equal chance of being selected. For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members not including Lisa. To choose a simple random sample of size three from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out three names. A more technological way is for Lisa to first list the last names of the members of her class together with a two-digit number, as in Table 1.7.

ID	Name	ID	Name	ID	Name
00	Anselmo	11	King	22	Roquero
01	Bautista	12	Legeny	23	Roth
02	Bayani	13	Lisa	24	Rowell
03	Cheng	14	Lundquist	25	Salangsang
04	Cuarismo	15	Macierz	26	Slade
05	Cuningham	16	Motogawa	27	Stratcher
06	Fontecha	17	Okimoto	28	Tallai
07	Hong	18	Patel	29	Tran
08	Hoobler	19	Price	30	Wai
09	Jiao	20	Quizon	31	Wood
10	Khan	21	Reyes		

Table 1.7 Class Roster

Lisa can use a table of random numbers (found in many statistics books and mathematical handbooks), a calculator, or a computer to generate random numbers. The most common random number generators are five digit numbers where each digit is a unique number from 0 to 9. For this example, suppose Lisa chooses to generate random numbers from a calculator. The numbers generated are as follows:

.94360, .99832, .14669, .51470, .40581, .73381, .04399.

Lisa reads two-digit groups until she has chosen three class members (That is, she reads .94360 as the groups 94, 43, 36, 60.) Each random number may only contribute one class member. If she needed to, Lisa could have generated more random numbers.

The table below shows how Lisa reads two-digit numbers form each random number. Each two-digit number in the table would represent each student in the roster above in Table 1.7.

Random number	Numbers read by Lisa			
.94360	94	43	36	60
.99832	99	98	83	32
.14669	14	46	66	69
.51470	51	14	47	70
.40581	40	05	58	81
.73381	73	33	38	81
.04399	04	39	39	99

Table 1.8 Lisa randomly generated the decimals in the Random Number column. She then used each consecutive number in each decimal to make the numbers she read. Some of the read numbers correspond with the ID numbers given to the students in her class (e.g., 14 = Lundquist in Table 1.7)

The random numbers .94360 and .99832 do not contain appropriate two digit numbers. However the third random number, .14669, contains 14 (the fourth random number also contains 14), the fifth random number contains 05, and the seventh random number contains 04. The two-digit number 14 corresponds to Lundquist, 05 corresponds to Cuningham, and 04 corresponds to Cuarismo. Besides herself, Lisa's group will consist of Lundquist, Cuningham, and Cuarismo.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

To generate random numbers perform the following steps:

- Press MATH.
- Arrow over to PRB.
- Press 5:randInt(0, 30).
- Press ENTER for the first random number.
- Press ENTER two more times for the other two random numbers. If there is a repeat press ENTER again.

Note—randInt(0, 30, 3) will generate three random numbers.

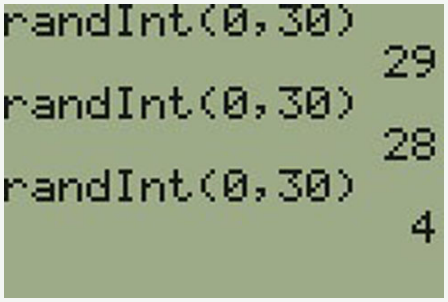


Figure 1.12

Besides simple random sampling, there are other forms of sampling that involve a chance process for getting the sample. **Other well-known random sampling methods are the stratified sample, the cluster sample, and the**

systematic sample.

To choose a stratified sample, divide the population into groups called strata and then the sample is selected by picking the same number of values from each strata until the desired sample size is reached. For example, you could stratify (group) your high school student population by year (freshmen, sophomore, juniors, and seniors) and then choose a proportionate simple random sample from each stratum (each year) to get a stratified random sample. To choose a simple random sample from each year, number each student of the first year, number each student of the second year, and do the same for the remaining years. Then use simple random sampling to choose proportionate numbers of students from the first year and do the same for each of the remaining years. Those numbers picked from the first year, picked from the second year, and so on represent the students who make up the stratified sample.

To choose a cluster sample, divide the population into clusters (groups) and then randomly select some of the clusters. All the members from these clusters are in the cluster sample. For example, if you randomly sample four homeroom classes from your student population, the four classes make up the cluster sample. Each class is a cluster. Number each cluster, and then choose four different numbers using random sampling. All the students of the four classes with those numbers are the cluster sample. So, unlike a stratified example, a cluster sample may not contain an equal number of randomly chosen students from each class.

To choose a systematic sample, establish and follow a rule. The most common way to select a systematic sample is to list the members of the population and choose every n^{th} entry from a random starting point. For example, suppose you have 100,000 individuals in your population and you want to choose a sample of 1,000. Use a random number generator to select your starting point. Now, 100,000/1,000 = 100, so to ensure coverage throughout the list, choose every 100^{th} entry in the list. When you reach the end of the list, continue the count from the beginning until you have selected the complete sample.

A type of sampling that is non-random is convenience sampling. **Convenience sampling** involves using results that are readily available. For example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. The results of convenience sampling may be very good in some cases and highly biased (favor certain outcomes) in others.

Sampling data should be done very carefully. Collecting data carelessly can have devastating results. Surveys mailed to households and then returned may be very biased. They may favor a certain group. It is better for the person conducting the survey to select the sample respondents.

When you analyze data, it is important to be aware of **sampling errors** and nonsampling errors. The actual process of sampling causes sampling errors. For example, the sample may not be large enough. Factors not related to the sampling process cause **nonsampling errors**. A defective counting device can cause a nonsampling error.

In reality, a sample will never be exactly representative of the population so there will always be some sampling error. As a rule, the larger the sample, the smaller the sampling error.

In statistics, **a sampling bias** is created when a sample is collected from a population and some members of the population are not as likely to be chosen as others. Remember, each member of the population should have an equally likely chance of being chosen. When a sampling bias happens, there can be incorrect conclusions drawn about the population that is being studied. For instance, if a survey of all students is conducted only during noon lunchtime hours is biased. This is because the students who do not have a noon lunchtime would not be included.

Critical Evaluation

We need to evaluate the statistical studies we read about critically and analyze them before accepting the results of the studies. Common problems to be aware of include the following:

- Problems with samples: —A sample must be representative of the population. A sample that is not representative of the population is biased. Biased samples that are not representative of the population give results that are inaccurate and not reliable. Reliability in statistical measures must also be considered when analyzing data. Reliability refers to the consistency of a measure. A measure is reliable when the same results are produced given the same circumstances.
- Self-selected samples—Responses only by people who choose to respond, such as internet surveys, are often unreliable.
- Sample size issues—: Samples that are too small may be unreliable. Larger samples are better, if possible. In some situations, having small samples is unavoidable and can still be used to draw conclusions. Examples include crash testing cars or medical testing for rare conditions.
- Undue influence—: collecting data or asking questions in a way that influences the response.

- Non-response or refusal of subject to participate: —The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can affect the results.
- Causality: —A relationship between two variables does not mean that one causes the other to occur. They may be related (correlated) because of their relationship through a different variable.
- Self-funded or self-interest studies—: A study performed by a person or organization in order to support their claim. Is the study impartial? Read the study carefully to evaluate the work. Do not automatically assume that the study is good, but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.
- Misleading use of data—: These can be improperly displayed graphs, incomplete data, or lack of context.

COLLABORATIVE EXERCISE

As a class, determine whether or not the following samples are representative. If they are not, discuss the reasons.

1. To find the average GPA of all students in a high school, use all honor students at the university as the sample.
2. To find out the most popular cereal among young people under the age of 10, stand outside a large supermarket for three hours and speak to every twentieth child under age 10 who enters the supermarket.
3. To find the average annual income of all adults in the United States, sample U.S. congressmen. Create a cluster sample by considering each state as a stratum (group). By using simple random sampling, select states to be part of the cluster. Then survey every U.S. congressman in the cluster.
4. To determine the proportion of people taking public transportation to work, survey 20 people in New York City. Conduct the survey by sitting in Central Park on a bench and interviewing every person who sits next to you.
5. To determine the average cost of a two-day stay in a hospital in Massachusetts, survey 100 hospitals across the state using simple random sampling.

EXAMPLE 1.12

Problem

A study is done to determine the average tuition that private high school students pay per semester. Each student in the following samples is asked how much tuition he or she paid for the fall semester. What is the type of sampling in each case?

a. A sample of 100 high school students is taken by organizing the students' names by classification (freshman, sophomore, junior, or senior) and then selecting 25 students from each.
b. A random number generator is used to select a student from the alphabetical listing of all high school students in the fall semester. Starting with that student, every 50th student is chosen until 75 students are included in the sample.
c. A completely random method is used to select 75 students. Each high school student in the fall semester has the same probability of being chosen at any stage of the sampling process.
d. The freshman, sophomore, junior, and senior years are numbered one, two, three, and four, respectively. A random number generator is used to pick two of those years. All students in those two years are in the sample.
e. An administrative assistant is asked to stand in front of the library one Wednesday and to ask the first 100 undergraduate students he encounters what they paid for tuition the fall semester. Those 100 students are the sample.

Solution

a. stratified, b. systematic, c. simple random, d. cluster, e. convenience

TRY IT 1.12

You are going to use the random number generator to generate different types of samples from the data.

This table displays six sets of quiz scores (each quiz counts 10 points) for an elementary statistics class.

#1	#2	#3	#4	#5	#6
5	7	10	9	8	3
10	5	9	8	7	6
9	10	8	6	7	9
9	10	10	9	8	9
7	8	9	5	7	4
9	9	9	10	8	7
7	7	10	9	8	8
8	8	9	10	8	8
9	7	8	7	7	8
8	8	10	9	8	7

Table 1.9 Scores for quizzes #1-6 for 10 students in a statistics class. Each quiz is out of 10 points.

Instructions: Use the Random Number Generator to pick samples.

1. Create a stratified sample by column. Pick three quiz scores randomly from each column.
   a. Number each row one through 10.
   b. On your calculator, press Math and arrow over to PRB.
   c. For column 1, Press 5:randInt(and enter 1,10). Press ENTER. Record the number. Press ENTER 2 more times (even the repeats). Record these numbers. Record the three quiz scores in column one that correspond to these three numbers.
   d. Repeat for columns two through six.
   e. These 18 quiz scores are a stratified sample.
2. Create a cluster sample by picking two of the columns. Use the column numbers: one through six.
   a. Press MATH and arrow over to the PRB function.
   b. Press 5:randInt (“and then enter “1,6). Press ENTER.
   c. The number the calculator displays names the first column of quiz scores to include in your sample. Press ENTER.
   d. The next number the calculator displays identifies the second column, or cluster, of data to include in the sample, giving a total of 20 quiz scores.
3. Create a simple random sample of 15 quiz scores.
   a. Use the numbering one through 60.
   b. Press MATH. Arrow over to PRB. Press 5:randInt(1, 60).
   c. Press ENTER 15 times and record the numbers.
   d. Record the quiz scores that correspond to these numbers.
   e. These 15 quiz scores are the systematic sample.
4. Create a systematic sample of 12 quiz scores.

a. Use the numbering one through 60.
b. Press MATH. Arrow over to PRB. Press 5:randInt(1, 60).
c. Press ENTER. Record the number and the first quiz score. From that number, count ten quiz scores and record that quiz score. Keep counting ten quiz scores and recording the quiz score until you have a sample of 12 quiz scores. You may wrap around (go back to the beginning).

EXAMPLE 1.13

Problem

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

a. A soccer coach selects six players from a group of boys aged eight to ten, seven players from a group of boys aged 11 to 12, and three players from a group of boys aged 13 to 14 to form a recreational soccer team.
b. A pollster interviews all human resource personnel in five different high tech companies.
c. A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.
d. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.
e. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.
f. A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on average.

Solution

a. stratified b. cluster c. stratified d. systematic e. simple random f. convenience

TRY IT 1.13

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

A high school principal polls 50 freshmen, 50 sophomores, 50 juniors, and 50 seniors regarding policy changes for after school activities.

If we were to examine two samples representing the same population, even if we used random sampling methods for the samples, they would not be exactly the same. Just as there is variation in data, there is variation in samples. As you become accustomed to sampling, the variability will begin to seem natural.

EXAMPLE 1.14

Suppose ABC college has 10,000 upperclassman (junior and senior level) students (the population). We are interested in the average amount of money an upperclassmen spends on books in the fall term. Asking all 10,000 upperclassmen is an almost impossible task.

Suppose we take two different samples.

First, we use convenience sampling and survey ten upperclassman students from a first term organic chemistry class. Many of these students are taking first term calculus in addition to the organic chemistry class. The amount of money they spend on books is as follows:

$128, $87, $173, $116, $130, $204, $147, $189, $93, $153.

The second sample is taken using a list of seniors who take P.E. classes and taking every fifth senior on the list, for a total of ten seniors. They spend the following:

$50, $40, $36, $15, $50, $100, $40, $53, $22, $22.

It is unlikely that any student is in both samples.

Problem

a. Do you think that either of these samples is representative of (or is characteristic of) the entire 10,000 part-time student population?

Solution

a. No. The first sample probably consists of science-oriented students. Besides the chemistry course, some of them are also taking first-term calculus. Books for these classes tend to be expensive. Most of these students are, more than likely, paying more than the average part-time student for their books. The second sample is a group of seniors who are, more than likely, taking courses for health and interest. The amount of money they spend on books is probably much less than the average part-time student. Both samples are biased. Also, in both cases, not all students have a chance to be in either sample.

Problem

b. Since these samples are not representative of the entire population, is it wise to use the results to describe the entire population?

Solution

b. No. For these samples, each member of the population did not have an equally likely chance of being chosen.

Now, suppose we take a third sample. We choose ten different part-time students from the disciplines of chemistry, math, English, psychology, sociology, history, nursing, physical education, art, and early childhood development. We assume that these are the only disciplines in which part-time students at ABC College are enrolled and that an equal number of part-time students are enrolled in each of the disciplines. Each student is chosen using simple random sampling. Using a calculator, random numbers are generated and a student from a particular discipline is selected if he or she has a corresponding number. The students spend the following amounts:

$180, $50, $150, $85, $260, $75, $180, $200, $200, $150.

Problem

c. Is the sample biased?

Solution

c. The sample is unbiased, but a larger sample would be recommended to increase the likelihood that the sample will be close to representative of the population. However, for a biased sampling technique, even a large sample runs the risk of not being representative of the population.

Students often ask if it is *good enough* to take a sample, instead of surveying the entire population. If the survey is done well, the answer is yes.

TRY IT 1.14

A local radio station has a fan base of 20,000 listeners. The station wants to know if its audience would prefer more music or more talk shows. Asking all 20,000 listeners is an almost impossible task.

The station uses convenience sampling and surveys the first 200 people they meet at one of the station's music concert events. Twenty-four people said they'd prefer more talk shows, and 176 people said they'd prefer more music.

Do you think that this sample is representative of (or is characteristic of) the entire 20,000 listener population?

Variation in Data

Variation is present in any set of data. For example, 16-ounce cans of beverage may contain more or less than 16 ounces of liquid. In one study, eight 16 ounce cans were measured and produced the following amount (in ounces) of beverage:

15.8, 16.1, 15.2, 14.8, 15.8, 15.9, 16.0, 15.5.

Measurements of the amount of beverage in a 16-ounce can may vary because different people make the measurements or because the exact amount, 16 ounces of liquid, was not put into the cans. Manufacturers regularly run tests to determine if the amount of beverage in a 16-ounce can falls within the desired range.

Be aware that as you take data, your data may vary somewhat from the data someone else is taking for the same purpose. This is completely natural. However, if two or more of you are taking the same data and get very different results, it is time for you and the others to reevaluate your data-taking methods and your accuracy.

Variation in Samples

It was mentioned previously that two or more **samples** from the same **population**, taken randomly, and having close to the same characteristics of the population will likely be different from each other. Suppose Doreen and Jung both decide to study the average amount of time students at their high school sleep each night. Doreen and Jung each take samples of 500 students. Doreen uses systematic sampling and Jung uses cluster sampling. Doreen's sample will be different from Jung's sample. Even if Doreen and Jung used the same sampling method, in all likelihood their samples would be different. Neither would be wrong, however.

Think about what contributes to making Doreen's and Jung's samples different.

If Doreen and Jung took larger samples, that is, the number of data values is increased, their sample results (the average amount of time a student sleeps) might be closer to the actual population average. But still, their samples would be, in all likelihood, different from each other. This is called **sampling variability**. In other words, it refers to how much a statistic varies from sample to sample within a population. The larger the sample size, the smaller the variability between samples will be. So, the large sample size makes for a better, more reliable statistic.

Size of a Sample

The size of a sample (often called the number of observations) is important. The examples you have seen in this book so far have been small. Samples of only a few hundred observations, or even smaller, are sufficient for many purposes. In polling, samples that are from 1,200–1,500 observations are considered large enough and good enough if the survey is random and is well done. You will learn why when you study confidence intervals.

Be aware that many large samples are biased. For example, internet surveys are invariably biased, because people choose to respond or not.

COLLABORATIVE EXERCISE

Divide into groups of two, three, or four. Your instructor will give each group one six-sided die. Try this experiment twice. Roll one fair die (six-sided) 20 times. Record the number of ones, twos, threes, fours, fives, and sixes you get in Table 1.10 and Table 1.11 (*frequency* is the number of times a particular face of the die occurs)

Face on Die	Frequency
1	
2	
3	
4	
5	
6	

Table 1.10 First Experiment (20 rolls)

Face on Die	Frequency
1	
2	
3	
4	
5	
6	

Table 1.11 Second Experiment (20 rolls)

Did the two experiments have the same results? Probably not. If you did the experiment a third time, do you expect the results to be identical to the first or second experiment? Why or why not?

Which experiment had the correct results? They both did. The job of the statistician is to see through the variability and draw appropriate conclusions.

1.3 Frequency, Frequency Tables, and Levels of Measurement

Once you have a set of data, you will need to organize it so that you can analyze how frequently each datum occurs in the set. However, when calculating the frequency, you may need to round your answers so that they are as precise as possible.

Answers and Rounding Off

A simple way to round off answers is to carry your final answer one more decimal place than was present in the original data. Round off only the final answer. Do not round off any intermediate results, if possible. If it becomes necessary to round off intermediate results, carry them to at least twice as many decimal places as the final answer. Expect that some of your answers will vary from the text due to rounding errors.

It is not necessary to reduce most fractions in this course. Especially in Probability Topics, the chapter on probability, it is more helpful to leave an answer as an unreduced fraction.

Levels of Measurement

The way a set of data is measured is called its **level of measurement**. Correct statistical procedures depend on a researcher being familiar with levels of measurement. Not every statistical operation can be used with every set of data. Data can be classified into four levels of measurement. They are as follows (from lowest to highest level):

- **Nominal scale level**
- **Ordinal scale level**
- **Interval scale level**
- **Ratio scale level**

Data that is measured using a **nominal scale** is **qualitative (categorical)**. Categories, colors, names, labels, and favorite foods along with yes or no responses are examples of nominal level data. Nominal scale data are not ordered. For example, trying to classify people according to their favorite food does not make any sense. Putting pizza first and sushi second is not meaningful.

Smartphone companies are another example of nominal scale data. The data are the names of the companies that make smartphones, but there is no agreed upon order of these brands, even though people may have personal preferences. Nominal scale data cannot be used in calculations.

Data that is measured using an **ordinal scale** is similar to nominal scale data but there is a big difference. The ordinal

scale data can be ordered. An example of ordinal scale data is a list of the top five national parks in the United States. The top five national parks in the United States can be ranked from one to five but we cannot measure differences between the data.

Another example of using the ordinal scale is a cruise survey where the responses to questions about the cruise are *excellent*, *good*, *satisfactory*, and *unsatisfactory*. These responses are ordered from the most desired response to the least desired. But the differences between two pieces of data cannot be measured. Like the nominal scale data, ordinal scale data cannot be used in calculations.

Like data measured on an ordinal scale, data that are measured on an **interval scale** have a definite ordering. While ordinal scale data are categorical, or qualitative, interval scale data are numerical, or quantitative. It is possible to calculate differences between values measured on an interval scale. There is no minimum or "zero" value, however.

Temperature scales like Celsius (C) and Fahrenheit (F) are measured by using the interval scale. In both temperature measurements, 40° is equal to 100° minus 60°. Differences make sense. But 0 degrees does not represent a minimum value. In both scales, 0 is not the absolute lowest temperature. Temperatures like –10 °F and –15 °C exist and are colder than 0.

Data that are measured using the **ratio scale** give the most information. Ratio scale data is like interval scale data, but there is a minimum value and ratios can be calculated. For example, four multiple choice statistics final exam scores are 80, 68, 20 and 92 (out of a possible 100 points). The exams are machine-graded.

The data can be put in order from lowest to highest 20, 68, 80, 92.

The differences between the data have meaning. The score 92 is more than the score 68 by 24 points. Ratios can be calculated. The minimum score is 0. So 80 is four times 20. The score of 80 is four times better than the score of 20.

Frequency

Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows: 5, 6, 3, 3, 2, 4, 7, 5, 2, 3, 5, 6, 5, 4, 4, 3, 5, 2, 5, 3.

Table 1.12 lists the different data values in ascending order and their frequencies.

DATA VALUE	FREQUENCY
2	3
3	5
4	3
5	6
6	2
7	1

Table 1.12 Frequency Table of Student Work Hours

A **frequency** is the number of times a value of the data occurs. According to Table 1.12, there are three students who work two hours, five students who work three hours, and so on. The sum of the values in the frequency column, 20, represents the total number of students included in the sample.

A **relative frequency** is the ratio (fraction or proportion) of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes. To find the relative frequencies, divide each frequency by the total number of students in the sample, in this case, 20. Relative frequencies can be written as fractions, percents, or decimals.

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or .15
3	5	$\frac{5}{20}$ or .25
4	3	$\frac{3}{20}$ or .15
5	6	$\frac{6}{20}$ or .30
6	2	$\frac{2}{20}$ or .10
7	1	$\frac{1}{20}$ or .05

Table 1.13 Frequency Table of Student Work Hours with Relative Frequencies

The sum of the values in the relative frequency column of Table 1.13 is $\frac{20}{20}$, or 1.

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row, as shown in Table 1.14.

In the first row, the cumulative frequency is simply .15 because it is the only one. In the second row, the relative frequency was .25, so adding that to .15, we get a relative frequency of .40. Continue adding the relative frequencies in each row to get the rest of the column.

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or .15	.15
3	5	$\frac{5}{20}$ or .25	.15 + .25 = .40
4	3	$\frac{3}{20}$ or .15	.40 + .15 = .55
5	6	$\frac{6}{20}$ or .30	.55 + .30 = .85
6	2	$\frac{2}{20}$ or .10	.85 + .10 = .95
7	1	$\frac{1}{20}$ or .05	.95 + .05 = 1.00

Table 1.14 Frequency Table of Student Work Hours with Relative and Cumulative Relative Frequencies

The last entry of the cumulative relative frequency column is one, indicating that one hundred percent of the data has been accumulated.

NOTE

Because of rounding, the relative frequency column may not always sum to one, and the last entry in the cumulative relative frequency column may not be one. However, they each should be close to one.

The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. The heights are **continuous** data since height is measured.
60, 60.5, 61, 61, 61.5,
63.5, 63.5, 63.5,
64, 64, 64, 64, 64, 64, 64, 64.5, 64.5, 64.5, 64.5, 64.5, 64.5, 64.5, 64.5,
66, 66, 66, 66, 66, 66, 66, 66, 66, 66, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 67, 67, 67, 67, 67, 67, 67,
67, 67, 67, 67, 67, 67.5, 67.5, 67.5, 67.5, 67.5, 67.5, 67.5,
68, 68, 69, 69, 69, 69, 69, 69, 69, 69, 69, 69, 69.5, 69.5, 69.5, 69.5, 69.5,
70, 70, 70, 70, 70, 70, 70.5, 70.5, 70.5, 71, 71, 71,
72, 72, 72, 72.5, 72.5, 73, 73.5,
74

Table 1.15 summarizes the heights in this sample. Since heights are expressed in tenths, the frequency table will use labels measured in hundredths. This ensures that no data value will coincide with the upper or lower limit of an interval.

HEIGHTS (INCHES)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
59.95–61.95	5	$\frac{5}{100}$ = .05	.05
61.95–63.95	3	$\frac{3}{100}$ = .03	.05 + .03 = .08
63.95–65.95	15	$\frac{15}{100}$ = .15	.08 + .15 = .23
65.95–67.95	40	$\frac{40}{100}$ = .40	.23 + .40 = .63
67.95–69.95	17	$\frac{17}{100}$ = .17	.63 + .17 = .80
69.95–71.95	12	$\frac{12}{100}$ = .12	.80 + .12 = .92
71.95–73.95	7	$\frac{7}{100}$ = .07	.92 + .07 = .99
73.95–75.95	1	$\frac{1}{100}$ = .01	.99 + .01 = 1.00
	Total = 100	**Total = 1.00**	

Table 1.15 Frequency Table of Soccer Player Height

The data in this table have been **grouped** into the following intervals:

- 59.95–61.95 inches
- 61.95–63.95 inches
- 63.95–65.95 inches
- 65.95–67.95 inches
- 67.95–69.95 inches
- 69.95–71.95 inches
- 71.95–73.95 inches
- 73.95–75.95 inches

Note

This example is used again in Descriptive Statistics, where the method used to compute the intervals will be explained.

In this sample, there are **five** players whose heights fall within the interval 59.95–61.95 inches, **three** players whose heights fall within the interval 61.95–63.95 inches, **15** players whose heights fall within the interval 63.95–65.95 inches, **40** players whose heights fall within the interval 65.95–67.95 inches, **17** players whose heights fall within the interval 67.95–69.95 inches, **12** players whose heights fall within the interval 69.95–71.95, **seven** players whose heights fall within the interval 71.95–73.95, and **one** player whose heights fall within the interval 73.95–75.95. All heights fall between the endpoints of an interval and not at the endpoints.

EXAMPLE 1.15

Problem

From Table 1.15, find the percentage of heights that are less than 65.95 inches.

Solution

If you look at the first, second, and third rows, the heights are all less than 65.95 inches. There are 5 + 3 + 15 = 23 players whose heights are less than 65.95 inches. The percentage of heights less than 65.95 inches is then $\frac{23}{100}$ or 23 percent. This percentage is the cumulative relative frequency entry in the third row.

TRY IT 1.15

Table 1.16 shows the amount, in inches, of annual rainfall in a sample of towns.

Rainfall (Inches)	Frequency	Relative Frequency	Cumulative Relative Frequency
2.95–4.97	6	$\frac{6}{50}$ = .12	.12
4.97–6.99	7	$\frac{7}{50}$ = .14	.12 + .14 = .26
6.99–9.01	15	$\frac{15}{50}$ = .30	.26 + .30 = .56
9.01–11.03	8	$\frac{8}{50}$ = .16	.56 + .16 = .72
11.03–13.05	9	$\frac{9}{50}$ = .18	.72 + .18 = .90
13.05–15.07	5	$\frac{5}{50}$ = .10	.90 + .10 = 1.00
	Total = 50	Total = 1.00	

Table 1.16

From Table 1.16, find the percentage of rainfall that is less than 9.01 inches.

EXAMPLE 1.16

Problem

From Table 1.15, find the percentage of heights that fall between 61.95 and 65.95 inches.

Solution

Add the relative frequencies in the second and third rows: .03 + .15 = .18 or 18 percent.

TRY IT 1.16

From Table 1.16, find the percentage of rainfall that is between 6.99 and 13.05 inches.

EXAMPLE 1.17

Problem

Use the heights of the 100 male semiprofessional soccer players in Table 1.15. Fill in the blanks and check your answers.

a. The percentage of heights that are from 67.95–71.95 inches is _____.
b. The percentage of heights that are from 67.95–73.95 inches is _____.
c. The percentage of heights that are more than 65.95 inches is _____.
d. The number of players in the sample who are between 61.95 and 71.95 inches tall is _____.
e. What kind of data are the heights?
f. Describe how you could gather this data (the heights) so that the data are characteristic of all male semiprofessional soccer players.

Remember, you **count frequencies**. To find the relative frequency, divide the frequency by the total number of data values. To find the cumulative relative frequency, add all of the previous relative frequencies to the relative frequency for the current row.

Solution

a. 29 percent
b. 36 percent
c. 77 percent
d. 87
e. quantitative continuous
f. get rosters from each team and choose a simple random sample from each

TRY IT 1.17

From Table 1.16, find the number of towns that have rainfall between 2.95 and 9.01 inches.

COLLABORATIVE EXERCISE

In your class, have someone conduct a survey of the number of siblings (brothers and sisters) each student has. Create a frequency table. Add to it a relative frequency column and a cumulative relative frequency column. Answer the following questions:

1. What percentage of the students in your class have no siblings?
2. What percentage of the students have from one to three siblings?
3. What percentage of the students have fewer than three siblings?

EXAMPLE 1.18

Nineteen people were asked how many miles, to the nearest mile, they commute to work each day. The data are as follows: 2; 5; 7; 3; 2; 10; 18; 15; 20; 7; 10; 18; 5; 12; 13; 12; 4; 5; 10. Table 1.17 was produced.

DATA	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
3	3	$\frac{3}{19}$	.1579
4	1	$\frac{1}{19}$	.2105
5	3	$\frac{3}{19}$	.1579
7	2	$\frac{2}{19}$	.2632
10	3	$\frac{4}{19}$	.4737
12	2	$\frac{2}{19}$	.7895
13	1	$\frac{1}{19}$	.8421
15	1	$\frac{1}{19}$	.8948
18	1	$\frac{1}{19}$	.9474
20	1	$\frac{1}{19}$	1.0000

Table 1.17 Frequency of Commuting Distances

Problem

a. Is the table correct? If it is not correct, what is wrong?
b. True or False: Three percent of the people surveyed commute three miles. If the statement is not correct, what should it be? If the table is incorrect, make the corrections.
c. What fraction of the people surveyed commute five or seven miles?
d. What fraction of the people surveyed commute 12 miles or more? Less than 12 miles? Between five and 13 miles (not including five and 13 miles)?

Solution

a. No. The frequency column sums to 18, not 19. Not all cumulative relative frequencies are correct. The table entries for data values 2, 3, 10, and 18 are incorrect. This affects cumulative relative frequency for most values.
b. False. The frequency for three miles should be one; for two miles (left out), two. The cumulative relative frequency column should read 1052, .1579, .2105, .3684, .4737, .6316, .7368, .7895, .8421, .9474, 1.0000.
c. $\frac{5}{19}$
d. $\frac{7}{19}$, $\frac{12}{19}$, $\frac{7}{19}$

TRY IT 1.18

Table 1.16 represents the amount, in inches, of annual rainfall in a sample of towns. What fraction of towns surveyed get between 11.03 and 13.05 inches of rainfall each year?

EXAMPLE 1.19

Table 1.18 contains the total number of deaths worldwide as a result of earthquakes for the period from 2000 to 2012.

Year	Total Number of Deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003
2006	6,605
2007	712
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	823,856

Table 1.18

Problem

Answer the following questions:

a. What is the frequency of deaths measured from 2006 through 2009?
b. What percentage of deaths occurred after 2009?
c. What is the relative frequency of deaths that occurred in 2003 or earlier?
d. What is the percentage of deaths that occurred in 2004?
e. What kind of data are the numbers of deaths?
f. The Richter scale is used to quantify the energy produced by an earthquake. Examples of Richter scale numbers are 2.3, 4.0, 6.1, and 7.0. What kind of data are these numbers?

Solution

a. 97,118 (11.8 percent)
b. 41.6 percent
c. 67,092/823,356 or 0.081 or 8.1 percent
d. 27.8 percent
e. quantitative discrete
f. quantitative continuous

TRY IT 1.19

Table 1.19 contains the total number of fatal motor vehicle traffic crashes in the United States for the period from 1994–2011.

Year	Total Number of Crashes	Year	Total Number of Crashes
1994	36,254	2004	38,444
1995	37,241	2005	39,252
1996	37,494	2006	38,648
1997	37,324	2007	37,435
1998	37,107	2008	34,172
1999	37,140	2009	30,862
2000	37,526	2010	30,296
2001	37,862	2011	29,757
2002	38,491	Total	653,782
2003	38,477		

Table 1.19

Answer the following questions:

a. What is the frequency of deaths measured from 2000 through 2004?
b. What percentage of deaths occurred after 2006?
c. What is the relative frequency of deaths that occurred in 2000 or before?
d. What is the percentage of deaths that occurred in 2011?
e. What is the cumulative relative frequency for 2006? Explain what this number tells you about the data.

1.4 Experimental Design and Ethics

Does aspirin reduce the risk of heart attacks? Is one brand of fertilizer more effective at growing roses than another? Is fatigue as dangerous to a driver as speeding? Questions like these are answered using randomized experiments. In this module, you will learn important aspects of experimental design. Proper study design ensures the production of reliable, accurate data.

The purpose of an experiment is to investigate the relationship between two variables. In an experiment, there is the **explanatory variable** which affects the **response variable**. In a randomized experiment, the researcher manipulates the explanatory variable and then observes the response variable. Each value of the explanatory variable used in an experiment is called a **treatment**.

You want to investigate the effectiveness of vitamin E in preventing disease. You recruit a group of subjects and ask them if they regularly take vitamin E. You notice that the subjects who take vitamin E exhibit better health on average than those who do not. Does this prove that vitamin E is effective in disease prevention? It does not. There are many differences between the two groups compared in addition to vitamin E consumption. People who take vitamin E regularly often take other steps to improve their health: exercise, diet, other vitamin supplements. Any one of these factors could be influencing health. As described, this study does not prove that vitamin E is the key to disease

prevention.

Additional variables that can cloud a study are called **lurking variables**. In order to prove that the explanatory variable is causing a change in the response variable, it is necessary to isolate the explanatory variable. The researcher must design her experiment in such a way that there is only one difference between groups being compared: the planned treatments. This is accomplished by the **random assignment** of experimental units to treatment groups. When subjects are assigned treatments randomly, all of the potential lurking variables are spread equally among the groups. At this point the only difference between groups is the one imposed by the researcher. Different outcomes measured in the response variable, therefore, must be a direct result of the different treatments. In this way, an experiment can prove a cause-and-effect connection between the explanatory and response variables.

Confounding occurs when the effects of multiple factors on a response cannot be separated, for instance, if a student guesses on the even-numbered questions on an exam and sits in a favorite spot on exam day. Why does the student get a high test scores on the exam? It could be the increased study time or sitting in the favorite spot or both. Confounding makes it difficult to draw valid conclusions about the effect of each factor on the outcome. The way around this is to test several outcomes with one method (treatment). This way, we know which treatment really works.

The power of suggestion can have an important influence on the outcome of an experiment. Studies have shown that the expectation of the study participant can be as important as the actual medication. In one study of performance-enhancing substances, researchers noted the following:

*Results showed that believing one had taken the substance resulted in [*performance*] times almost as fast as those associated with consuming the substance itself. In contrast, taking the substance without knowledge yielded no significant performance increment.*[1]

When participation in a study prompts a physical response from a participant, it is difficult to isolate the effects of the explanatory variable. To counter the power of suggestion, researchers set aside one treatment group as a **control group**. This group is given a **placebo** treatment, a treatment that cannot influence the response variable. The control group helps researchers balance the effects of being in an experiment with the effects of the active treatments. Of course, if you are participating in a study and you know that you are receiving a pill that contains no actual medication, then the power of suggestion is no longer a factor. **Blinding** in a randomized experiment designed to reduce bias by hiding information. When a person involved in a research study is blinded, he does not know who is receiving the active treatment(s) and who is receiving the placebo treatment. A **double-blind experiment** is one in which both the subjects and the researchers involved with the subjects are blinded.

Sometimes, it is neither possible nor ethical for researchers to conduct experimental studies. For example, if you want to investigate whether malnutrition affects elementary school performance in children, it would not be appropriate to assign an experimental group to be malnourished. In these cases, **observational studies** or **surveys** may be used. In an observational study, the researcher does not directly manipulate the independent variable. Instead, he or she takes recordings and measurements of naturally occurring phenomena. By sorting these data into control and experimental conditions, the relationship between the dependent and independent variables can be drawn. In a survey, a researcher's measurements consist of questionnaires that are answered by the research participants.

EXAMPLE 1.20

Problem

Researchers want to investigate whether taking aspirin regularly reduces the risk of a heart attack. 400 men between the ages of 50 and 84 are recruited as participants. The men are divided randomly into two groups: one group will take aspirin, and the other group will take a placebo. Each man takes one pill each day for three years, but he does not know whether he is taking aspirin or the placebo. At the end of the study, researchers count the number of men in each group who have had heart attacks.

Identify the following values for this study: population, sample, experimental units, explanatory variable, response variable, treatments.

Solution

The ***population*** is men aged 50 to 84.
The ***sample*** is the 400 men who participated.
The ***experimental units*** are the individual men in the study.

1 McClung, M. and Collins, D. (2007 June). "Because I know it will!" Placebo effects of an ergogenic aid on athletic performance. *Journal of Sport & Exercise Psychology, 29(3), 382-94.*

The ***explanatory variable*** is oral medication.
The ***treatments*** are aspirin and a placebo.
The ***response variable*** is whether a subject had a heart attack.

EXAMPLE 1.21

Problem

The Smell & Taste Treatment and Research Foundation conducted a study to investigate whether smell can affect learning. Subjects completed mazes multiple times while wearing masks. They completed the pencil and paper mazes three times wearing floral-scented masks, and three times with unscented masks. Participants were assigned at random to wear the floral mask during the first three trials or during the last three trials. For each trial, researchers recorded the time it took to complete the maze and the subject's impression of the mask's scent: positive, negative, or neutral.

a. Describe the explanatory and response variables in this study.
b. What are the treatments?
c. Identify any lurking variables that could interfere with this study.
d. Is it possible to use blinding in this study?

Solution

a. The explanatory variable is scent, and the response variable is the time it takes to complete the maze.
b. There are two treatments: a floral-scented mask and an unscented mask.
c. All subjects experienced both treatments. The order of treatments was randomly assigned so there were no differences between the treatment groups. Random assignment eliminates the problem of lurking variables.
d. Subjects will clearly know whether they can smell flowers or not, so subjects cannot be blinded in this study. Researchers timing the mazes can be blinded, though. The researcher who is observing a subject will not know which mask is being worn.

EXAMPLE 1.22

Problem

A researcher wants to study the effects of birth order on personality. Explain why this study could not be conducted as a randomized experiment. What is the main problem in a study that cannot be designed as a randomized experiment?

Solution

The explanatory variable is birth order. You cannot randomly assign a person's birth order. Random assignment eliminates the impact of lurking variables. When you cannot assign subjects to treatment groups at random, there will be differences between the groups other than the explanatory variable.

TRY IT 1.22

You are concerned about the effects of texting on driving performance. Design a study to test the response time of drivers while texting and while driving only. How many seconds does it take for a driver to respond when a leading car hits the brakes?

a. Describe the explanatory and response variables in the study.
b. What are the treatments?
c. What should you consider when selecting participants?
d. Your research partner wants to divide participants randomly into two groups: one to drive without distraction and one to text and drive simultaneously. Is this a good idea? Why or why not?
e. Identify any lurking variables that could interfere with this study.
f. How can blinding be used in this study?

Ethics

The widespread misuse and misrepresentation of statistical information often gives the field a bad name. Some say that "numbers don't lie," but the people who use numbers to support their claims often do.

A recent investigation of famous social psychologist, Diederik Stapel, has led to the retraction of his articles from some of the world's top journals including, *Journal of Experimental Social Psychology, Social Psychology, Basic and Applied Social Psychology, British Journal of Social Psychology,* and the magazine *Science*. Diederik Stapel is a former professor at Tilburg University in the Netherlands. Over the past two years, an extensive investigation involving three universities where Stapel has worked concluded that the psychologist is guilty of fraud on a colossal scale. Falsified data taints over 55 papers he authored and 10 Ph.D. dissertations that he supervised.

Stapel did not deny that his deceit was driven by ambition. But it was more complicated than that, he told me. He insisted that he loved social psychology but had been frustrated by the messiness of experimental data, which rarely led to clear conclusions. His lifelong obsession with elegance and order, he said, led him to concoct results that journals found attractive. "It was a quest for aesthetics, for beauty—instead of the truth," he said. He described his behavior as an addiction that drove him to carry out acts of increasingly daring fraud.[2]

The committee investigating Stapel concluded that he is guilty of several practices including

- creating datasets, which largely confirmed the prior expectations,
- altering data in existing datasets,
- changing measuring instruments without reporting the change, and
- misrepresenting the number of experimental subjects.

Clearly, it is never acceptable to falsify data the way this researcher did. Sometimes, however, violations of ethics are not as easy to spot.

Researchers have a responsibility to verify that proper methods are being followed. The report describing the investigation of Stapel's fraud states that, "statistical flaws frequently revealed a lack of familiarity with elementary statistics."[3] Many of Stapel's co-authors should have spotted irregularities in his data. Unfortunately, they did not know very much about statistical analysis, and they simply trusted that he was collecting and reporting data properly.

Many types of statistical fraud are difficult to spot. Some researchers simply stop collecting data once they have just enough to prove what they had hoped to prove. They don't want to take the chance that a more extensive study would complicate their lives by producing data contradicting their hypothesis.

Professional organizations, like the American Statistical Association, clearly define expectations for researchers. There are even laws in the federal code about the use of research data.

When a statistical study uses human participants, as in medical studies, both ethics and the law dictate that researchers should be mindful of the safety of their research subjects. The U.S. Department of Health and Human Services oversees federal regulations of research studies with the aim of protecting participants. When a university or other research institution engages in research, it must ensure the safety of all human subjects. For this reason, research institutions establish oversight committees known as **Institutional Review Boards (IRB)**. All planned studies must be approved in advance by the IRB. Key protections that are mandated by law include the following:

- Risks to participants must be minimized and reasonable with respect to projected benefits.
- Participants must give **informed consent**. This means that the risks of participation must be clearly explained to the subjects of the study. Subjects must consent in writing, and researchers are required to keep documentation of their consent.
- Data collected from individuals must be guarded carefully to protect their privacy.

These ideas may seem fundamental, but they can be very difficult to verify in practice. Is removing a participant's name from the data record sufficient to protect privacy? Perhaps the person's identity could be discovered from the data that remains. What happens if the study does not proceed as planned and risks arise that were not anticipated? When is informed consent really necessary? Suppose your doctor wants a blood sample to check your cholesterol level. Once the sample has been tested, you expect the lab to dispose of the remaining blood. At that point the blood becomes biological waste. Does a researcher have the right to take it for use in a study?

It is important that students of statistics take time to consider the ethical questions that arise in statistical studies. How

2 Bhattacharjee, Y. (2013, April 26). The mind of a con man. *The New York Times*. Retrieved from http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html?_r=3&src=dayp&.
3 Tillburg University. (2012, Nov. 28). Flawed science: the fraudulent research practices of social psychologist Diederik Stapel. Retrieved from https://www.tilburguniversity.edu/upload/3ff904d7-547b-40ae-85fe-bea38e05a34a_Final%20report%20Flawed%20Science.pdf.

prevalent is fraud in statistical studies? You might be surprised—and disappointed. There is a website (http://openstaxcollege.org/l/40introone) dedicated to cataloging retractions of study articles that have been proven fraudulent. A quick glance will show that the misuse of statistics is a bigger problem than most people realize.

Vigilance against fraud requires knowledge. Learning the basic theory of statistics will empower you to analyze statistical studies critically.

EXAMPLE 1.23

Problem

Describe the unethical behavior in each example and describe how it could impact the reliability of the resulting data. Explain how the problem should be corrected.

A researcher is collecting data in a community.

a. She selects a block where she is comfortable walking because she knows many of the people living on the street.
b. No one seems to be home at four houses on her route. She does not record the addresses and does not return at a later time to try to find residents at home.
c. She skips four houses on her route because she is running late for an appointment. When she gets home, she fills in the forms by selecting random answers from other residents in the neighborhood.

Solution

a. By selecting a convenient sample, the researcher is intentionally selecting a sample that could be biased. Claiming that this sample represents the community is misleading. The researcher needs to select areas in the community at random.
b. Intentionally omitting relevant data will create bias in the sample. Suppose the researcher is gathering information about jobs and child care. By ignoring people who are not home, she may be missing data from working families that are relevant to her study. She needs to make every effort to interview all members of the target sample.
c. It is never acceptable to fake data. Even though the responses she uses are *real* responses provided by other participants, the duplication is fraudulent and can create bias in the data. She needs to work diligently to interview everyone on her route.

TRY IT 1.23

Describe the unethical behavior, if any, in each example and describe how it could impact the reliability of the resulting data. Explain how the problem should be corrected.

A study is commissioned to determine the favorite brand of fruit juice among teens in California.

a. The survey is commissioned by the seller of a popular brand of apple juice.
b. There are only two types of juice included in the study: apple juice and cranberry juice.
c. Researchers allow participants to see the brand of juice as samples are poured for a taste test.
d. Twenty-five percent of participants prefer Brand X, 33 percent prefer Brand Y and 42 percent have no preference between the two brands. Brand X references the study in a commercial saying “Most teens like Brand X as much as or more than Brand Y.”

1.5 Data Collection Experiment

Stats Lab

Data Collection Experiment

Student Learning Outcomes

- The student will demonstrate the systematic sampling technique.
- The student will construct relative frequency tables.
- The student will interpret results and their differences from different data groupings.

Movie Survey

Get a class roster/list. Randomly mark a person's name, and then mark every fourth name on the list until you get 12 names. You may have to go back to the start of the list. For each name marked, record the number of movies they saw at the theater last month.

Order the Data

Complete the two relative frequency tables below using your class data.

Number of Movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0			
1			
2			
3			
4			
5			
6			
7+			

Table 1.20 Frequency of Number of Movies Viewed

Number of Movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0–1			
2–3			
4–5			
6–7+			

Table 1.21 Frequency of Number of Movies Viewed

1. Using the tables, find the percent of data that is at most two. Which table did you use and why?
2. Using the tables, find the percent of data that is at most three. Which table did you use and why?
3. Using the tables, find the percent of data that is more than two. Which table did you use and why?
4. Using the tables, find the percent of data that is more than three. Which table did you use and why?

Discussion Questions

1. Is one of the tables *more correct* than the other? Why or why not?
2. In general, how could you group the data differently? Are there any advantages to either way of grouping the data?
3. Why did you switch between tables, if you did, when answering the question above?

1.6 Sampling Experiment

Stats Lab

Sampling Experiment

Student Learning Outcomes

- The student will demonstrate the simple random, systematic, stratified, and cluster sampling techniques.
- The student will explain the details of each procedure used.

In this lab, you will be asked to pick several random samples of restaurants. In each case, describe your procedure briefly, including how you might have used the random number generator, and then list the restaurants in the sample you obtained.

Note

The following section contains restaurants stratified by city into columns and grouped horizontally by entree cost (clusters).

Restaurants Stratified by City and Entree Cost

Entree Cost	Under $10	$10 to under $15	$15 to under $20	Over $20
San Jose	El Abuelo Taq, Pasta Mia, Emma's Express, Bamboo Hut	Emperor's Guard, Creekside Inn	Agenda, Gervais, Miro's	Blake's, Eulipia, Hayes Mansion, Germania
Palo Alto	Senor Taco, Tuscan Garden, Taxi's	Ming's, P.A. Joe's, Stickney's	Scott's Seafood, Poolside Grill, Fish Market	Sundance Mine, Maddalena's, Sally's
Los Gatos	Mary's Patio, Mount Everest, Sweet Pea's, Andele Taqueria	Lindsey's, Willow Street	Toll House	Charter House, La Maison Du Cafe
Mountain View	Maharaja, New Ma's, Thai-Rific, Garden Fresh	Amber Indian, La Fiesta, Fiesta del Mar, Dawit	Austin's, Shiva's, Mazeh	Le Petit Bistro
Cupertino	Hobees, Hung Fu, Samrat, China Express	Santa Barb. Grill, Mand. Gourmet, Bombay Oven, Kathmandu West	Fontana's, Blue Pheasant	Hamasushi, Helios
Sunnyvale	Chekijababi, Taj India, Full Throttle, Tia Juana, Lemon Grass	Pacific Fresh, Charley Brown's, Cafe Cameroon, Faz, Aruba's	Lion & Compass, The Palace, Beau Sejour	
Santa Clara	Rangoli, Armadillo Willy's, Thai Pepper, Pasand	Arthur's, Katie's Cafe, Pedro's, La Galleria	Birk's, Truya Sushi, Valley Plaza	Lakeside, Mariani's

Table 1.22 Restaurants Used in Sample

A Simple Random Sample

Pick a **simple random sample** of 15 restaurants.

1. Describe your procedure.
2. Complete the table with your sample.

1. _____	6. _____	11. _____
2. _____	7. _____	12. _____
3. _____	8. _____	13. _____
4. _____	9. _____	14. _____
5. _____	10. _____	15. _____

Table 1.23

A Systematic Sample

Pick a **systematic sample** of 15 restaurants.

1. Describe your procedure.
2. Complete the table with your sample.

1. _____	6. _____	11. _____
2. _____	7. _____	12. _____
3. _____	8. _____	13. _____
4. _____	9. _____	14. _____
5. _____	10. _____	15. _____

Table 1.24

A Stratified Sample

Pick a **stratified sample**, by city, of 20 restaurants. Use 25 percent of the restaurants from each stratum. Round to the nearest whole number.

1. Describe your procedure.
2. Complete the table with your sample.

1. _____	6. _____	11. _____	16. _____
2. _____	7. _____	12. _____	17. _____
3. _____	8. _____	13. _____	18. _____
4. _____	9. _____	14. _____	19. _____
5. _____	10. _____	15. _____	20. _____

Table 1.25

A Stratified Sample

Pick a **stratified sample**, by entree cost, of 21 restaurants. Use 25 percent of the restaurants from each stratum. Round

to the nearest whole number.

1. Describe your procedure.
2. Complete the table with your sample.

1. _____	6. _____	11. _____	16. _____
2. _____	7. _____	12. _____	17. _____
3. _____	8. _____	13. _____	18. _____
4. _____	9. _____	14. _____	19. _____
5. _____	10. _____	15. _____	20. _____
			21. _____

Table 1.26

A Cluster Sample

Pick a **cluster sample** of restaurants from two cities. The number of restaurants will vary.

1. Describe your procedure.
2. Complete the table with your sample.

1. _____	6. _____	11. _____	16. _____	21. _____
2. _____	7. _____	12. _____	17. _____	22. _____
3. _____	8. _____	13. _____	18. _____	23. _____
4. _____	9. _____	14. _____	19. _____	24. _____
5. _____	10. _____	15. _____	20. _____	25. _____

Table 1.27

Key Terms

average also called mean; a number that describes the central tendency of the data
blinding not telling participants which treatment a subject is receiving
categorical variable variables that take on values that are names or labels
cluster sampling a method for selecting a random sample and dividing the population into groups (clusters); use simple random sampling to select a set of clusters; every individual in the chosen clusters is included in the sample
continuous random variable a random variable (RV) whose outcomes are measured; the height of trees in the forest is a continuous RV
control group a group in a randomized experiment that receives an inactive treatment but is otherwise managed exactly as the other groups
convenience sampling a nonrandom method of selecting a sample; this method selects individuals that are easily accessible and may result in biased data
cumulative relative frequency the term applies to an ordered set of observations from smallest to largest. The cumulative relative frequency is the sum of the relative frequencies for all values that are less than or equal to the given value
data a set of observations (a set of possible outcomes); most data can be put into two groups: **qualitative** (an attribute whose value is indicated by a label) or **quantitative** (an attribute whose value is indicated by a number) Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (such as the number of students of a given ethnic group in a class or the number of books on a shelf). Data is continuous if it is the result of measuring (such as distance traveled or weight of luggage)
discrete random variable a random variable (RV) whose outcomes are counted
double-blinding the act of blinding both the subjects of an experiment and the researchers who work with the subjects
experimental unit any individual or object to be measured
explanatory variable the independent variable in an experiment; the value controlled by researchers
frequency the number of times a value of the data occurs
informed consent any human subject in a research study must be cognizant of any risks or costs associated with the study; the subject has the right to know the nature of the treatments included in the study, their potential risks, and their potential benefits; consent must be given freely by an informed, fit participant
institutional review board a committee tasked with oversight of research programs that involve human subjects
lurking variable a variable that has an effect on a study even though it is neither an explanatory variable nor a response variable
mathematical models a description of a phenomenon using mathematical concepts, such as equations, inequalities, distributions, etc.
nonsampling error an issue that affects the reliability of sampling data other than natural variation; it includes a variety of human errors including poor study design, biased sampling methods, inaccurate information provided by study participants, data entry errors, and poor analysis
numerical Variable variables that take on values that are indicated by numbers
observational study a study in which the independent variable is not manipulated by the researcher
parameter a number that is used to represent a population characteristic and that generally cannot be determined easily
placebo an inactive treatment that has no real effect on the explanatory variable
population all individuals, objects, or measurements whose properties are being studied
probability a number between zero and one, inclusive, that gives the likelihood that a specific event will occur
proportion the number of successes divided by the total number in the sample
qualitative data see data
quantitative data see data
random assignment the act of organizing experimental units into treatment groups using random methods
random sampling a method of selecting a sample that gives every member of the population an equal chance of being selected
relative frequency the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes to the total number of outcomes
reliability the consistency of a measure; a measure is reliable when the same results are produced given the same circumstances
representative sample a subset of the population that has the same characteristics as the population
response variable the dependent variable in an experiment; the value that is measured for change at the end of an experiment
sample a subset of the population studied

sampling bias not all members of the population are equally likely to be selected
sampling error the natural variation that results from selecting a sample to represent a larger population; this variation decreases as the sample size increases, so selecting larger samples reduces sampling error
sampling with replacement once a member of the population is selected for inclusion in a sample, that member is returned to the population for the selection of the next individual
sampling without replacement a member of the population may be chosen for inclusion in a sample only once; if chosen, the member is not returned to the population before the next selection
simple random sampling a straightforward method for selecting a random sample; give each member of the population a number
Use a random number generator to select a set of labels. These randomly selected labels identify the members of your sample
statistic a numerical characteristic of the sample; a statistic estimates the corresponding population parameter
statistical models a description of a phenomenon using probability distributions that describe the expected behavior of the phenomenon and the variability in the expected observations
stratified sampling a method for selecting a random sample used to ensure that subgroups of the population are represented adequately; divide the population into groups (strata). Use simple random sampling to identify a proportionate number of individuals from each stratum
survey a study in which data is collected as reported by individuals.
systematic sampling a method for selecting a random sample; list the members of the population
Use simple random sampling to select a starting point in the population. Let k = (number of individuals in the population)/(number of individuals needed in the sample). Choose every kth individual in the list starting with the one that was randomly selected. If necessary, return to the beginning of the population list to complete your sample
treatments different values or components of the explanatory variable applied in an experiment
validity refers to how much a measure or conclusion accurately reflects real world
variable a characteristic of interest for each person or object in a population

Chapter Review

1.1 Definitions of Statistics, Probability, and Key Terms

The mathematical theory of statistics is easier to learn when you know the language. This module presents important terms that will be used throughout the text.

1.2 Data, Sampling, and Variation in Data and Sampling

Data are individual items of information that come from a population or sample. Data may be classified as qualitative (categorical), quantitative continuous, or quantitative discrete.

Because it is not practical to measure the entire population in a study, researchers use samples to represent the population. A random sample is a representative group from the population chosen by using a method that gives each individual in the population an equal chance of being included in the sample. Random sampling methods include simple random sampling, stratified sampling, cluster sampling, and systematic sampling. Convenience sampling is a nonrandom method of choosing a sample that often produces biased data.

Samples that contain different individuals result in different data. This is true even when the samples are well-chosen and representative of the population. When properly selected, larger samples model the population more closely than smaller samples. There are many different potential problems that can affect the reliability of a sample. Statistical data needs to be critically analyzed, not simply accepted.

1.3 Frequency, Frequency Tables, and Levels of Measurement

Some calculations generate numbers that are artificially precise. It is not necessary to report a value to eight decimal places when the measures that generated that value were only accurate to the nearest tenth. Round your final answer to one more decimal place than was present in the original data. This means that if you have data measured to the nearest tenth of a unit, report the final statistic to the nearest hundredth. Expect that some of your answers will vary from the text due to rounding errors.

In addition to rounding your answers, you can measure your data using the following four levels of measurement:

- **Nominal scale level** data that cannot be ordered nor can it be used in calculations
- **Ordinal scale level** data that can be ordered; the differences cannot be measured
- **Interval scale level** data with a definite ordering but no starting point; the differences can be measured, but there is no such thing as a ratio

- **Ratio scale level** data with a starting point that can be ordered; the differences have meaning and ratios can be calculated

When organizing data, it is important to know how many times a value appears. How many statistics students study five hours or more for an exam? What percent of families on our block own two pets? Frequency, relative frequency, and cumulative relative frequency are measures that answer questions like these.

1.4 Experimental Design and Ethics

A poorly designed study will not produce reliable data. There are certain key components that must be included in every experiment. To eliminate lurking variables, subjects must be assigned randomly to different treatment groups. One of the groups must act as a control group, demonstrating what happens when the active treatment is not applied. Participants in the control group receive a placebo treatment that looks exactly like the active treatments but cannot influence the response variable. To preserve the integrity of the placebo, both researchers and subjects may be blinded. When a study is designed properly, the only difference between treatment groups is the one imposed by the researcher. Therefore, when groups respond differently to different treatments, the difference must be due to the influence of the explanatory variable.

"An ethics problem arises when you are considering an action that benefits you or some cause you support, hurts or reduces benefits to others, and violates some rule."[4] Ethical violations in statistics are not always easy to spot. Professional associations and federal agencies post guidelines for proper conduct. It is important that you learn basic statistical procedures so that you can recognize proper data analysis.

Practice

1.1 Definitions of Statistics, Probability, and Key Terms

1. Below is a two-way table showing the types of college sports played by men and women.

	Soccer	Basketball	Lacrosse	Total
Women	8	8	4	20
Men	4	12	4	20
Total	12	20	8	40

Table 1.28

Given these data, calculate the marginal distributions of college sports for the people surveyed.

2. Below is a two-way table showing the types of college sports played by men and women.

	Soccer	Basketball	Lacrosse	Total
Women	8	8	4	20
Men	4	12	4	20
Total	12	20	8	40

Table 1.29

Given these data, calculate the conditional distributions for the subpopulation of women who play college sports.

4 Gelman, A. (2013, May 1). Open data and open methods. *Ethics and Statistics*. Retrieved from http://www.stat.columbia.edu/~gelman/research/published/ChanceEthics1.pdf.

Use the following information to answer the next five exercises. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new viral antibody drug is currently under study. It is given to patients once the virus's symptoms have revealed themselves. Of interest is the average (mean) length of time in months patients live once they start the treatment. Two researchers each follow a different set of 40 patients with the viral disease from the start of treatment until their deaths. The following data (in months) are collected.

Researcher A
3; 4; 11; 15; 16; 17; 22; 44; 37; 16; 14; 24; 25; 15; 26; 27; 33; 29; 35; 44; 13; 21; 22; 10; 12; 8; 40; 32; 26; 27; 31; 34; 29; 17; 8; 24; 18; 47; 33; 34

Researcher B
3; 14; 11; 5; 16; 17; 28; 41; 31; 18; 14; 14; 26; 25; 21; 22; 31; 2; 35; 44; 23; 21; 21; 16; 12; 18; 41; 22; 16; 25; 33; 34; 29; 13; 18; 24; 23; 42; 33; 29

Determine what the key terms refer to in the example for Researcher A.

3. population

4. sample

5. parameter

6. statistic

7. variable

1.2 Data, Sampling, and Variation in Data and Sampling

8. *Number of times per week* is what type of data?
a. qualitative (categorical); b. quantitative discrete; c. quantitative continuous

Use the following information to answer the next four exercises: A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in San Antonio, Texas. The first house in the neighborhood around the park was selected randomly, and then the resident of every eighth house in the neighborhood around the park was interviewed.

9. The sampling method was
a. simple random; b. systematic; c. stratified; d. cluster

10. *Duration (amount of time)* is what type of data?
a. qualitative (categorical); b. quantitative discrete; c. quantitative continuous

11. The colors of the houses around the park are what kind of data?
a. qualitative (categorical); b. quantitative discrete; c. quantitative continuous

12. The population is _____.

13. Table 1.30 contains the total number of deaths worldwide as a result of earthquakes from 2000–2012.

Year	Total Number of Deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003

Table 1.30

Year	Total Number of Deaths
2006	6,605
2007	712
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	**823,856**

Table 1.30

Use Table 1.30 to answer the following questions.

a. What is the proportion of deaths between 2007–2012?
b. What percent of deaths occurred before 2001?
c. What is the percent of deaths that occurred in 2003 or after 2010?
d. What is the fraction of deaths that happened before 2012?
e. What kind of data is the number of deaths?
f. Earthquakes are quantified according to the amount of energy they produce (examples are 2.1, 5.0, 6.7). What type of data is that?
g. What contributed to the large number of deaths in 2010? In 2004? Explain.
h. If you were asked to present these data in an oral presentation, what type of graph would you choose to present and why? Explain what features you would point out on the graph during your presentation.

For the following four exercises, determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

14. A group of test subjects is divided into twelve groups; then four of the groups are chosen at random.

15. A market researcher polls every tenth person who walks into a store.

16. The first 50 people who walk into a sporting event are polled on their television preferences.

17. A computer generates 100 random numbers, and 100 people whose names correspond with the numbers on the list are chosen.

Use the following information to answer the next seven exercises: Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new viral antibody drug is currently under study. It is given to patients once the virus's symptoms have revealed themselves. Of interest is the average (mean) length of time in months patients live once starting the treatment. Two researchers each follow a different set of 40 patients with the viral disease from the start of treatment until their deaths. The following data (in months) are collected:

Researcher A: 3; 4; 11; 15; 16; 17; 22; 44; 37; 16; 14; 24; 25; 15; 26; 27; 33; 29; 35; 44; 13; 21; 22; 10; 12; 8; 40; 32; 26; 27; 31; 34; 29; 17; 8; 24; 18; 47; 33; 34

Researcher B: 3; 14; 11; 5; 16; 17; 28; 41; 31; 18; 14; 14; 26; 25; 21; 22; 31; 2; 35; 44; 23; 21; 21; 16; 12; 18; 41; 22; 16; 25; 33; 34; 29; 13; 18; 24; 23; 42; 33; 29

18. Complete the tables using the data provided.

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Relative Frequency
.5–6.5			
6.5–12.5			
12.5–18.5			
18.5–24.5			
24.5–30.5			
30.5–36.5			
36.5–42.5			
42.5–48.5			

Table 1.31 Researcher A

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Relative Frequency
.5–6.5			
6.5–12.5			
12.5–18.5			
18.5–24.5			
24.5–30.5			
30.5–36.5			
36.5-45.5			

Table 1.32 Researcher B

19. Determine what the key term data refers to in the above example for Researcher A.

20. List two reasons why the data may differ.

21. Can you tell if one researcher is correct and the other one is incorrect? Why?

22. Would you expect the data to be identical? Why or why not?

23. Suggest at least two methods the researchers might use to gather random data.

24. Suppose that the first researcher conducted his survey by randomly choosing one state in the nation and then randomly picking 40 patients from that state. What sampling method would that researcher have used?

25. Suppose that the second researcher conducted his survey by choosing 40 patients he knew. What sampling method would that researcher have used? What concerns would you have about this data set, based upon the data collection method?

Use the following data to answer the next five exercises: Two researchers are gathering data on hours of video games played by school-aged children and young adults. They each randomly sample different groups of 150 students from the same school. They collect the following data:

Hours Played per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0–2	26	.17	.17
2–4	30	.20	.37
4–6	49	.33	.70
6–8	25	.17	.87
8–10	12	.08	.95
10–12	8	.05	1

Table 1.33 Researcher A

Hours Played per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0–2	48	.32	.32
2–4	51	.34	.66
4–6	24	.16	.82
6–8	12	.08	.90
8–10	11	.07	.97
10–12	4	.03	1

Table 1.34 Researcher B

26. Give a reason why the data may differ.

27. Would the sample size be large enough if the population is the students in the school?

28. Would the sample size be large enough if the population is school-aged children and young adults in the United States?

29. Researcher A concludes that most students play video games between four and six hours each week. Researcher B concludes that most students play video games between two and four hours each week. Who is correct?

30. Suppose you were asked to present the data from researchers A and B in an oral presentation. When would a pie graph be appropriate? When would a bar graph more desirable? Explain which features you would point out on each type of graph and what potential display problems you would try to avoid.

31. As part of a way to reward students for participating in the survey, the researchers gave each student a gift card to a video game store. Would this affect the data if students knew about the award before the study?

Use the following data to answer the next five exercises: A pair of studies was performed to measure the effectiveness of a new software program designed to help stroke patients regain their problem-solving skills. Patients were asked to use the software program twice a day, once in the morning, and once in the evening. The studies observed 200 stroke patients recovering over a period of several weeks. The first study collected the data in Table 1.35. The second study collected the data in Table 1.36.

Group	Showed Improvement	No Improvement	Deterioration
Used program	142	43	15
Did not use program	72	110	18

Table 1.35

Group	Showed Improvement	No Improvement	Deterioration
Used program	105	74	19
Did not use program	89	99	12

Table 1.36

32. Given what you know, which study is correct?

33. The first study was performed by the company that designed the software program. The second study was performed by the American Medical Association. Which study is more reliable?

34. Both groups that performed the study concluded that the software works. Is this accurate?

35. The company takes the two studies as proof that their software causes mental improvement in stroke patients. Is this a fair statement?

36. Patients who used the software were also a part of an exercise program whereas patients who did not use the software were not. Does this change the validity of the conclusions from Exercise 1.34?

37. Is a sample size of 1,000 a reliable measure for a population of 5,000?

38. Is a sample of 500 volunteers a reliable measure for a population of 2,500?

39. A question on a survey reads: "Do you prefer the delicious taste of Brand X or the taste of Brand Y?" Is this a fair question?

40. Is a sample size of two representative of a population of five?

41. Is it possible for two experiments to be well run with similar sample sizes to get different data?

1.3 Frequency, Frequency Tables, and Levels of Measurement

42. What type of measure scale is being used? Nominal, ordinal, interval or ratio.

a. High school soccer players classified by their athletic ability: superior, average, above average
b. Baking temperatures for various main dishes: 350, 400, 325, 250, 300
c. The colors of crayons in a 24-crayon box
d. Social security numbers
e. Incomes measured in dollars
f. A satisfaction survey of a social website by number: 1 = very satisfied, 2 = somewhat satisfied, 3 = not satisfied
g. Preferred TV shows: comedy, drama, science fiction, sports, news
h. Time of day on an analog watch
i. The distance in miles to the closest grocery store
j. The dates 1066, 1492, 1644, 1947, and 1944
k. The heights of 21–65-year-old women
l. Common letter grades: A, B, C, D, and F

1.4 Experimental Design and Ethics

43. Design an experiment. Identify the explanatory and response variables. Describe the population being studied and the experimental units. Explain the treatments that will be used and how they will be assigned to the experimental units. Describe how blinding and placebos may be used to counter the power of suggestion.

44. Discuss potential violations of the rule requiring informed consent.
   a. Inmates in a correctional facility are offered good behavior credit in return for participation in a study.
   b. A research study is designed to investigate a new children's allergy medication.
   c. Participants in a study are told that the new medication being tested is highly promising, but they are not told that only a small portion of participants will receive the new medication. Others will receive placebo treatments and traditional treatments.

Homework

1.1 Definitions of Statistics, Probability, and Key Terms

45. For each of the following situations, indicate whether it would be best modeled with a mathematical model or a statistical model. Explain your answers.
   a. driving time from New York to Florida
   b. departure time of a commuter train at rush hour
   c. distance from your house to school
   d. temperature of a refrigerator at any given time
   e. weight of a bag of rice at the store

For each of the following eight exercises, identify: a. the population, b. the sample, c. the parameter, d. the statistic, e. the variable, and f. the data. Give examples where appropriate.

46. A fitness center is interested in the mean amount of time a client exercises in the center each week.

47. Ski resorts are interested in the mean age that children take their first ski and snowboard lessons. They need this information to plan their ski classes optimally.

48. A cardiologist is interested in the mean recovery period of her patients who have had heart attacks.

49. Insurance companies are interested in the mean health costs each year of their clients, so that they can determine the costs of health insurance.

50. A politician is interested in the proportion of voters in his district who think he is doing a new good job.

51. A marriage counselor is interested in the proportion of clients she counsels who stay married.

52. Political pollsters may be interested in the proportion of people who will vote for a particular cause.

53. A marketing company is interested in the proportion of people who will buy a particular product.

Use the following information to answer the next three exercises: A Lake Tahoe Community College instructor is interested in the mean number of days Lake Tahoe Community College math students are absent from class during a quarter.

54. What is the population she is interested in?
   a. all Lake Tahoe Community College students
   b. all Lake Tahoe Community College English students
   c. all Lake Tahoe Community College students in her classes
   d. all Lake Tahoe Community College math students

55. Consider the following
   X = number of days a Lake Tahoe Community College math student is absent.
   In this case, X is an example of which of the following?
   a. variable
   b. population
   c. statistic
   d. data

56. The instructor's sample produces a mean number of days absent of 3.5 days. This value is an example of which of the following?
   a. parameter
   b. data
   c. statistic
   d. variable

1.2 Data, Sampling, and Variation in Data and Sampling

For the following exercises, identify the type of data that would be used to describe a response (quantitative discrete, quantitative continuous, or qualitative), and give an example of the data.

57. number of tickets sold to a concert

58. percent of body fat

59. favorite baseball team

60. time in line to buy groceries

61. number of students enrolled at Evergreen Valley College

62. most-watched television show

63. brand of toothpaste

64. distance to the closest movie theatre

65. age of executives in Fortune 500 companies

66. number of competing computer spreadsheet software packages

Use the following information to answer the next two exercises: A study was done to determine the age, number of times per week, and the duration (amount of time) of resident use of a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every 8th house in the neighborhood around the park was interviewed.

67. *Number of times per week* is what type of data?
   a. qualitative
   b. quantitative discrete
   c. quantitative continuous

68. *Duration (amount of time)* is what type of data?
   a. qualitative
   b. quantitative discrete
   c. quantitative continuous

69. Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. Suppose an airline conducts a survey. Over Thanksgiving weekend, it surveys six flights from Boston to Salt Lake City to determine the number of babies on the flights. It determines the amount of safety equipment needed by the result of that study.
   a. Using complete sentences, list three things wrong with the way the survey was conducted.
   b. Using complete sentences, list three ways that you would improve the survey if it were to be repeated.

70. Suppose you want to determine the mean number of students per statistics class in your state. Describe a possible sampling method in three to five complete sentences. Make the description detailed.

71. Suppose you want to determine the mean number of cans of soda drunk each month by students in their twenties at your school. Describe a possible sampling method in three to five complete sentences. Make the description detailed.

72. List some practical difficulties involved in getting accurate results from a telephone survey.

73. List some practical difficulties involved in getting accurate results from a mailed survey.

74. With your classmates, brainstorm some ways you could overcome these problems if you needed to conduct a phone or mail survey.

75. The instructor takes her sample by gathering data on five randomly selected students from each Lake Tahoe Community College math class. The type of sampling she used is which of the following?
   a. cluster sampling
   b. stratified sampling
   c. simple random sampling
   d. convenience sampling

76. A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every eighth house in the neighborhood around the park was interviewed. The sampling method was which of the following?
   a. simple random
   b. systematic
   c. stratified
   d. cluster

77. Name the sampling method used in each of the following situations:
   a. A woman in the airport is handing out questionnaires to travelers asking them to evaluate the airport's service. She does not ask travelers who are hurrying through the airport with their hands full of luggage, but instead asks all travelers who are sitting near gates and not taking naps while they wait.
   b. A teacher wants to know if her students are doing homework, so she randomly selects rows two and five and then calls on all students in row two and all students in row five to present the solutions to homework problems to the class.
   c. The marketing manager for an electronics chain store wants information about the ages of its customers. Over the next two weeks, at each store location, 100 randomly selected customers are given questionnaires to fill out asking for information about age, as well as about other variables of interest.
   d. The librarian at a public library wants to determine what proportion of the library users are children. The librarian has a tally sheet on which she marks whether books are checked out by an adult or a child. She records this data for every fourth patron who checks out books.
   e. A political party wants to know the reaction of voters to a debate between the candidates. The day after the debate, the party's polling staff calls 1,200 randomly selected phone numbers. If a registered voter answers the phone or is available to come to the phone, that registered voter is asked whom he or she intends to vote for and whether the debate changed his or her opinion of the candidates.

78. A *random survey* was conducted of 3,274 people of the *microprocessor generation*—people born since 1971, the year the microprocessor was invented. It was reported that 48 percent of those individuals surveyed stated that if they had $2,000 to spend, they would use it for computer equipment. Also, 66 percent of those surveyed considered themselves relatively savvy computer users.
   a. Do you consider the sample size large enough for a study of this type? Why or why not?
   b. Based on your *gut feeling*, do you believe the percents accurately reflect the U.S. population for those individuals born since 1971? If not, do you think the percents of the population are actually higher or lower than the sample statistics? Why?
      Additional information: The survey, reported by Intel Corporation, was filled out by individuals who visited the Los Angeles Convention Center to see the Smithsonian Institute's road show called "America's Smithsonian."
   c. With this additional information, do you feel that all demographic and ethnic groups were equally represented at the event? Why or why not?
   d. With the additional information, comment on how accurately you think the sample statistics reflect the population parameters.

79. The Well-Being Index is a survey that follows trends of U.S. residents on a regular basis. There are six areas of health and wellness covered in the survey: Life Evaluation, Emotional Health, Physical Health, Healthy Behavior, Work Environment, and Basic Access. Some of the questions used to measure the Index are listed below. Identify the type of data obtained from each question used in this survey: qualitative, quantitative discrete, or quantitative continuous.
   a. Do you have any health problems that prevent you from doing any of the things people your age can normally do?
   b. During the past 30 days, for about how many days did poor health keep you from doing your usual activities?
   c. In the last seven days, on how many days did you exercise for 30 minutes or more?

d. Do you have health insurance coverage?

80. In advance of the 1936 presidential election, a magazine released the results of an opinion poll predicting that the republican candidate Alf Landon would win by a large margin. The magazine sent post cards to approximately 10,000,000 prospective voters. These prospective voters were selected from the subscription list of the magazine, from automobile registration lists, from phone lists, and from club membership lists. Approximately 2,300,000 people returned the postcards.
   a. Think about the state of the United States in 1936. Explain why a sample chosen from magazine subscription lists, automobile registration lists, phone books, and club membership lists was not representative of the population of the United States at that time.
   b. What effect does the low response rate have on the reliability of the sample?
   c. Are these problems examples of sampling error or nonsampling error?
   d. During the same year, another pollster conducted a poll of 30,000 prospective voters. These researchers used a method they called *quota sampling* to obtain survey answers from specific subsets of the population. Quota sampling is an example of which sampling method described in this module?

81. Crime-related and demographic statistics for 47 US states in 1960 were collected from government agencies, including the FBI's *Uniform Crime Report*. One analysis of this data found a strong connection between education and crime indicating that higher levels of education in a community correspond to higher crime rates. Which of the potential problems with samples discussed in Data, Sampling, and Variation in Data and Sampling could explain this connection?

82. A website that allows anyone to create and respond to polls had a question posted on April 15 which asked: "Do you feel happy paying your taxes when members of the Obama administration are allowed to ignore their tax liabilities?"[5]
As of April 25, 11 people responded to this question. Each participant answered "NO!"
Which of the potential problems with samples discussed in this module could explain this connection?

83. A scholarly article about response rates begins with the following quote:
"Declining contact and cooperation rates in random digit dial (RDD) national telephone surveys raise serious concerns about the validity of estimates drawn from such research."[6]
The Pew Research Center for People and the Press admits
"The percentage of people we interview—out of all we try to interview—has been declining over the past decade or more."[7]
   a. What are some reasons for the decline in response rate over the past decade?
   b. Explain why researchers are concerned with the impact of the declining response rate on public opinion polls.

1.3 Frequency, Frequency Tables, and Levels of Measurement

84. Fifty part-time students were asked how many courses they were taking this term. The (incomplete) results are shown below.

# of Courses	Frequency	Relative Frequency	Cumulative Relative Frequency
1	30	.6	
2	15		
3			

Table 1.37 Part-time Student Course Loads

   a. Fill in the blanks in Table 1.37.
   b. What percent of students take exactly two courses?

5 lastbaldeagle. Retrieved from http://www.youpolls.com/details.aspx?id=12328.
6 Keeter, S., et al. (2006). Gauging the impact of growing nonresponse on estimates from a national RDD telephone survey. *Public Opinion Quarterly, 70*(5). Retrieved from http://hbanaszak.mjr.uw.edu.pl/TempTxt/Links/GAUGING%20THE%20IMPACT%20OF%20GROWING.pdf.
7 Pew Research Center. (n.d.). Frequently asked questions. Retrieved from http://www.pewresearch.org/methodology/u-s-survey-research/frequently-asked-questions/#dont-you-have-trouble-getting-people-to-answer-your-polls.

c. What percent of students take one or two courses?

85. Sixty adults with gum disease were asked the number of times per week they used to floss before their diagnosis. The (incomplete) results are shown in Table 1.38.

# Flossing per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0	27	.4500	
1	18		
3			.9333
6	3	.0500	
7	1	.0167	

Table 1.38 Flossing Frequency for Adults with Gum Disease

a. Fill in the blanks in Table 1.38.
b. What percent of adults flossed six times per week?
c. What percent flossed at most three times per week?

86. Nineteen immigrants to the United States were asked how many years, to the nearest year, they have lived in the United States The data are as follows: 2, 5, 7, 2, 2, 10, 20, 15, 0, 7, 0, 20, 5, 12, 15, 12, 4, 5, 10. Table 1.39 was produced.

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
0	2	$\frac{2}{19}$	.1053
2	3	$\frac{3}{19}$	.2632
4	1	$\frac{1}{19}$	.3158
5	3	$\frac{3}{19}$	.4737
7	2	$\frac{2}{19}$	.5789
10	2	$\frac{2}{19}$	.6842
12	2	$\frac{2}{19}$	.7895
15	1	$\frac{1}{19}$	.8421
20	1	$\frac{1}{19}$	1.0000

Table 1.39 Frequency of Immigrant Survey Responses

a. Fix the errors in Table 1.39. Also, explain how someone might have arrived at the incorrect number(s).
b. Explain what is wrong with this statement: “47 percent of the people surveyed have lived in the United States for 5 years.”
c. Fix the statement in **b** to make it correct.

d. What fraction of the people surveyed have lived in the United States five or seven years?
e. What fraction of the people surveyed have lived in the United States at most 12 years?
f. What fraction of the people surveyed have lived in the United States fewer than 12 years?
g. What fraction of the people surveyed have lived in the United States from five to 20 years, inclusive?

87. How much time does it take to travel to work? Table 1.40 shows the mean commute time by state for workers at least 16 years old who are not working at home. Find the mean travel time, and round off the answer properly.

24.0	24.3	25.9	18.9	27.5	17.9	21.8	20.9	16.7	27.3
18.2	24.7	20.0	22.6	23.9	18.0	31.4	22.3	24.0	25.5
24.7	24.6	28.1	24.9	22.6	23.6	23.4	25.7	24.8	25.5
21.2	25.7	23.1	23.0	23.9	26.0	16.3	23.1	21.4	21.5
27.0	27.0	18.6	31.7	23.3	30.1	22.9	23.3	21.7	18.6

Table 1.40

88. A business magazine published data on the best small firms in 2012. These were firms which had been publicly traded for at least a year, have a stock price of at least $5 per share, and have reported annual revenue between $5 million and $1 billion. Table 1.41 shows the ages of the chief executive officers for the first 60 ranked firms.

Age	Frequency	Relative Frequency	Cumulative Relative Frequency
40–44	3		
45–49	11		
50–54	13		
55–59	16		
60–64	10		
65–69	6		
70–74	1		

Table 1.41

a. What is the frequency for CEO ages between 54 and 65?
b. What percentage of CEOs are 65 years or older?
c. What is the relative frequency of ages under 50?
d. What is the cumulative relative frequency for CEOs younger than 55?
e. Which graph shows the relative frequency and which shows the cumulative relative frequency?

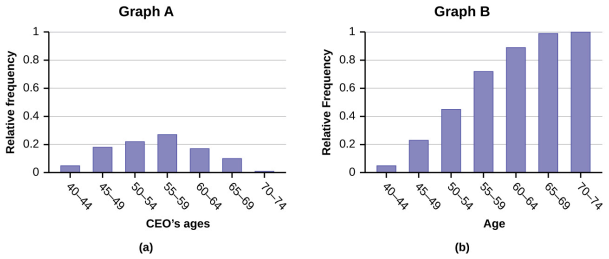


Figure 1.13

Use the following information to answer the next two exercises: Table 1.42 contains data on hurricanes that have made direct hits on the United States. Between 1851-2004. A hurricane is given a strength category rating based on the minimum wind speed generated by the storm.

Category	Number of Direct Hits	Relative Frequency	Cumulative Frequency
1	109	.3993	.3993
2	72	.2637	.6630
3	71	.2601	
4	18		.9890
5	3	.0110	1.0000
	Total = 273		

Table 1.42 Frequency of Hurricane Direct Hits

89. What is the relative frequency of direct hits that were category 4 hurricanes?
a. .0768
b. .0659
c. .2601
d. not enough information to calculate

90. What is the relative frequency of direct hits that were AT MOST a category 3 storm?
a. .3480
b. .9231
c. .2601
d. .3370

1.4 Experimental Design and Ethics

91. How does sleep deprivation affect your ability to drive? A recent study measured the effects on 19 professional drivers. Each driver participated in two experimental sessions: one after normal sleep and one after 27 hours of total sleep deprivation. The treatments were assigned in random order. In each session, performance was measured on a variety of tasks including a driving simulation.
Use key terms from this module to describe the design of this experiment.

92. An advertisement for Acme Investments displays the two graphs in Figure 1.14 to show the value of Acme's product in comparison with the Other Guy's product. Describe the potentially misleading visual effect of these comparison graphs. How can this be corrected?

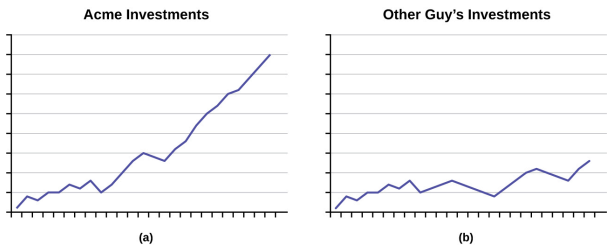


Figure 1.14 As the graphs show, Acme consistently outperforms the Other Guys!

93. The graph in Figure 1.15 shows the number of complaints for six different airlines as reported to the U.S. Department of Transportation in February 2013. Alaska, Pinnacle, and Airtran Airlines have far fewer complaints reported than American, Delta, and United. Can we conclude that American, Delta, and United are the worst airline carriers since they have the most complaints?

Figure 1.15

94. An epidemiologist is studying the spread of the common cold among college students. He is interested in how the temperature of the dorm room correlates with the incidence of new infections. How can he design an observational study to answer this question? If he chooses to use surveys in his measurements, what type of questions should he include in the survey?

Bringing It Together: Homework

95. Seven hundred and seventy-one distance learning students at Long Beach City College responded to surveys in the 2010–11 academic year. Highlights of the summary report are listed in Table 1.43.

Have computer at home	96%
Unable to come to campus for classes	65%
Age 41 or over	24%
Would like LBCC to offer more DL courses	95%
Took DL classes due to a disability	17%

Table 1.43 LBCC Distance Learning Survey Results

Live at least 16 miles from campus	13%
Took DL courses to fulfill transfer requirements	71%

Table 1.43 LBCC Distance Learning Survey Results

a. What percent of the students surveyed do not have a computer at home?
b. About how many students in the survey live at least 16 miles from campus?
c. If the same survey were done at Great Basin College in Elko, Nevada, do you think the percentages would be the same? Why?

96. Several online textbook retailers advertise that they have lower prices than on-campus bookstores. However, an important factor is whether the Internet retailers actually have the textbooks that students need in stock. Students need to be able to get textbooks promptly at the beginning of the college term. If the book is not available, then a student would not be able to get the textbook at all, or might get a delayed delivery if the book is back ordered.

A college newspaper reporter is investigating textbook availability at online retailers. He decides to investigate one textbook for each of the following seven subjects: calculus, biology, chemistry, physics, statistics, geology, and general engineering. He consults textbook industry sales data and selects the most popular nationally used textbook in each of these subjects. He visits websites for a random sample of major online textbook sellers and looks up each of these seven textbooks to see if they are available in stock for quick delivery through these retailers. Based on his investigation, he writes an article in which he draws conclusions about the overall availability of all college textbooks through online textbook retailers.
Write an analysis of his study that addresses the following issues: Is his sample representative of the population of all college textbooks? Explain why or why not. Describe some possible sources of bias in this study, and how it might affect the results of the study. Give some suggestions about what could be done to improve the study.

References

1.1 Definitions of Statistics, Probability, and Key Terms

The Data and Story Library. Retrieved from http://lib.stat.cmu.edu/DASL/Stories/CrashTestDummies.html.

1.2 Data, Sampling, and Variation in Data and Sampling

Gallup. Retrieved from http://www.well-beingindex.com/.

Gallup. Retrieved from http://www.gallup.com/poll/110548/gallup-presidential-election-trialheat-trends-19362004.aspx#4.

Gallup. Retrieved from http://www.gallup.com/175196/gallup-healthways-index-methodology.aspx.

Data from http://www.bookofodds.com/Relationships-Society/Articles/A0374-How-George-Gallup-Picked-the-President.

LBCC Distance Learning (DL) Program. Retrieved from http://de.lbcc.edu/reports/2010-11/future/highlights.html#focus.

Lusinchi, D. (2012). "President" Landon and the 1936 Literary Digest poll: Were automobile and telephone owners to blame? Social Science History 36(1), 23-54. Retrieved from https://muse.jhu.edu/article/471582/pdf.

The Data and Story Library. Retrieved from http://lib.stat.cmu.edu/DASL/Datafiles/USCrime.html.

The Mercury News. Retrieved from http://www.mercurynews.com/.

Virtual Laboratories in Probability and Statistics. Retrieved from http://www.math.uah.edu/stat/data/LiteraryDigest.html.

The Mercury News. Retrieved from http://www.mercurynews.com/.

1.3 Frequency, Frequency Tables, and Levels of Measurement

Levels of Measurement. Retrieved from http://cnx.org/content/m10809/latest/.

National Hurricane Center. Retrieved from http://www.nhc.noaa.gov/gifs/table5.gif.

ThoughtCo. Retrieved from https://www.thoughtco.com/levels-of-measurement-in-statistics-3126349.

U.S. Census Bureau. Retrieved from https://www.census.gov/quickfacts/table/PST045216/00.

Levels of measurement. Retrieved from https://www.cos.edu/Faculty/georgew/Tutorial/Data_Levels.htm.

1.4 Experimental Design and Ethics

Econoclass.com. Retrieved from http://www.econoclass.com/misleadingstats.html.

Bloomberg Businessweek. Retrieved from www.businessweek.com.

Ethics in statistics. Retrieved from http://cnx.org/content/m15555/latest/.

Forbes. Retrieved from www.forbes.com.

Forbes. http://www.forbes.com/best-small-companies/list/.

Harvard School of Public Health. Retrieved from https://www.hsph.harvard.edu/nutritionsource/vitamin-e/.

Jacskon, M.L., et al. (2013). Cognitive components of simulated driving performance: Sleep loss effect and predictors. Accident Analysis and Prevention Journal 50, 438-44. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22721550.

International Business Times. Retrieved from http://www.ibtimes.com/daily-dose-aspirin-helps-reduce-heart-attacks-study-300443.

National Highway Traffic Safety Administration. Retrieved from http://www-fars.nhtsa.dot.gov/Main/index.aspx.

Athleteinme.com. Retrieved from http://www.athleteinme.com/ArticleView.aspx?id=1053.

The Data and Story Library. Retrieved from http://lib.stat.cmu.edu/DASL/Stories/ScentsandLearning.html.

U.S. Department of Health and Human Services. Retrieved from https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html.

U.S. Department of Transportation. Retrieved from http://www.dot.gov/airconsumer/april-2013-air-travel-consumer-report.

U.S. Geological Survey. Retrieved from http://earthquake.usgs.gov/earthquakes/eqarchives/year/.

Solutions

1. soccer = 12/40 = ;
basketball = 20/40 = ;
lacrosse = 8/40 = 0.2

2. women who play soccer = 8/20 = ;
women who play basketball = 8/20 = ;
women who play lacrosse = 4/20 = ;

3. patients with the virus

5. The average length of time (in months) patients live after treatment.

7. X = the length of time (in months) patients live after treatment

9. b

11. a

13. a. .5242
b. .03 percent
c. 6.86 percent
d. $\frac{823,088}{823,856}$
e. quantitative discrete
f. quantitative continuous
g. In both years, underwater earthquakes produced massive tsunamis.
h. Answers may vary. Sample answer: A bar graph with one bar for each year, in order, would be best since it would show the change in the number of deaths from year to year. In my presentation, I would point out that the scale of the graph is in thousands, and I would discuss which specific earthquakes were responsible for the

greatest numbers of deaths in those years.

15. systematic

17. simple random

19. values for X, such as 3, 4, 11, and so on

21. No, we do not have enough information to make such a claim.

23. Take a simple random sample from each group. One way is by assigning a number to each patient and using a random number generator to randomly select patients.

25. This would be convenience sampling and is not random.

27. Yes, the sample size of 150 would be large enough to reflect a population of one school.

29. Even though the specific data support each researcher's conclusions, the different results suggest that more data need to be collected before the researchers can reach a conclusion.

30. Answers may vary. Sample answer: A pie graph would be best for showing the percentage of students that fall into each *Hours Played* category. A bar graph would be more desirable if knowing the total numbers of students in each category is important. I would be sure that the colors used on the two pie graphs are the same for each category and are clearly distinguishable when displayed. The percentages should be legible, and the pie graph should be large enough to show the smaller sections clearly. For the bar graph, I would display the bars in chronological order and make sure that the colors used for each researcher's data are clearly distinguishable. The numbers and the scale should be legible and clear when the bar graph is displayed.

32. There is not enough information given to judge if either one is correct or incorrect.

34. The software program seems to work because the second study shows that more patients improve while using the software than not. Even though the difference is not as large as that in the first study, the results from the second study are likely more reliable and still show improvement.

36. Yes, because we cannot tell if the improvement was due to the software or the exercise; the data is confounded, and a reliable conclusion cannot be drawn. New studies should be performed.

38. No, even though the sample is large enough, the fact that the sample consists of volunteers makes it a self-selected sample, which is not reliable.

40. No, even though the sample is a large portion of the population, two responses are not enough to justify any conclusions. Because the population is so small, it would be better to include everyone in the population to get the most accurate data.

42.
a. ordinal
b. interval
c. nominal
d. nominal
e. ratio
f. ordinal
g. nominal
h. interval
i. ratio
j. interval
k. ratio
l. ordinal

44. a. Inmates may not feel comfortable refusing participation, or may feel obligated to take advantage of the promised benefits. They may not feel truly free to refuse participation.
b. Parents can provide consent on behalf of their children, but children are not competent to provide consent for themselves.
c. All risks and benefits must be clearly outlined. Study participants must be informed of relevant aspects of the study in order to give appropriate consent.

45. a. statistical model: The time any journey takes from New York to Florida is variable and depends on traffic and other driving conditions.
b. statistical model: Although trains try to leave on time, the exact time of departure differs slightly from day to day.
c. mathematical model: The distance from your house to school is the same every day and can be precisely determined.
d. statistical model: The temperature of a refrigerator fluctuates as the compressor turns on and off.
e. statistical model: The fill weight of a bag of rice is different for each bag. Manufacturers spend considerable effort to minimize the variance from bag to bag.

47. a. all children who take ski or snowboard lessons
b. a group of these children
c. the population mean age of children who take their first snowboard lesson
d. the sample mean age of children who take their first snowboard lesson
e. X = the age of one child who takes his or her first ski or snowboard lesson
f. values for X, such as 3, 7, and so on

49. a. the clients of the insurance companies
b. a group of the clients
c. the mean health costs of the clients
d. the mean health costs of the sample
e. X = the health costs of one client
f. values for X, such as 34, 9, 82, and so on

51. a. all the clients of this counselor
b. a group of clients of this marriage counselor
c. the proportion of all her clients who stay married
d. the proportion of the sample of the counselor's clients who stay married
e. X = the number of couples who stay married
f. yes, no

53. a. all people (maybe in a certain geographic area, such as the United States)
b. a group of the people
c. the proportion of all people who will buy the product
d. the proportion of the sample who will buy the product
e. X = the number of people who will buy it
f. buy, not buy

55. a

57. quantitative discrete, 150

59. qualitative, Oakland A's

61. quantitative discrete, 11,234 students

63. qualitative, Crest

65. quantitative continuous, 47.3 years

67. b

69. a. The survey was conducted using six similar flights.
The survey would not be a true representation of the entire population of air travelers.
Conducting the survey on a holiday weekend will not produce representative results.
b. Conduct the survey during different times of the year.
Conduct the survey using flights to and from various locations.
Conduct the survey on different days of the week.

71. Answers will vary. Sample Answer: You could use a systematic sampling method. Stop the tenth person as they leave one of the buildings on campus at 9:50 in the morning. Then stop the tenth person as they leave a different building on campus at 1:50 in the afternoon.

73. Answers will vary. Sample Answer: Many people will not respond to mail surveys. If they do respond to the surveys, you can't be sure who is responding. In addition, mailing lists can be incomplete.

75. b

77. convenience cluster stratified systematic simple random

79. a. qualitative
b. quantitative discrete
c. quantitative discrete
d. qualitative

81. Causality: The fact that two variables are related does not guarantee that one variable is influencing the other. We cannot assume that crime rate impacts education level or that education level impacts crime rate.
Confounding: There are many factors that define a community other than education level and crime rate. Communities with high crime rates and high education levels may have other lurking variables that distinguish them from communities with lower crime rates and lower education levels. Because we cannot isolate these variables of interest, we cannot draw valid conclusions about the connection between education and crime. Possible lurking variables include police expenditures, unemployment levels, region, average age, and size.

83. a. Possible reasons: increased use of caller id, decreased use of landlines, increased use of private numbers, voice mail, privacy managers, hectic nature of personal schedules, decreased willingness to be interviewed
b. When a large number of people refuse to participate, then the sample may not have the same characteristics of the population. Perhaps the majority of people willing to participate are doing so because they feel strongly about the subject of the survey.

85. a.

# Flossing per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0	27	.4500	.4500
1	18	.3000	.7500
3	11	.1833	.9333

Table 1.44

# Flossing per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
6	3	.0500	.9833
7	1	.0167	1

Table 1.44

b. 5.00 percent
c. 93.33 percent

87. The sum of the travel times is 1,173.1. Divide the sum by 50 to calculate the mean value: 23.462. Because each state's travel time was measured to the nearest tenth, round this calculation to the nearest hundredth: 23.46.

89. b

91. Explanatory variable: amount of sleep
Response variable: performance measured in assigned tasks
Treatments: normal sleep and 27 hours of total sleep deprivation
Experimental Units: 19 professional drivers
Lurking variables: none – all drivers participated in both treatments
Random assignment: treatments were assigned in random order; this eliminated the effect of any *learning* that may take place during the first experimental session
Control/Placebo: completing the experimental session under normal sleep conditions
Blinding: researchers evaluating subjects' performance must not know which treatment is being applied at the time

93. You cannot assume that the numbers of complaints reflect the quality of the airlines. The airlines shown with the greatest number of complaints are the ones with the most passengers. You must consider the appropriateness of methods for presenting data; in this case displaying totals is misleading.

94. He can observe a population of 100 college students on campus. He can collect data about the temperature of their dorm rooms and track how many of them catch a cold. If he uses a survey, the temperature of the dorm rooms can be determined from the survey. He can also ask them to self-report when they catch a cold.

96. Answers will vary. Sample answer: The sample is not representative of the population of all college textbooks. Two reasons why it is not representative are that he only sampled seven subjects and he only investigated one textbook in each subject. There are several possible sources of bias in the study. The seven subjects that he investigated are all in mathematics and the sciences; there are many subjects in the humanities, social sciences, and other subject areas, for example: literature, art, history, psychology, sociology, business, that he did not investigate at all. It may be that different subject areas exhibit different patterns of textbook availability, but his sample would not detect such results.
He also looked only at the most popular textbook in each of the subjects he investigated. The availability of the most popular textbooks may differ from the availability of other textbooks in one of two ways:

- The most popular textbooks may be more readily available online, because more new copies are printed, and more students nationwide are selling back their used copies
- The most popular textbooks may be harder to find available online, because more student demand exhausts the supply more quickly.

In reality, many college students do not use the most popular textbook in their subject, and this study gives no useful information about the situation for those less popular textbooks.
He could improve this study by

- expanding the selection of subjects he investigates so that it is more representative of all subjects studied by college students, and
- expanding the selection of textbooks he investigates within each subject to include a mixed representation of both the most popular and less popular textbooks.

2 DESCRIPTIVE STATISTICS

Figure 2.1 When you have a large amount of data, you will need to organize it in a way that makes sense. These ballots from an election are rolled together with similar ballots to keep them organized. (credit: William Greeson)

Chapter Objectives

By the end of this chapter, the student should be able to do the following:

- Display data graphically and interpret the following graphs: stem-and-leaf plots, line graphs, bar graphs, frequency polygons, time series graphs, histograms, box plots, and dot plots
- Recognize, describe, and calculate the measures of location of data with quartiles and percentiles
- Recognize, describe, and calculate the measures of the center of data with mean, median, and mode
- Recognize, describe, and calculate the measures of the spread of data with variance, standard deviation, and range

Introduction

Once you have a data collection, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.

In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called *descriptive statistics.* You will learn how to calculate and, even more important, how to interpret these measurements and graphs.

A statistical graph is a tool that helps you learn about the shape or distribution of a sample or a population. A graph can be a more effective way of presenting data than a mass of numbers because we can see where data values cluster and where there are only a few data values. Newspapers and the internet use graphs to show trends and to enable readers to compare facts and figures quickly. Statisticians often graph data first to get a picture of the data. Then more formal

tools may be applied.

Some of the types of graphs that are used to summarize and organize data are the dot plot, the bar graph, the histogram, the stem-and-leaf plot, the frequency polygon—a type of broken line graph—the pie chart, and the box plot. In this chapter, we will briefly look at stem-and-leaf plots, line graphs, and bar graphs as well as frequency polygons, time series graphs, and dot plots. Our emphasis will be on histograms and box plots.

NOTE

This book contains instructions for constructing a histogram and a box plot for the TI-83+ and TI-84 calculators. The Texas Instruments (TI) website (http://education.ti.com/educationportal/sites/US/sectionHome/support.html) provides additional instructions for using these calculators.

2.1 Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

One simple graph, the **stem-and-leaf graph** or **stemplot**, comes from the field of exploratory data analysis. It is a good choice when the data sets are small. To create the plot, divide each observation of data into a stem and a leaf. The stem consists of the leading digit(s), while the leaf consists of a **final significant digit**. For example, 23 has stem two and leaf three. The number 432 has stem 43 and leaf two. Likewise, the number 5,432 has stem 543 and leaf two. The decimal 9.3 has stem nine and leaf three. Write the stems in a vertical line from smallest to largest. Draw a vertical line to the right of the stems. Then write the leaves in increasing order next to their corresponding stem. Make sure the leaves show a space between values, so that the exact data values may be easily determined. The frequency of data values for each stem provides information about the shape of the distribution.

EXAMPLE 2.1

For Susan Dean's spring precalculus class, scores for the first exam were as follows (smallest to largest):
33, 42, 49, 49, 53, 55, 55, 61, 63, 67, 68, 68, 69, 69, 72, 73, 74, 78, 80, 83, 88, 88, 88, 90, 92, 94, 94, 94, 94, 96, 100

Stem	Leaf
3	3
4	2 9 9
5	3 5 5
6	1 3 7 8 8 9 9
7	2 3 4 8
8	0 3 8 8 8
9	0 2 4 4 4 4 6
10	0

Table 2.1 Stem-and-Leaf Graph

The stemplot shows that most scores fell in the 60s, 70s, 80s, and 90s. Eight out of the 31 scores or approximately 26 percent $\left(\frac{8}{31}\right)$ were in the 90s or 100, a fairly high number of As.

TRY IT 2.1

For the Park City basketball team, scores for the last 30 games were as follows (smallest to largest):
32, 32, 33, 34, 38, 40, 42, 42, 43, 44, 46, 47, 47, 48, 48, 48, 49, 50, 50, 51, 52, 52, 52, 53, 54, 56, 57, 57, 60, 61
Construct a stemplot for the data.

The stemplot is a quick way to graph data and gives an exact picture of the data. You want to look for an overall pattern and any outliers. An **outlier** is an observation of data that does not fit the rest of the data. It is sometimes called an **extreme value.** When you graph an outlier, it will appear not to fit the pattern of the graph. Some outliers are due to mistakes, for example, writing 50 instead of 500, while others may indicate that something unusual is happening. It takes some background information to explain outliers, so we will cover them in more detail later.

EXAMPLE 2.2

The data are the distances (in kilometers) from a home to local supermarkets. Create a stemplot using the data.
1.1, 1.5, 2.3, 2.5, 2.7, 3.2, 3.3, 3.3, 3.5, 3.8, 4.0, 4.2, 4.5, 4.5, 4.7, 4.8, 5.5, 5.6, 6.5, 6.7, 12.3

Problem
Do the data seem to have any concentration of values?

The leaves are to the right of the decimal.

Solution
The value 12.3 may be an outlier. Values appear to concentrate at 3 and 4 kilometers.

Stem	Leaf
1	1 5
2	3 5 7
3	2 3 3 5 8
4	0 2 5 5 7 8
5	5 6
6	5 7
7	
8	
9	
10	
11	
12	3

Table 2.2

TRY IT 2.2

The data below show the distances (in miles) from the homes of high school students to the school. Create a stemplot using the following data and identify any outliers.

0.5, 0.7, 1.1, 1.2, 1.2, 1.3, 1.3, 1.5, 1.5, 1.7, 1.7, 1.8, 1.9, 2.0, 2.2, 2.5, 2.6, 2.8, 2.8, 2.8, 3.5, 3.8, 4.4, 4.8, 4.9, 5.2, 5.5, 5.7, 5.8, 8.0

EXAMPLE 2.3

Problem

A **side-by-side stem-and-leaf plot** allows a comparison of the two data sets in two columns. In a side-by-side stem-and-leaf plot, two sets of leaves share the same stem. The leaves are to the left and the right of the stems. Table 2.3 and Table 2.4 show the ages of presidents at their inauguration and at their death. Construct a side-by-side stem-and-leaf plot using these data.

President	Age	President	Age	President	Age
Washington	57	Lincoln	52	Hoover	54
J. Adams	61	A. Johnson	56	F. Roosevelt	51
Jefferson	57	Grant	46	Truman	60
Madison	57	Hayes	54	Eisenhower	62
Monroe	58	Garfield	49	Kennedy	43
J. Q. Adams	57	Arthur	51	L. Johnson	55
Jackson	61	Cleveland	47	Nixon	56
Van Buren	54	B. Harrison	55	Ford	61
W. H. Harrison	68	Cleveland	55	Carter	52
Tyler	51	McKinley	54	Reagan	69
Polk	49	T. Roosevelt	42	G.H.W. Bush	64
Taylor	64	Taft	51	Clinton	47
Fillmore	50	Wilson	56	G. W. Bush	54
Pierce	48	Harding	55	Obama	47
Buchanan	65	Coolidge	51		

Table 2.3 Presidential Ages at Inauguration

President	Age	President	Age	President	Age
Washington	67	Lincoln	56	Hoover	90
J. Adams	90	A. Johnson	66	F. Roosevelt	63
Jefferson	83	Grant	63	Truman	88
Madison	85	Hayes	70	Eisenhower	78
Monroe	73	Garfield	49	Kennedy	46
J. Q. Adams	80	Arthur	56	L. Johnson	64
Jackson	78	Cleveland	71	Nixon	81
Van Buren	79	B. Harrison	67	Ford	93
W. H. Harrison	68	Cleveland	71	Reagan	93
Tyler	71	McKinley	58		
Polk	53	T. Roosevelt	60		
Taylor	65	Taft	72		
Fillmore	74	Wilson	67		
Pierce	64	Harding	57		
Buchanan	77	Coolidge	60		

Table 2.4 Presidential Age at Death

Solution

Ages at Inauguration		Ages at Death
9 9 8 7 7 7 6 3 2	4	6 9
8 7 7 7 7 6 6 6 5 5 5 5 4 4 4 4 4 2 1 1 1 1 1 0	5	3 6 6 7 7 8
9 5 4 4 2 1 1 1 0	6	0 0 3 3 4 4 5 6 7 7 7 8
	7	0 0 1 1 1 4 7 8 8 9
	8	0 1 3 5 8
	9	0 0 3 3

Table 2.5

Notice that the leaf values increase in order, from right to left, for leaves shown to the left of the stem, while the leaf values increase in order from left to right, for leaves shown to the right of the stem.

TRY IT 2.3

The table shows the number of wins and losses a sports team has had in 42 seasons. Create a side-by-side stem-and-leaf plot of these wins and losses.

Losses	Wins	Year	Losses	Wins	Year
34	48	1968–1969	41	41	1989–1990
34	48	1969–1970	39	43	1990–1991
46	36	1970–1971	44	38	1991–1992
46	36	1971–1972	39	43	1992–1993
36	46	1972–1973	25	57	1993–1994
47	35	1973–1974	40	42	1994–1995
51	31	1974–1975	36	46	1995–1996
53	29	1975–1976	26	56	1996–1997
51	31	1976–1977	32	50	1997–1998
41	41	1977–1978	19	31	1998–1999
36	46	1978–1979	54	28	1999–2000
32	50	1979–1980	57	25	2000–2001
51	31	1980–1981	49	33	2001–2002
40	42	1981–1982	47	35	2002–2003
39	43	1982–1983	54	28	2003–2004
42	40	1983–1984	69	13	2004–2005
48	34	1984–1985	56	26	2005–2006
32	50	1985–1986	52	30	2006–2007
25	57	1986–1987	45	37	2007–2008
32	50	1987–1988	35	47	2008–2009
30	52	1988–1989	29	53	2009–2010

Table 2.6

Another type of graph that is useful for specific data values is a **line graph**. In the particular line graph shown in Example 2.4, the ***x*-axis** (horizontal axis) consists of **data values** and the ***y*-axis** (vertical axis) consists of **frequency points**. The frequency points are connected using line segments.

EXAMPLE 2.4

In a survey, 40 mothers were asked how many times per week a teenager must be reminded to do his or her chores. The results are shown in Table 2.7 and in Figure 2.2.

Number of Times Teenager Is Reminded	Frequency
0	2
1	5
2	8
3	14
4	7
5	4

Table 2.7

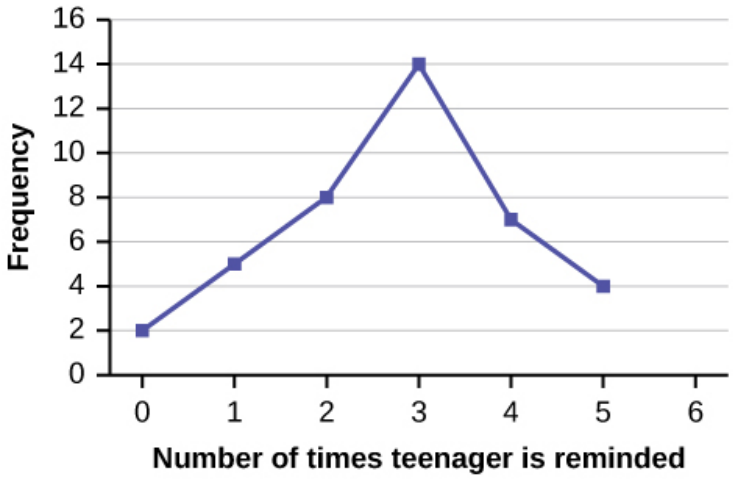


Figure 2.2

TRY IT 2.4

In a survey, 40 people were asked how many times per year they had their car in the shop for repairs. The results are shown in Table 2.8. Construct a line graph.

Number of Times in Shop	Frequency
0	7
1	10
2	14
3	9

Table 2.8

Bar graphs consist of bars that are separated from each other. The bars can be rectangles, or they can be rectangular boxes, used in three-dimensional plots, and they can be vertical or horizontal. The **bar graph** shown in Example 2.5 has age-groups represented on the ***x*-axis** and proportions on the ***y*-axis**.

EXAMPLE 2.5

Problem

By the end of 2011, a social media site had more than 146 million users in the United States. Table 2.9 shows three age-groups, the number of users in each age-group, and the proportion (percentage) of users in each age-group. Construct a bar graph using this data.

Age-Groups	Number of Site Users	Proportion (%) of Site Users
13–25	65,082,280	45%
26–44	53,300,200	36%
45–64	27,885,100	19%

Table 2.9

Solution

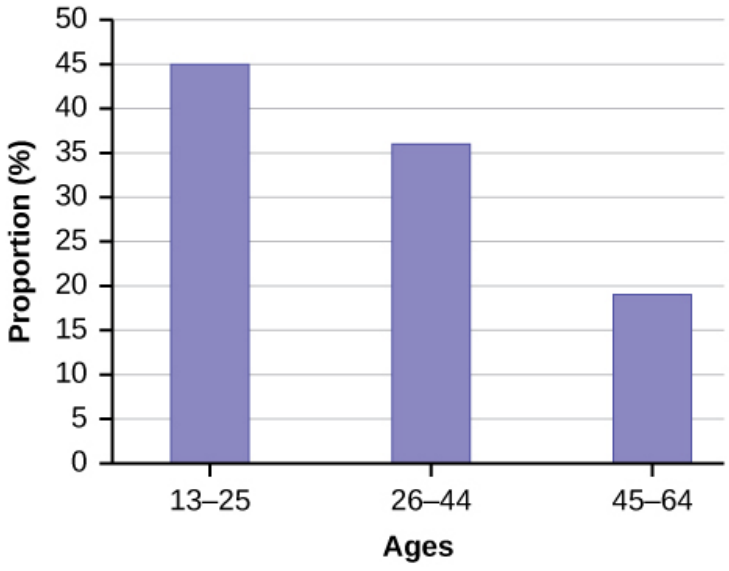


Figure 2.3

TRY IT 2.5

The population in Park City is made up of children, working-age adults, and retirees. Table 2.10 shows the three age-groups, the number of people in the town from each age-group, and the proportion (%) of people in each age-group. Construct a bar graph showing the proportions.

Age-Groups	Number of People	Proportion of Population
Children	67,059	19%
Working-age adults	152,198	43%
Retirees	131,662	38%

Table 2.10

EXAMPLE 2.6

Problem

The columns in Table 2.11 contain the race or ethnicity of students in U.S. public schools for the class of 2011, percentages for the Advanced Placement (AP) examinee population for that class, and percentages for the overall student population. Create a bar graph with the student race or ethnicity (qualitative data) on the *x*-axis and the AP examinee population percentages on the *y*-axis.

Race/Ethnicity	AP Examinee Population	Overall Student Population
1 = Asian, Asian American, or Pacific Islander	10.3%	5.7%
2 = Black or African American	9.0%	14.7%
3 = Hispanic or Latino	17.0%	17.6%
4 = American Indian or Alaska Native	0.6%	1.1%
5 = White	57.1%	59.2%
6 = Not reported/other	6.0%	1.7%

Table 2.11

Solution

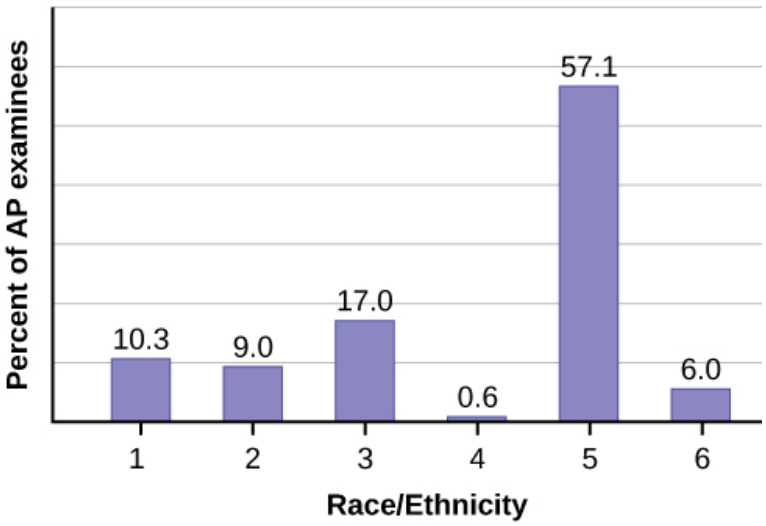


Figure 2.4

TRY IT 2.6

Park City is broken down into six voting districts. The table shows the percentage of the total registered voter population that lives in each district as well as the percentage of the entire population that lives in each district. Construct a bar graph that shows the registered voter population by district.

District	Registered Voter Population	Overall City Population
1	15.5%	19.4%
2	12.2%	15.6%
3	9.8%	9.0%
4	17.4%	18.5%
5	22.8%	20.7%
6	22.3%	16.8%

Table 2.12

EXAMPLE 2.7

Problem

Table 2.13 is a two-way table showing the types of pets owned by men and women.

	Dogs	Cats	Fish	Total
Men	4	2	2	8
Women	4	6	2	12
Total	8	8	4	20

Table 2.13

Given these data, calculate the marginal distributions of pets for the people surveyed.

Solution

$\text{Dogs} = 8/20 = 0.4$

$\text{Cats} = 8/20 = 0.4$

$\text{Fish} = 4/20 = 0.2$

Note—The sum of all the marginal distributions must equal one. In this case, $0.4 + 0.4 + 0.2 = 1$; therefore, the solution *checks*.

EXAMPLE 2.8

Problem

Table 2.14 is a two-way table showing the types of pets owned by men and women.

	Dogs	Cats	Fish	Total
Men	4	2	2	8
Women	4	6	2	12
Total	8	8	4	20

Table 2.14

Given these data, calculate the conditional distributions for the subpopulation of men who own each pet type.

Solution

$\text{Men who own dogs} = 4/8 = 0.5$

$\text{Men who own cats} = 2/8 = 0.25$

$\text{Men who own fish} = 2/8 = 0.25$

Note—The sum of all the conditional distributions must equal one. In this case, $0.5 + 0.25 + 0.25 = 1$; therefore, the solution *checks*.

2.2 Histograms, Frequency Polygons, and Time Series Graphs

For most of the work you do in this book, you will use a histogram to display the data. One advantage of a histogram is that it can readily display large data sets.

A **histogram** consists of contiguous (adjoining) boxes. It has both a horizontal axis and a vertical axis. The horizontal axis is more or less a number line, labeled with what the data represents, for example, distance from your home to school. The vertical axis is labeled either **frequency** or **relative frequency** (or percent frequency or probability). The graph will have the same shape with either label. The histogram (like the stemplot) can give you the shape of the data, the center, and the spread of the data. The shape of the data refers to the shape of the distribution, whether normal, approximately

normal, or skewed in some direction, whereas the center is thought of as the middle of a data set, and the spread indicates how far the values are dispersed about the center. In a skewed distribution, the mean is pulled toward the tail of the distribution.

The relative frequency is equal to the frequency for an observed value of the data divided by the total number of data values in the sample. Remember, frequency is defined as the number of times an answer occurs. If

- f = frequency,
- n = total number of data values (or the sum of the individual frequencies), and
- RF = relative frequency,

then

$$RF = \frac{f}{n}.$$

For example, if three students in Mr. Ahab's English class of 40 students received from ninety to 100 percent, then $f = 3$, $n = 40$, and $RF = \frac{f}{n} = \frac{3}{40} = 0.075$. Thus, 7.5 percent of the students received 90 to 100 percent. Ninety to 100 percent is a quantitative measures.

To construct a histogram, first decide how many **bars** or **intervals**, also called classes, represent the data. Many histograms consist of five to 15 bars or classes for clarity. The width of each bar is also referred to as the bin size, which may be calculated by dividing the range of the data values by the desired number of bins (or bars). There is not a set procedure for determining the number of bars or bar width/bin size; however, consistency is key when determining which data values to place inside each interval.

EXAMPLE 2.9

The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. The heights are **continuous** data since height is measured.
60, 60.5, 61, 61, 61.5,
63.5, 63.5, 63.5,
64, 64, 64, 64, 64, 64, 64, 64.5, 64.5, 64.5, 64.5, 64.5, 64.5, 64.5, 64.5,
66, 66, 66, 66, 66, 66, 66, 66, 66, 66, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 67, 67, 67, 67, 67, 67, 67,
67, 67, 67, 67, 67, 67.5, 67.5, 67.5, 67.5, 67.5, 67.5, 67.5,
68, 68, 69, 69, 69, 69, 69, 69, 69, 69, 69, 69, 69.5, 69.5, 69.5, 69.5, 69.5,
70, 70, 70, 70, 70, 70, 70.5, 70.5, 70.5, 71, 71, 71,
72, 72, 72, 72.5, 72.5, 73, 73.5,
74

The smallest data value is 60, and the largest data value is 74. To make sure each is included in an interval, we can use 59.95 as the smallest value and 74.05 as the largest value, subtracting and adding .05 to these values, respectively. We have a small range here of 14.1 (74.05 – 59.95), so we will want a fewer number of bins; let"s say eight. So, 14.1 divided by eight bins gives a bin size (or interval size) of approximately 1.76.

NOTE

We will round up to two and make each bar or class interval two units wide. Rounding up to two is a way to prevent a value from falling on a boundary. Rounding to the next number is often necessary even if it goes against the standard rules of rounding. For this example, using 1.76 as the width would also work. A guideline that is followed by some for the width of a bar or class interval is to take the square root of the number of data values and then round to the nearest whole number, if necessary. For example, if there are 150 values of data, take the square root of 150 and round to 12 bars or intervals.

The boundaries are as follows:

- 59.95
- 59.95 + 2 = 61.95
- 61.95 + 2 = 63.95

- 63.95 + 2 = 65.95
- 65.95 + 2 = 67.95
- 67.95 + 2 = 69.95
- 69.95 + 2 = 71.95
- 71.95 + 2 = 73.95
- 73.95 + 2 = 75.95

The heights 60 through 61.5 inches are in the interval 59.95–61.95. The heights that are 63.5 are in the interval 61.95–63.95. The heights that are 64 through 64.5 are in the interval 63.95–65.95. The heights 66 through 67.5 are in the interval 65.95–67.95. The heights 68 through 69.5 are in the interval 67.95–69.95. The heights 70 through 71 are in the interval 69.95–71.95. The heights 72 through 73.5 are in the interval 71.95–73.95. The height 74 is in the interval 73.95–75.95.

The following histogram displays the heights on the *x*-axis and relative frequency on the *y*-axis.

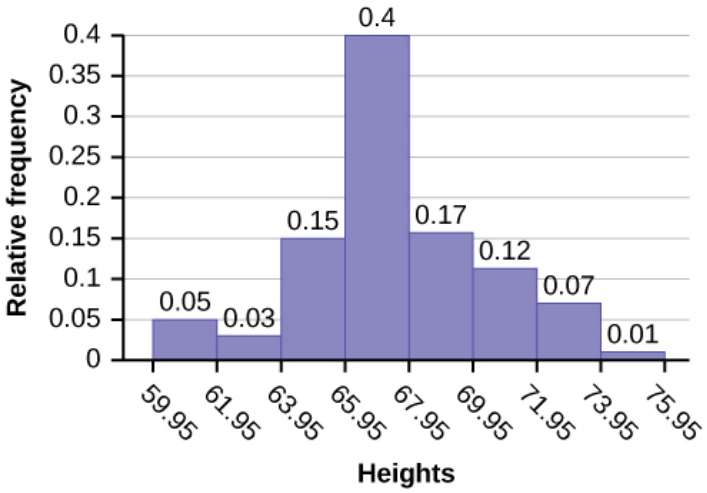


Figure 2.5

Interval	Frequency	Relative Frequency
59.95–61.95	5	5/100 = 0.05
61.95–63.95	3	3/100 = 0.03
63.95–65.95	15	15/100 = 0.15
65.95–67.95	40	40/100 = 0.40
67.95–69.95	17	17/100 = 0.17
69.95–71.95	12	12/100 = 0.12
71.95–73.95	7	7/100 = 0.07
73.95–75.95	1	1/100 = 0.01

Table 2.15

TRY IT 2.9

The following data are the shoe sizes of 50 male students. The sizes are continuous data since shoe size is measured. Construct a histogram and calculate the width of each bar or class interval. Use six bars on the histogram.
9, 9, 9.5, 9.5, 10, 10, 10, 10, 10, 10, 10.5, 10.5, 10.5, 10.5, 10.5, 10.5, 10.5, 10.5,
11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11.5, 11.5, 11.5, 11.5, 11.5, 11.5, 11.5,
12, 12, 12, 12, 12, 12, 12, 12.5, 12.5, 12.5, 12.5, 14

EXAMPLE 2.10

The following data are the number of books bought by 50 part-time college students at ABC College. The number of books is **discrete data** since books are counted.
1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
4, 4, 4, 4, 4, 4,
5, 5, 5, 5, 5,
6, 6

Eleven students buy one book. Ten students buy two books. Sixteen students buy three books. Six students buy four books. Five students buy five books. Two students buy six books.

Problem
Calculate the width of each bar/bin size/interval size.

Solution
The smallest data value is 1, and the largest data value is 6. To make sure each is included in an interval, we can use 0.5 as the smallest value and 6.5 as the largest value by subtracting and adding 0.5 to these values. We have a small range here of 6 (6.5 –– 0.5), so we will want a fewer number of bins; let"s say six this time. So, six divided by six bins gives a bin size (or interval size) of one.

Notice that we may choose different rational numbers to add to, or subtract from, our maximum and minimum values when calculating bin size. In the previous example, we added and subtracted .05, while this time, we added and subtracted .5. Given a data set, you will be able to determine what is appropriate and reasonable.

The following histogram displays the number of books on the *x*-axis and the frequency on the *y*-axis.

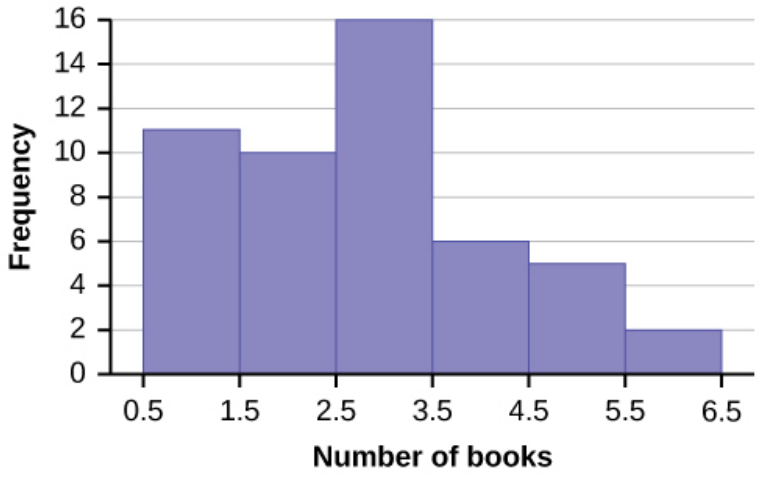


Figure 2.6

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Go to Appendix G. There are calculator instructions for entering data and for creating a customized histogram. Create the histogram for Example 2.10.

- Press Y=. Press CLEAR to delete any equations.
- Press STAT 1:EDIT. If L1 has data in it, arrow up into the name L1, press CLEAR and then arrow down. If necessary, do the same for L2.
- Into L1, enter 1, 2, 3, 4, 5, 6. Note that these values represent the numbers of books.
- Into L2, enter 11, 10, 16, 6, 5, 2. Note that these numbers represent the frequencies for the numbers of books.
- Press WINDOW. Set Xmin = .5, Xscl = (6.5 – .5)/6, Ymin = –1, Ymax = 20, Yscl = 1, Xres = 1. The window settings are chosen to accurately and completely show the data value range and the frequency range.
- Press second Y=. Start by pressing 4:Plotsoff ENTER.
- Press second Y=. Press 1:Plot1. Press ENTER. Arrow down to TYPE. Arrow to the third picture (histogram). Press ENTER.
- Arrow down to Xlist: Enter L1 (2^{nd} 1). Arrow down to Freq. Enter L2 (second 2).
- Press GRAPH.
- Use the TRACE key and the arrow keys to examine the histogram.

TRY IT 2.10

The following data are the number of sports played by 50 student athletes. The number of sports is discrete data since sports are counted.

1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
3, 3, 3, 3, 3, 3, 3, 3
Twenty student athletes play one sport. Twenty-two student athletes play two sports. Eight student athletes play three sports. Calculate a desired bin size for the data. Create a histogram and clearly label the endpoints of the intervals.

EXAMPLE 2.11

Problem
Using this data set, construct a histogram.

Number of Hours My Classmates Spent Playing Video Games on Weekends				
9.95	10	2.25	16.75	0
19.5	22.5	7.5	15	12.75
5.5	11	10	20.75	17.5
23	21.9	24	23.75	18
20	15	22.9	18.8	20.5

Table 2.16

Solution

Figure 2.7

Some values in this data set fall on boundaries for the class intervals. A value is counted in a class interval if it falls on the left boundary but not if it falls on the right boundary. Different researchers may set up histograms for the same data in different ways. There is more than one correct way to set up a histogram.

TRY IT 2.11

The following data represent the number of employees at various restaurants in New York City. Using this data, create a histogram.

22, 35, 15, 26, 40, 28, 18, 20, 25, 34, 39, 42, 24, 22, 19, 27, 22, 34, 40, 20, 38, 28

COLLABORATIVE EXERCISE

Count the money (bills and change) in your pocket or purse. Your instructor will record the amounts. As a class, construct a histogram displaying the data. Discuss how many intervals you think would be appropriate. You may want to experiment with the number of intervals.

Frequency Polygons

Frequency polygons are analogous to line graphs, and just as line graphs make continuous data visually easy to interpret, so too do frequency polygons.

To construct a frequency polygon, first examine the data and decide on the number of intervals and resulting interval size, for both the *x*-axis and *y*-axis. The *x*-axis will show the lower and upper bound for each interval, containing the data values, whereas the *y*-axis will represent the frequencies of the values. Each data point represents the frequency for each interval. For example, if an interval has three data values in it, the frequency polygon will show a 3 at the upper endpoint of that interval. After choosing the appropriate intervals, begin plotting the data points. After all the points are plotted, draw line segments to connect them.

EXAMPLE 2.12

A frequency polygon was constructed from the frequency table below.

Frequency Distribution for Calculus Final Test Scores			
Lower Bound	**Upper Bound**	**Frequency**	**Cumulative Frequency**
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45
79.5	89.5	40	85
89.5	99.5	15	100

Table 2.17

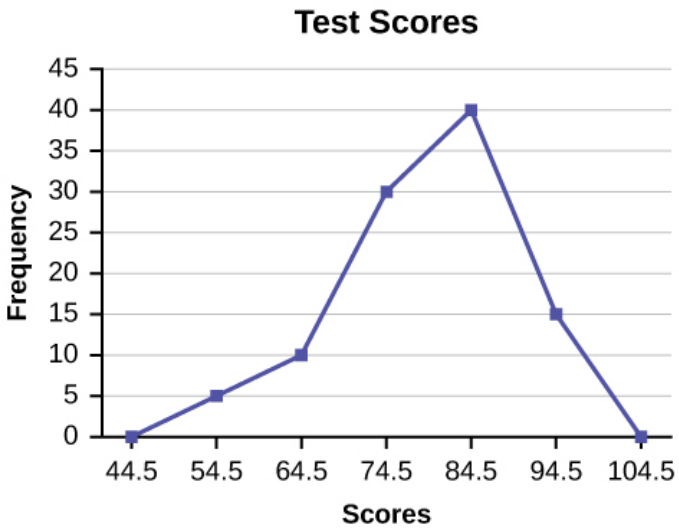


Figure 2.8

Notice that each point represents frequency for a particular interval. These points are located halfway between the lower bound and upper bound. In fact, the horizontal axis, or *x*-axis, shows only these midpoint values. For the interval 49.5–59.5 the value 54.5 is represented by a point, showing the correct frequency of 5. For the interval occurring before 49.5–59.5, (as well as 39.5–49.5), the value of the midpoint, or 44.5, is represented by a point, showing a frequency of 0, since we do not have any values in that range. The same idea applies to the last interval of 99.5–109.5, which has a midpoint of 104.5 and correctly shows a point representing a frequency of 0. Looking at the graph, we say that this distribution is skewed because one side of the graph does not mirror the other side.

TRY IT 2.12

Construct a frequency polygon of U.S. presidents' ages at inauguration shown in Table 2.18.

Age at Inauguration	Frequency
41.5–46.5	4
46.5–51.5	11
51.5–56.5	14
56.5–61.5	9
61.5–66.5	4
66.5–71.5	2

Table 2.18

Frequency polygons are useful for comparing distributions. This comparison is achieved by overlaying the frequency polygons drawn for different data sets.

EXAMPLE 2.13

We will construct an overlay frequency polygon comparing the scores from Example 2.12 with the students' final numeric grades.

Frequency Distribution for Calculus Final Test Scores			
Lower Bound	**Upper Bound**	**Frequency**	**Cumulative Frequency**
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45
79.5	89.5	40	85
89.5	99.5	15	100

Table 2.19

Frequency Distribution for Calculus Final Grades			
Lower Bound	**Upper Bound**	**Frequency**	**Cumulative Frequency**
49.5	59.5	10	10
59.5	69.5	10	20
69.5	79.5	30	50

Table 2.20

Frequency Distribution for Calculus Final Grades			
Lower Bound	Upper Bound	Frequency	Cumulative Frequency
79.5	89.5	45	95
89.5	99.5	5	100

Table 2.20

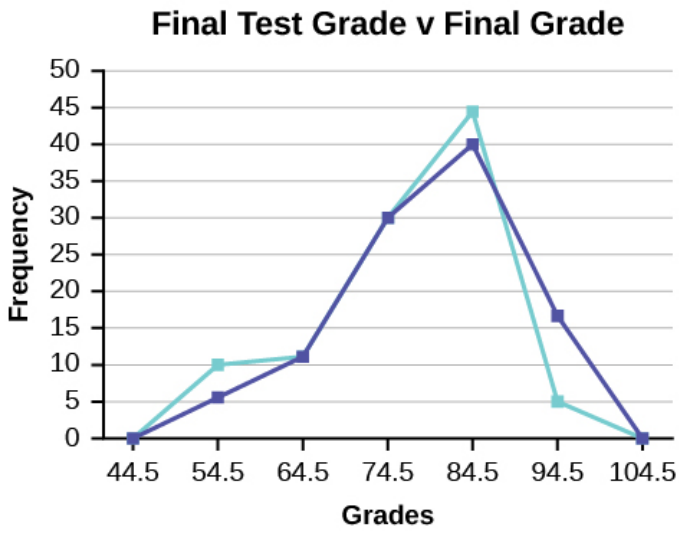


Figure 2.9

Suppose that we want to study the temperature range of a region for an entire month. Every day at noon, we note the temperature and write this down in a log. A variety of statistical studies could be done with these data. We could find the mean or the median temperature for the month. We could construct a histogram displaying the number of days that temperatures reach a certain range of values. However, all of these methods ignore a portion of the data that we have collected.

One feature of the data that we may want to consider is that of time. Since each date is paired with the temperature reading for the day, we don't have to think of the data as being random. We can instead use the times given to impose a chronological order on the data. A graph that recognizes this ordering and displays the changing temperature as the month progresses is called a time series graph.

Constructing a Time Series Graph

To construct a time series graph, we must look at both pieces of our **paired data set**. We start with a standard Cartesian coordinate system. The horizontal axis is used to plot the date or time increments, and the vertical axis is used to plot the values of the variable that we are measuring. By using the axes in that way, we make each point on the graph correspond to a date and a measured quantity. The points on the graph are typically connected by straight lines in the order in which they occur.

EXAMPLE 2.14

Problem

The following data show the Annual Consumer Price Index each month for 10 years. Construct a time series graph for the Annual Consumer Price Index data only.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul
2003	181.7	183.1	184.2	183.8	183.5	183.7	183.9
2004	185.2	186.2	187.4	188.0	189.1	189.7	189.4
2005	190.7	191.8	193.3	194.6	194.4	194.5	195.4
2006	198.3	198.7	199.8	201.5	202.5	202.9	203.5
2007	202.416	203.499	205.352	206.686	207.949	208.352	208.299
2008	211.080	211.693	213.528	214.823	216.632	218.815	219.964
2009	211.143	212.193	212.709	213.240	213.856	215.693	215.351
2010	216.687	216.741	217.631	218.009	218.178	217.965	218.011
2011	220.223	221.309	223.467	224.906	225.964	225.722	225.922
2012	226.665	227.663	229.392	230.085	229.815	229.478	229.104

Table 2.21

Year	Aug	Sep	Oct	Nov	Dec	Annual
2003	184.6	185.2	185.0	184.5	184.3	184.0
2004	189.5	189.9	190.9	191.0	190.3	188.9
2005	196.4	198.8	199.2	197.6	196.8	195.3
2006	203.9	202.9	201.8	201.5	201.8	201.6
2007	207.917	208.490	208.936	210.177	210.036	207.342
2008	219.086	218.783	216.573	212.425	210.228	215.303
2009	215.834	215.969	216.177	216.330	215.949	214.537
2010	218.312	218.439	218.711	218.803	219.179	218.056
2011	226.545	226.889	226.421	226.230	225.672	224.939
2012	230.379	231.407	231.317	230.221	229.601	229.594

Table 2.22

Solution

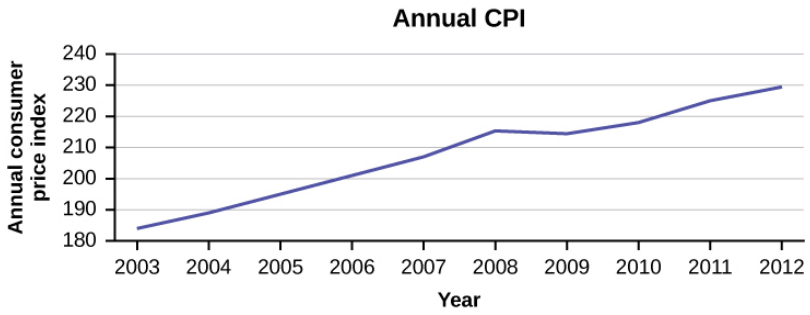


Figure 2.10 The annual amounts are plotted for each year. Then, consecutive points are connected with a line.

TRY IT 2.14

The following table is a portion of a data set from a banking website. Use the table to construct a time series graph for CO_2 emissions for the United States.

	CO_2 Emissions		
	Ukraine	United Kingdom	United States
2003	352,259	540,640	5,681,664
2004	343,121	540,409	5,790,761
2005	339,029	541,990	5,826,394
2006	327,797	542,045	5,737,615
2007	328,357	528,631	5,828,697
2008	323,657	522,247	5,656,839
2009	272,176	474,579	5,299,563

Table 2.23

Uses of a Time Series Graph

Time series graphs are important tools in various applications of statistics. When a researcher records values of the same variable over an extended period of time, it is sometimes difficult for him or her to discern any trend or pattern. However, once the same data points are displayed graphically, some features jump out. Time series graphs make trends easy to spot.

2.3 Measures of the Location of the Data

The common measures of location are **quartiles** and **percentiles**.

Quartiles are special percentiles. The **first quartile**, Q_1, is the same as the 25^{th} percentile, and the third quartile, Q_3, is the same as the 75^{th} percentile. The median, M, is called both the second quartile and the 50^{th} percentile.

To calculate quartiles and percentiles, you must order the data from smallest to largest. Quartiles divide ordered data into quarters. Percentiles divide ordered data into hundredths. Recall that a percent means one-hundredth. So, percentiles mean the data is divided into 100 sections. To score in the 90^{th} percentile of an exam does not mean,

necessarily, that you received 90 percent on a test. It means that 90 percent of test scores are the same as or less than your score and that 10 percent of the test scores are the same as or greater than your test score.

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively. One instance in which colleges and universities use percentiles is when SAT results are used to determine a minimum testing score that will be used as an acceptance factor. For example, suppose Duke accepts SAT scores at or above the 75th percentile. That translates into a score of at least 1220.

Percentiles are mostly used with very large populations. Therefore, if you were to say that 90 percent of the test scores are less, and not the same or less, than your score, it would be acceptable because removing one particular data value is not significant.

The **median** is a number that measures the *center* of the data. You can think of the median as the *middle value*, but it does not actually have to be one of the observed values. It is a number that separates ordered data into halves. Half the values are the same number or smaller than the median, and half the values are the same number or larger. For example, consider the following data:
1, 11.5, 6, 7.2, 4, 8, 9, 10, 6.8, 8.3, 2, 2, 10, 1
Ordered from smallest to largest:
1, 1, 2, 2, 4, 6, 6.8, 7.2, 8, 8.3, 9, 10, 10, 11.5

When a data set has an even number of data values, the median is equal to the average of the two middle values when the data are arranged in ascending order (least to greatest). When a data set has an odd number of data values, the median is equal to the middle value when the data are arranged in ascending order.

Since there are 14 observations (an even number of data values), the median is between the seventh value, 6.8, and the eighth value, 7.2. To find the median, add the two values together and divide by two.

$$\frac{6.8 + 7.2}{2} = 7 \quad \textbf{2.1}$$

The median is seven. Half of the values are smaller than seven and half of the values are larger than seven.

Quartiles are numbers that separate the data into quarters. Quartiles may or may not be part of the data. To find the quartiles, first find the median, or second, quartile. The first quartile, Q_1, is the middle value of the lower half of the data, and the third quartile, Q_3, is the middle value, or median, of the upper half of the data. To get the idea, consider the same data set:
1, 1, 2, 2, 4, 6, 6.8, 7.2, 8, 8.3, 9, 10, 10, 11.5

The data set has an even number of values (14 data values), so the median will be the average of the two middle values (the average of 6.8 and 7.2), which is calculated as $\frac{6.8+7.2}{2}$ and equals 7.

So, the median, or second quartile (Q_2), is 7.

The first quartile is the median of the lower half of the data, so if we divide the data into seven values in the lower half and seven values in the upper half, we can see that we have an odd number of values in the lower half. Thus, the median of the lower half, or the first quartile (Q_1) will be the middle value, or 2. Using the same procedure, we can see that the median of the upper half, or the third quartile (Q_3) will be the middle value of the upper half, or 9.

The quartiles are illustrated below:

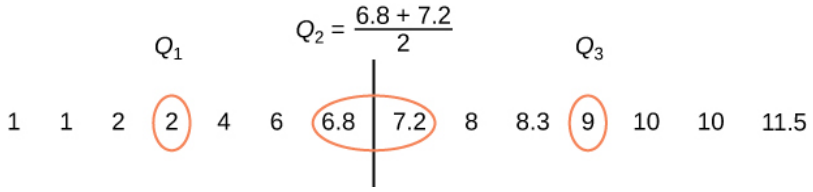


Figure 2.11

The **interquartile range** is a number that indicates the spread of the middle half, or the middle 50 percent of the data. It is the difference between the third quartile (Q_3) and the first quartile (Q_1)

$IQR = Q_3 - Q_1$. The IQR for this data set is calculated as 9 minus 2, or 7.

The IQR can help to determine potential **outliers**. **A value is suspected to be a potential outlier if it is less than 1.5 × *IQR* below the first quartile or more than 1.5 × *IQR* above the third quartile**. Potential outliers always require further investigation.

NOTE

A potential outlier is a data point that is significantly different from the other data points. These special data points may be errors or some kind of abnormality, or they may be a key to understanding the data.

EXAMPLE 2.15

Problem

For the following 13 real estate prices, calculate the *IQR* and determine if any prices are potential outliers. Prices are in dollars.
389,950; 230,500; 158,000; 479,000; 639,000; 114,950; 5,500,000; 387,000; 659,000; 529,000; 575,000; 488,800; 1,095,000

Solution

Order the following data from smallest to largest:
114,950; 158,000; 230,500; 387,000; 389,950; 479,000; 488,800; 529,000; 575,000; 639,000; 659,000; 1,095,000; 5,500,000

M = 488,800

$Q_1 = \frac{230{,}500 + 387{,}000}{2} = 308{,}750$

$Q_3 = \frac{639{,}000 + 659{,}000}{2} = 649{,}000$

IQR = 649,000 – 308,750 = 340,250

(1.5)(IQR) = (1.5)(340,250) = 510,375

Q_1 – (1.5)(IQR) = 308,750 – 510,375 = –201,625

Q_3 + (1.5)(IQR) = 649,000 + 510,375 = 1,159,375

No house price is less than –201,625. However, 5,500,000 is more than 1,159,375. Therefore, 5,500,000 is a potential **outlier**.

TRY IT 2.15

For the 11 salaries, calculate the *IQR* and determine if any salaries are outliers. The following salaries are in dollars.

$33,000; $64,500; $28,000; $54,000; $72,000; $68,500; $69,000; $42,000; $54,000; $120,000; $40,500

In the example above, you just saw the calculation of the median, first quartile, and third quartile. These three values are part of the five number summary. The other two values are the minimum value (or min) and the maximum value (or max). The five number summary is used to create a box plot.

TRY IT 2.15

Find the interquartile range for the following two data sets and compare them.

Test Scores for Class *A*:
69, 96, 81, 79, 65, 76, 83, 99, 89, 67, 90, 77, 85, 98, 66, 91, 77, 69, 80, 94
Test Scores for Class *B*:
90, 72, 80, 92, 90, 97, 92, 75, 79, 68, 70, 80, 99, 95, 78, 73, 71, 68, 95, 100

EXAMPLE 2.16

Fifty statistics students were asked how much sleep they get per school night (rounded to the nearest hour). The results were as follows:

Amount of Sleep per School Night (Hours)	Frequency	Relative Frequency	Cumulative Relative Frequency
4	2	.04	.04
5	5	.10	.14
6	7	.14	.28
7	12	.24	.52
8	14	.28	.80
9	7	.14	.94
10	3	.06	1.00

Table 2.24

Find the 28^{th} percentile. Notice the .28 in the Cumulative Relative Frequency column. Twenty-eight percent of 50 data values is 14 values. There are 14 values less than the 28^{th} percentile. They include the two 4s, the five 5s, and the seven 6s. The 28^{th} percentile is between the last six and the first seven. **The 28^{th} percentile is 6.5.**

Find the median. Look again at the Cumulative Relative Frequency column and find .52. The median is the 50^{th} percentile or the second quartile. Fifty percent of 50 is 25. There are 25 values less than the median. They include the two 4s, the five 5s, the seven 6s, and 11 of the 7s. The median or 50^{th} percentile is between the 25^{th}, or seven, and 26^{th}, or seven, values. **The median is seven.**

Find the third quartile. The third quartile is the same as the 75^{th} percentile. You can *eyeball* this answer. If you look at the Cumulative Relative Frequency column, you find .52 and .80. When you have all the fours, fives, sixes, and sevens, you have 52 percent of the data. When you include all the 8s, you have 80 percent of the data. **The 75^{th} percentile, then, must be an eight**. Another way to look at the problem is to find 75 percent of 50, which is 37.5, and round up to 38. The third quartile, Q_3, is the 38^{th} value, which is an eight. You can check this answer by counting the values. There are 37 values below the third quartile and 12 values above.

TRY IT 2.16

Forty bus drivers were asked how many hours they spend each day running their routes (rounded to the nearest hour). Find the 65^{th} percentile.

Amount of Time Spent on Route (Hours)	Frequency	Relative Frequency	Cumulative Relative Frequency
2	12	.30	.30
3	14	.35	.65

Table 2.25

Amount of Time Spent on Route (Hours)	Frequency	Relative Frequency	Cumulative Relative Frequency
4	10	.25	.90
5	4	.10	1.00

Table 2.25

EXAMPLE 2.17

Problem

Using Table 2.24:

a. Find the 80th percentile.
b. Find the 90th percentile.
c. Find the first quartile. What is another name for the first quartile?

Solution

Using the data from the frequency table, we have the following:

a. The 80th percentile is between the last eight and the first nine in the table (between the 40th and 41st values). Therefore, we need to take the mean of the 40th an 41st values. The 80th percentile $= \frac{8+9}{2} = 8.5$.
b. The 90th percentile will be the 45th data value (location is 0.90(50) = 45), and the 45th data value is nine.
c. Q_1 is also the 25th percentile. The 25th percentile location calculation: P_{25} = .25(50) = 12.5 ≈ 13, the 13th data value. Thus, the 25th percentile is six.

TRY IT 2.17

Refer to Table 2.25. Find the third quartile. What is another name for the third quartile?

COLLABORATIVE EXERCISE

Your instructor or a member of the class will ask everyone in class how many sweaters he or she owns. Answer the following questions:

1. How many students were surveyed?
2. What kind of sampling did you do?
3. Construct two different histograms. For each, starting value = _____ and ending value = _____.
4. Find the median, first quartile, and third quartile.
5. Construct a table of the data to find the following:
   a. The 10th percentile
   b. The 70th percentile
   c. The percentage of students who own fewer than four sweaters

A Formula for Finding the *k*th Percentile

If you were to do a little research, you would find several formulas for calculating the k^{th} percentile. Here is one of them.

k = the k^{th} percentile. It may or may not be part of the data.

i = the index (ranking or position of a data value)

n = the total number of data

- Order the data from smallest to largest.
- Calculate $i = \frac{k}{100}(n + 1)$.
- If i is an integer, then the k^{th} percentile is the data value in the i^{th} position in the ordered set of data.
- If i is not an integer, then round i up and round i down to the nearest integers. Average the two data values in these two positions in the ordered data set. The formula and calculation are easier to understand in an example.

EXAMPLE 2.18

Problem

Listed are 29 ages for Academy Award-winning best actors *in order from smallest to largest:*
18, 21, 22, 25, 26, 27, 29, 30, 31, 33, 36, 37, 41, 42, 47, 52, 55, 57, 58, 62, 64, 67, 69, 71, 72, 73, 74, 76, 77

a. Find the 70^{th} percentile.
b. Find the 83^{rd} percentile.

Solution

a. k = 70
i = the index
n = 29

$i = \frac{k}{100}(n + 1) = (\frac{70}{100})(29 + 1) = 21$. This equation tells us that i, or the position of the data value in the data set, is 21. So, we will count over to the 21^{st} position, which shows a data value of 64.

b. $k = 83^{rd}$ percentile
i = the index
n = 29

$i = \frac{k}{100}(n + 1) = (\frac{83}{100})(29 + 1) = 24.9$, which is **not** an integer. Round it down to 24 and up to 25. The age in the 24^{th} position is 71, and the age in the 25^{th} position is 72. Average 71 and 72. The 83^{rd} percentile is 71.5 years.

TRY IT 2.18

Listed are 29 ages for Academy Award-winning best actors *in order from smallest to largest:*

18, 21, 22, 25, 26, 27, 29, 30, 31, 33, 36, 37, 41, 42, 47, 52, 55, 57, 58, 62, 64, 67, 69, 71, 72, 73, 74, 76, 77
Calculate the 20^{th} percentile and the 55^{th} percentile.

NOTE
You can calculate percentiles using calculators and computers. There are a variety of online calculators.

A Formula for Finding the Percentile of a Value in a Data Set

- Order the data from smallest to largest.
- x = the number of data values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile.
- y = the number of data values equal to the data value for which you want to find the percentile.
- n = the total number of data.
- Calculate $\frac{x+.5y}{n}(100)$. Then round to the nearest integer.

EXAMPLE 2.19

Problem

Listed are 29 ages for Academy Award-winning best actors *in order from smallest to largest:*
18, 21, 22, 25, 26, 27, 29, 30, 31, 33, 36, 37, 41, 42, 47, 52, 55, 57, 58, 62, 64, 67, 69, 71, 72, 73, 74, 76, 77

a. Find the percentile for 58.

b. Find the percentile for 25.

Solution

a. Counting from the bottom of the list, there are 18 data values less than 58. There is one value of 58. $x = 18$ and $y = 1$. $\frac{x+.5y}{n}(100) = \frac{18+.5(1)}{29}(100) = 63.80$. Fifty-eight is the 64^{th} percentile.

b. Counting from the bottom of the list, there are three data values less than 25. There is one value of 25. $x = 3$ and $y = 1$. $\frac{x+.5y}{n}(100) = \frac{3+.5(1)}{29}(100) = 12.07$. Twenty-five is the 12^{th} percentile.

TRY IT 2.19

Listed are 30 ages for Academy Award-winning best actors *in order from smallest to largest:*

18, 21, 22, 25, 26, 27, 29, 30, 31, 31, 33, 36, 37, 41, 42, 47, 52, 55, 57, 58, 62, 64, 67, 69, 71, 72, 73, 74, 76, 77
Find the percentiles for 47 and 31.

Interpreting Percentiles, Quartiles, and Median

A percentile indicates the relative standing of a data value when data are sorted into numerical order from smallest to largest. Percentages of data values are less than or equal to the *p*th percentile. For example, 15 percent of data values are less than or equal to the 15^{th} percentile.

- Low percentiles always correspond to lower data values.
- High percentiles always correspond to higher data values.

A percentile may or may not correspond to a value judgment about whether it is *good* or *bad*. The interpretation of whether a certain percentile is *good* or *bad* depends on the context of the situation to which the data apply. In some situations, a low percentile would be considered *good;* in other contexts a high percentile might be considered *good*. In many situations, there is no value judgment that applies. A high percentile on a standardized test is considered good, while a lower percentile on body mass index might be considered good. A percentile associated with a person's height doesn't carry any value judgment.

Understanding how to interpret percentiles properly is important not only when describing data, but also when calculating probabilities in later chapters of this text.

Guideline

When writing the interpretation of a percentile in the context of the given data, make sure the sentence contains the following information:

- Information about the context of the situation being considered
- The data value (value of the variable) that represents the percentile
- The percentage of individuals or items with data values below the percentile
- The percentage of individuals or items with data values above the percentile

EXAMPLE 2.20

Problem

On a timed math test, the first quartile for time it took to finish the exam was 35 minutes. Interpret the first quartile in the context of this situation.

Solution

- Twenty-five percent of students finished the exam in 35 minutes or less.
- Seventy-five percent of students finished the exam in 35 minutes or more.
- A low percentile could be considered good, as finishing more quickly on a timed exam is desirable. If you take too

long, you might not be able to finish.

TRY IT 2.20

For the 100-meter dash, the third quartile for times for finishing the race was 11.5 seconds. Interpret the third quartile in the context of the situation.

EXAMPLE 2.21

Problem

On a 20-question math test, the 70th percentile for number of correct answers was 16. Interpret the 70th percentile in the context of this situation.

Solution

- Seventy percent of students answered 16 or fewer questions correctly.
- Thirty percent of students answered 16 or more questions correctly.
- A higher percentile could be considered good, as answering more questions correctly is desirable.

TRY IT 2.21

On a 60-point written assignment, the 80th percentile for the number of points earned was 49. Interpret the 80th percentile in the context of this situation.

EXAMPLE 2.22

Problem

At a high school, it was found that the 30th percentile of number of hours that students spend studying per week is seven hours. Interpret the 30th percentile in the context of this situation.

Solution

- Thirty percent of students study seven or fewer hours per week.
- Seventy percent of students study seven or more hours per week.
- In this example, there is not necessarily a *good* or *bad* value judgment associated with a higher or lower percentile, since the time a student studies per week is dependent on his/her needs.

TRY IT 2.22

During a season, the 40th percentile for points scored per player in a game is eight. Interpret the 40th percentile in the context of this situation.

EXAMPLE 2.23

A middle school is applying for a grant that will be used to add fitness equipment to the gym. The principal surveyed 15 anonymous students to determine how many minutes a day the students spend exercising. The results from the 15 anonymous students are shown:

0 minutes, 40 minutes, 60 minutes, 30 minutes, 60 minutes,

10 minutes, 45 minutes, 30 minutes, 300 minutes, 90 minutes,

30 minutes, 120 minutes, 60 minutes, 0 minutes, 20 minutes

Find the five values that make up the five number summary.

Min = 0

Q_1 = 20

Med = 40

Q_3 = 60

Max = 300

Listing the data in ascending order gives the following:

Figure 2.12

The minimum value is 0.

The maximum value is 300.

Since there are an odd number of data values, the median is the middle value of this data set as it is arranged in ascending order, or 40.

The first quartile is the median of the lower half of the scores and does not include the median. The lower half has seven data values; the median of the lower half will equal the middle value of the lower half, or 20.

The third quartile is the median of the upper half of the scores and does not include the median. The upper half also has seven data values; so the median of the upper half will equal the middle value of the upper half, or 60.

If you were the principal, would you be justified in purchasing new fitness equipment? Since 75 percent of the students exercise for 60 minutes or less daily, and since the *IQR* is 40 minutes (60 – 20 = 40), we know that half of the students surveyed exercise between 20 minutes and 60 minutes daily. This seems a reasonable amount of time spent exercising, so the principal would be justified in purchasing the new equipment.

However, the principal needs to be careful. The value 300 appears to be a potential outlier.

$Q_3 + 1.5(IQR) = 60 + (1.5)(40) = 120$.

The value 300 is greater than 120, so it is a potential outlier. If we delete it and calculate the five values, we get the following values:

Min = 0
Q_1 = 20
Q_3 = 60
Max = 120

We still have 75 percent of the students exercising for 60 minutes or less daily and half of the students exercising between 20 and 60 minutes a day. However, 15 students is a small sample, and the principal should survey more students to be sure of his survey results.

2.4 Box Plots

Box plots, also called **box-and-whisker plots** or **box-whisker plots**, give a good graphical image of the concentration of the data. They also show how far the extreme values are from most of the data. As mentioned previously, a box plot is constructed from five values: the minimum value, the first quartile, the median, the third quartile, and the maximum value. We use these values to compare how close other data values are to them.

To construct a box plot, use a horizontal or vertical number line and a rectangular box. The smallest and largest data values label the endpoints of the axis. The first quartile marks one end of the box, and the third quartile marks the other end of the box. Approximately **the middle 50 percent of the data fall inside the box.** The *whiskers* extend from the

ends of the box to the smallest and largest data values. A box plot easily shows the range of a data set, which is the difference between the largest and smallest data values (or the difference between the maximum and minimum). Unless the median, first quartile, and third quartile are the same value, the median will lie inside the box or between the first and third quartiles. The box plot gives a good, quick picture of the data.

NOTE

You may encounter box-and-whisker plots that have dots marking outlier values. In those cases, the whiskers are not extending to the minimum and maximum values.

Consider, again, this data set:

1, 1, 2, 2, 4, 6, 6.8, 7.2, 8, 8.3, 9, 10, 10, 11.5

The first quartile is two, the median is seven, and the third quartile is nine. The smallest value is one, and the largest value is 11.5. The following image shows the constructed box plot.

NOTE

See the calculator instructions on the TI website (https://education.ti.com/en/professional-development/webinars-and-tutorials/technology-tutorials) or in the appendix.

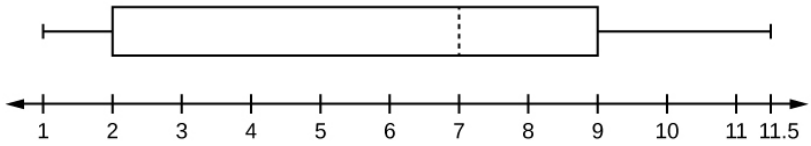


Figure 2.13

The two whiskers extend from the first quartile to the smallest value and from the third quartile to the largest value. The median is shown with a dashed line.

NOTE

It is important to start a box plot with a **scaled number line**. Otherwise, the box plot may not be useful.

EXAMPLE 2.24

The following data are the heights of 40 students in a statistics class:

59, 60, 61, 62, 62, 63, 63, 64, 64, 64, 65, 65, 65, 65, 65, 65, 65, 65, 65, 66, 66, 67, 67, 68, 68, 69, 70, 70, 70, 70, 70, 71, 71, 72, 72, 73, 74, 74, 75, 77.

Construct a box plot with the following properties. Calculator instructions for finding the five number summary follow this example:

- Minimum value = 59
- Maximum value = 77
- Q_1: First quartile = 64.5
- Q_2: Second quartile or median = 66
- Q_3: Third quartile = 70

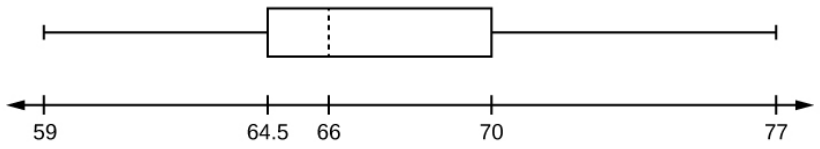


Figure 2.14

a. Each quarter has approximately 25 percent of the data.
b. The spreads of the four quarters are 64.5 – 59 = 5.5 (first quarter), 66 – 64.5 = 1.5 (second quarter), 70 – 66 = 4 (third quarter), and 77 – 70 = 7 (fourth quarter). So, the second quarter has the smallest spread, and the fourth quarter has the largest spread.
c. Range = maximum value – minimum value = 77 – 59 = 18.
d. Interquartile Range: $IQR = Q3 - Q1 = 70 - 64.5 = 5.5$.
e. The interval 59–65 has more than 25 percent of the data, so it has more data in it than the interval 66–70, which has 25 percent of the data.
f. The middle 50 percent (middle half) of the data has a range of 5.5 inches.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

To find the minimum, maximum, and quartiles:

Enter data into the list editor (Pres STAT 1:EDIT). If you need to clear the list, arrow up to the name L1, press CLEAR, and then arrow down.

Put the data values into the list L1.

Press STAT and arrow to CALC. Press 1:1-VarStats. Enter L1.

Press ENTER.

Use the down and up arrow keys to scroll.

Smallest value = 59.

Largest value = 77.

Q_1: First quartile = 64.5.

Q_2: Second quartile or median = 66.

Q_3: Third quartile = 70.

To construct the box plot:

Press 4:Plotsoff. Press ENTER.

Arrow down and then use the right arrow key to go to the fifth picture, which is the box plot. Press ENTER.

Arrow down to Xlist: Press 2^{nd} 1 for L1.

Arrow down to Freq: Press ALPHA. Press 1.

Press Zoom. Press 9: ZoomStat.

Press TRACE and use the arrow keys to examine the box plot.

TRY IT 2.24

The following data are the number of pages in 40 books on a shelf. Construct a box plot using a graphing calculator

and state the interquartile range.

136, 140, 178, 190, 205, 215, 217, 218, 232, 234, 240, 255, 270, 275, 290, 301, 303, 315, 317, 318, 326, 333, 343, 349, 360, 369, 377, 388, 391, 392, 398, 400, 402, 405, 408, 422, 429, 450, 475, 512

For some sets of data, some of the largest value, smallest value, first quartile, median, and third quartile may be the same. For instance, you might have a data set in which the median and the third quartile are the same. In this case, the diagram would not have a dotted line inside the box displaying the median. The right side of the box would display both the third quartile and the median. For example, if the smallest value and the first quartile were both one, the median and the third quartile were both five, and the largest value was seven, the box plot would look like the following:

Figure 2.15

In this case, at least 25 percent of the values are equal to one. Twenty-five percent of the values are between one and five, inclusive. At least 25 percent of the values are equal to five. The top 25 percent of the values fall between five and seven, inclusive.

EXAMPLE 2.25

Test scores for Mr. Ramirez's class held during the day are as follows:

99, 56, 78, 55.5, 32, 90, 80, 81, 56, 59, 45, 77, 84.5, 84, 70, 72, 68, 32, 79, 90.

Test scores for Ms. Park's class held during the evening are as follows:

98, 78, 68, 83, 81, 89, 88, 76, 65, 45, 98, 90, 80, 84.5, 85, 79, 78, 98, 90, 79, 81, 25.5.

Problem

a. Find the smallest and largest values, the median, and the first and third quartile for Mr. Ramirez's class.
b. Find the smallest and largest values, the median, and the first and third quartile for Ms. Park's class.
c. For each data set, what percentage of the data is between the smallest value and the first quartile? the first quartile and the median? the median and the third quartile? the third quartile and the largest value? What percentage of the data is between the first quartile and the largest value?
d. Create a box plot for each set of data. Use one number line for both box plots.
e. Which box plot has the widest spread for the middle 50 percent of the data,the data between the first and third quartiles? What does this mean for that set of data in comparison to the other set of data?

Solution

a. Min = 32
Q_1 = 56
M = 74.5
Q_3 = 82.5
Max = 99

b. Min = 25.5
Q_1 = 78
M = 81
Q_3 = 89
Max = 98

c. Mr. Ramirez's class: There are six data values ranging from 32 to 56: 30 percent. There are six data values ranging from 56 to 74.5: 30 percent. There are five data values ranging from 74.5 to 82.5: 25 percent. There are five data values ranging from 82.5 to 99: 25 percent. There are 16 data values between the first quartile, 56, and the largest value, 99: 80 percent. Ms. Park's class: There are seven data values ranging from 25.5 to 78: 32 percent. There are

seven data values ranging from 78 through 32 percent. There are seven data values ranging 81 to 89: 32 percent. There are six data values ranging from 89 to 98: 27 percent. There are 17 values between the first quartile, 78, and the largest value, 98: 77 percent.

d.

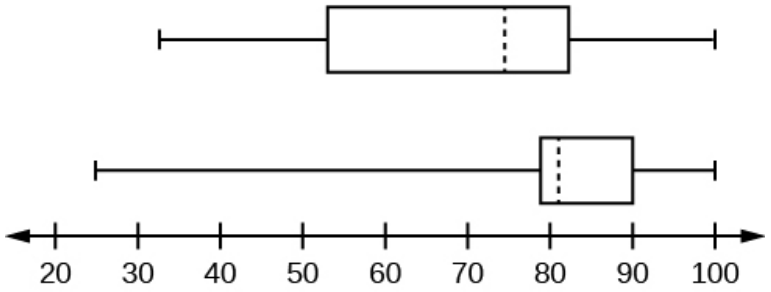


Figure 2.16

e. The first data set has the wider spread for the middle 50 percent of the data. The *IQR* for the first data set is greater than the *IQR* for the second set. This means that there is more variability in the middle 50 percent of the first data set.

TRY IT 2.25

The following data set shows the heights in inches for the boys in a class of 40 students:

66, 66, 67, 67, 68, 68, 68, 68, 68, 69, 69, 69, 70, 71, 72, 72, 72, 73, 73, 74.
The following data set shows the heights in inches for the girls in a class of 40 students:
61 61, 62, 62, 63, 63, 63, 65, 65, 65, 66, 66, 66, 67, 68, 68, 68, 69, 69, 69.
Construct a box plot using a graphing calculator for each data set, and state which box plot has the wider spread for the middle 50 percent of the data.

EXAMPLE 2.26

Graph a box-and-whisker plot for the following data values shown:

10, 10, 10, 15, 35, 75, 90, 95, 100, 175, 420, 490, 515, 515, 790

The five numbers used to create a box-and-whisker plot are as follows:

Min: 10
Q_1: 15
Med: 95
Q_3: 490
Max: 790

The following graph shows the box-and-whisker plot.

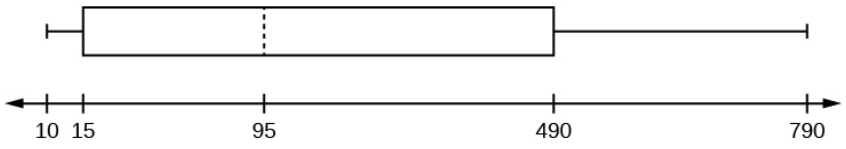


Figure 2.17

TRY IT 2.26

Follow the steps you used to graph a box-and-whisker plot for the data values shown:

0, 5, 5, 15, 30, 30, 45, 50, 50, 60, 75, 110, 140, 240, 330

2.5 Measures of the Center of the Data

The *center* of a data set is also a way of describing location. The two most widely used measures of the *center* of the data are the **mean** (average) and the **median**. To calculate the **mean weight** of 50 people, add the 50 weights together and divide by 50. To find the **median weight** of the 50 people, order the data and find the number that splits the data into two equal parts. The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

NOTE

The words *mean* and *average* are often used interchangeably. The substitution of one word for the other is common practice. The technical term is *arithmetic mean* and *average* is technically a center location. However, in practice among non statisticians, *average* is commonly accepted for *arithmetic mean*.

When each value in the data set is not unique, the mean can be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. The letter used to represent the **sample mean** is an *x* with a bar over it (pronounced "x bar"): $\bar{x}$. The sample mean is a statistic.

The Greek letter μ (pronounced "mew") represents the **population mean**. The population mean is a parameter. One of the requirements for the **sample mean** to be a good estimate of the **population mean** is for the sample taken to be truly random.

To see that both ways of calculating the mean are the same, consider the following sample:
1, 1, 1, 2, 2, 3, 4, 4, 4, 4, 4

$$\bar{x} = \frac{1+1+1+2+2+3+4+4+4+4+4}{11} = 2.7$$

$$\bar{x} = \frac{3(1)+2(2)+1(3)+5(4)}{11} = 2.7.$$

In the second example, the frequencies are 3(1) + 2(2) + 1(3) + 5(4).

You can quickly find the location of the median by using the expression $\frac{n+1}{2}$.

The letter *n* is the total number of data values in the sample. As discussed earlier, if *n* is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If *n* is an even number, the median is equal to the two middle values added together and divided by two after the data have been ordered. For example, if the total number of data values is 97, then $\frac{n+1}{2} = \frac{97+1}{2}$ = 49. The median is the 49^{th} value in the ordered data. If the total number of data values is 100, then $\frac{n+1}{2} = \frac{100+1}{2}$ = 50.5. The median occurs midway between the 50^{th} and 51^{st} values. The location of the median and the value of the median are **not** the same. The uppercase letter *M* is often used to represent the median. The next example illustrates the location of the median and the value of the median.

EXAMPLE 2.27

Problem

Data indicating the number of months a patient with a specific disease lives after taking a new antibody drug are as follows (smallest to largest):
3, 4, 8, 8, 10, 11, 12, 13, 14, 15, 15, 16, 16, 17, 17, 18, 21, 22, 22, 24, 24, 25, 26, 26, 27, 27, 29, 29, 31, 32, 33, 33, 34, 34, 35, 37, 40, 44, 44, 47
Calculate the mean and the median.

Solution

The calculation for the mean is

$$\bar{x} = [3 + 4 + (8)(2) + 10 + 11 + 12 + 13 + 14 + (15)(2) + (16)(2) + (17)(2) + 18 + 21 + (22)(2) + (24)(2) + 25 + (26)(2) + (27)(2) + (29)(2) + 31 + 32 + (33)(2) + (34)(2) + 35 + 37 + 40 + (44)(2) + 47]/40 = 23.6.$$

To find the median, *M*, first use the formula for the location. The location is $\frac{n+1}{2} = \frac{40+1}{2} = 20.5$.

Start from the smallest value and count up; the median is located between the 20th and 21st values (the two 24s):

3, 4, 8, 8, 10, 11, 12, 13, 14, 15, 15, 16, 16, 17, 17, 18, 21, 22, 22, 24, 24, 25, 26, 26, 27, 27, 29, 29, 31, 32, 33, 33, 34, 34, 35, 37, 40, 44, 44, 47

$$M = \frac{24+24}{2} = 24$$

USING THE TI-83, 83+, 84, 84+ CALCULATOR

To find the mean and the median:

Clear list L1. Pres STAT 4:ClrList. Enter 2nd 1 for list L1. Press ENTER.

Enter data into the list editor. Press STAT 1:EDIT.

Put the data values into list L1.

Press STAT and arrow to CALC. Press 1:1-VarStats. Press 2nd 1 for L1 and then ENTER.

Press the down and up arrow keys to scroll.

$\bar{x}$ = 23.6, *M* = 24

TRY IT 2.27

The following data show the number of months patients typically wait on a transplant list before getting surgery. The data are ordered from smallest to largest. Calculate the mean and median.

3, 4, 5, 7, 7, 7, 7, 8, 8, 9, 9, 10, 10, 10, 10, 10, 11, 12, 12, 13, 14, 14, 15, 15, 17, 17, 18, 19, 19, 19, 21, 21, 22, 22, 23, 24, 24, 24, 24

EXAMPLE 2.28

Problem

Suppose that in a small town of 50 people, one person earns $5,000,000 per year and the other 49 each earn $30,000. Which is the better measure of the *center*: the mean or the median?

Solution

$$\bar{x} = \frac{5{,}000{,}000 + 49(30{,}000)}{50} = 129{,}400$$

M = 30,000

There are 49 people who earn $30,000 and one person who earns $5,000,000.

The median is a better measure of the *center* than the mean because 49 of the values are 30,000 and one is 5,000,000. The 5,000,000 is an outlier. The 30,000 gives us a better sense of the middle of the data.

TRY IT 2.28

In a sample of 60 households, one house is worth $2,500,000. Half of the rest are worth $280,000, and all the others are worth $315,000. Which is the better measure of the *center*: the mean or the median?

Another measure of the center is the mode. The **mode** is the most frequent value. There can be more than one mode in a data set as long as those values have the same frequency and that frequency is the highest. A data set with two modes is called bimodal.

EXAMPLE 2.29

Statistics exam scores for 20 students are as follows:

50, 53, 59, 59, 63, 63, 72, 72, 72, 72, 72, 76, 78, 81, 83, 84, 84, 84, 90, 93

Problem
Find the mode.

Solution
The most frequent score is 72, which occurs five times. Mode = 72.

TRY IT 2.29

The number of books checked out from the library by 25 students are as follows:

0, 0, 0, 1, 2, 3, 3, 4, 4, 5, 5, 7, 7, 7, 7, 8, 8, 8, 9, 10, 10, 11, 11, 12, 12
Find the mode.

EXAMPLE 2.30

Five real estate exam scores are 430, 430, 480, 480, 495. The data set is bimodal because the scores 430 and 480 each occur twice.

When is the mode the best measure of the *center*? Consider a weight loss program that advertises a mean weight loss of six pounds the first week of the program. The mode might indicate that most people lose two pounds the first week, making the program less appealing.

NOTE

The mode can be calculated for qualitative data as well as for quantitative data. For example, if the data set is red, red, red, green, green, yellow, purple, black, blue, the mode is red.

Statistical software will easily calculate the mean, the median, and the mode. Some graphing calculators can also make these calculations. In the real world, people make these calculations using software.

TRY IT 2.30

Five credit scores are 680, 680, 700, 720, 720. The data set is bimodal because the scores 680 and 720 each occur twice. Consider the annual earnings of workers at a factory. The mode is $25,000 and occurs 150 times out of 301. The median is $50,000, and the mean is $47,500. What would be the best measure of the *center*?

The Law of Large Numbers and the Mean

The Law of Large Numbers says that if you take samples of larger and larger size from any population, then the mean $\overline{x}$ of the sample is very likely to get closer and closer to μ. This law is discussed in more detail later in the text.

Sampling Distributions and Statistic of a Sampling Distribution

You can think of a **sampling distribution** as a **relative frequency distribution** with a great many samples. See **Chapter 1: Sampling and Data** for a review of relative frequency. Suppose 30 randomly selected students were asked the number of movies they watched the previous week. The results are in the **relative frequency table** shown below.

Number of Movies	Relative Frequency
0	$\frac{5}{30}$
1	$\frac{15}{30}$
2	$\frac{6}{30}$
3	$\frac{3}{30}$
4	$\frac{1}{30}$

Table 2.26

A relative frequency distribution includes the relative frequencies of a number of samples.

Recall that a **statistic** is a number calculated from a sample. Statistic examples include the mean, the median, and the mode as well as others. The sample mean $\overline{x}$ is an example of a statistic that estimates the population mean μ.

Calculating the Mean of Grouped Frequency Tables

When only grouped data is available, you do not know the individual data values (we know only intervals and interval frequencies); therefore, you cannot compute an exact mean for the data set. What we must do is estimate the actual mean by calculating the mean of a frequency table. A frequency table is a data representation in which grouped data is displayed along with the corresponding frequencies. To calculate the mean from a grouped frequency table, we can apply the basic definition of mean: $mean = \frac{data\ sum}{number\ of\ data\ values}$. We simply need to modify the definition to fit within the restrictions of a frequency table.

Since we do not know the individual data values, we can instead find the midpoint of each interval. The midpoint is $\frac{lower\ boundary+upper\ boundary}{2}$. We can now modify the mean definition to be $Mean\ of\ Frequency\ Table = \frac{\sum fm}{\sum f}$, where f = the frequency of the interval, m = the midpoint of the interval, and sigma ($\sum$) is read as "sigma" and means to sum up. So this formula says that we will sum the products of each midpoint and the corresponding frequency and divide by the sum of all of the frequencies.

EXAMPLE 2.31

Problem

A frequency table displaying Professor Blount's last statistic test is shown. Find the best estimate of the class mean.

Grade Interval	Number of Students
50–56.5	1
56.5–62.5	0
62.5–68.5	4
68.5–74.5	4
74.5–80.5	2
80.5–86.5	3
86.5–92.5	4
92.5–98.5	1

Table 2.27

Solution

- Find the midpoints for all intervals.

Grade Interval	Midpoint
50–56.5	53.25
56.5–62.5	59.5
62.5–68.5	65.5
68.5–74.5	71.5
74.5–80.5	77.5
80.5–86.5	83.5
86.5–92.5	89.5
92.5–98.5	95.5

Table 2.28

- Calculate the sum of the product of each interval frequency and midpoint. $\sum fm$

  $53.25(1) + 59.5(0) + 65.5(4) + 71.5(4) + 77.5(2) + 83.5(3) + 89.5(4) + 95.5(1) = 1460.25$
- $\mu = \frac{\sum fm}{\sum f} = \frac{1460.25}{19} = 76.86$

TRY IT 2.31

Maris conducted a study on the effect that playing video games has on memory recall. As part of her study, she compiled the following data:

Hours Teenagers Spend on Video Games	Number of Teenagers
0–3.5	3
3.5–7.5	7
7.5–11.5	12
11.5–15.5	7
15.5–19.5	9

Table 2.29

What is the best estimate for the mean number of hours spent playing video games?

2.6 Skewness and the Mean, Median, and Mode

Consider the following data set:
4, 5, 6, 6, 6, 7, 7, 7, 7, 7, 7, 8, 8, 8, 9, 10

This data set can be represented by the following histogram. Each interval has width 1, and each value is located in the middle of an interval.

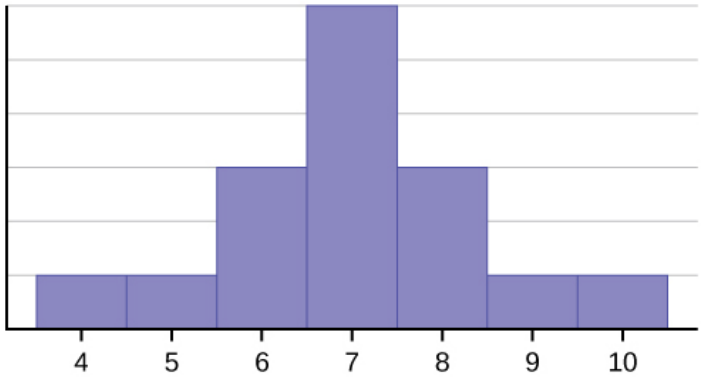


Figure 2.18

The histogram displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each seven for these data. **In a perfectly symmetrical distribution, the mean and the median are the same.** This example has one mode (unimodal), and the mode is the same as the mean and median. In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.

The histogram for the data: 4, 5, 6, 6, 6, 7, 7, 7, 7, 8 is not symmetrical. The right-hand side seems *chopped off* compared to the left-hand side. A distribution of this type is called **skewed to the left** because it is pulled out to the left. A skewed left distribution has more high values.

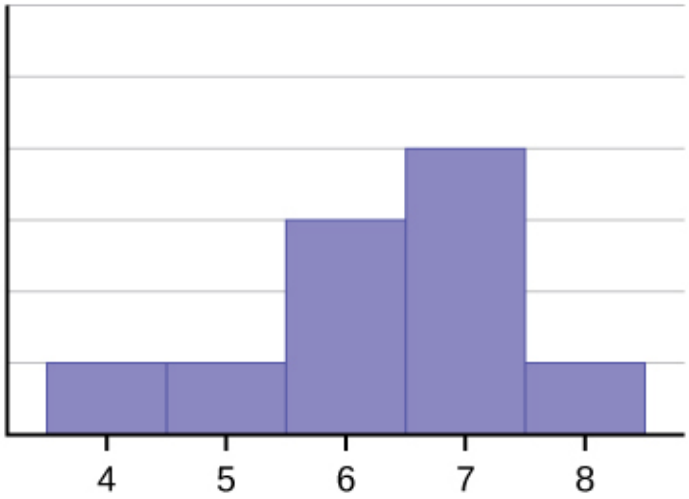


Figure 2.19

The mean is 6.3, the median is 6.5, and the mode is seven. **Notice that the mean is less than the median, and they are both less than the mode**. The mean and the median both reflect the skewing, but the mean reflects it more so. The mean is pulled toward the tail in a skewed distribution.

The histogram for the data: 6, 7, 7, 7, 7, 8, 8, 8, 9, 10 is also not symmetrical. It is **skewed to the right**. A skewed right distribution has more low values.

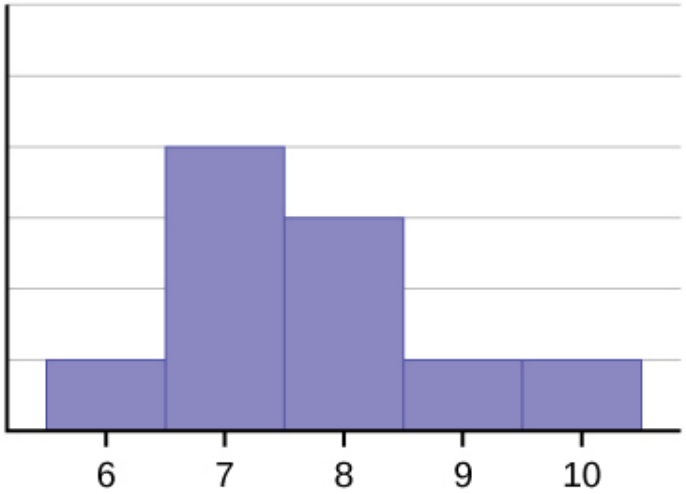


Figure 2.20

The mean is 7.7, the median is 7.5, and the mode is seven. Of the three statistics, **the mean is the largest, while the mode is the smallest**. Again, the mean reflects the skewing the most.

To summarize, generally if the distribution of data is skewed to the left, the mean is less than the median, which is often less than the mode. If the distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.

Skewness and symmetry become important when we discuss probability distributions in later chapters.

EXAMPLE 2.32

Problem

Statistics are used to compare and sometimes identify authors. The following lists show a simple random sample that compares the letter counts for three authors.

Terry: 7, 9, 3, 3, 3, 4, 1, 3, 2, 2

Davis: 3, 3, 3, 4, 1, 4, 3, 2, 3, 1

Maris: 2, 3, 4, 4, 4, 6, 6, 6, 8, 3

a. Make a dot plot for the three authors and compare the shapes.
b. Calculate the mean for each.
c. Calculate the median for each.
d. Describe any pattern you notice between the shape and the measures of center.

Solution

a.

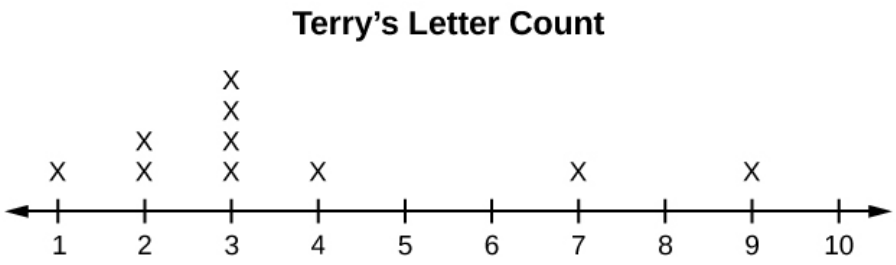


Figure 2.21 Terry's distribution has a right (positive) skew.

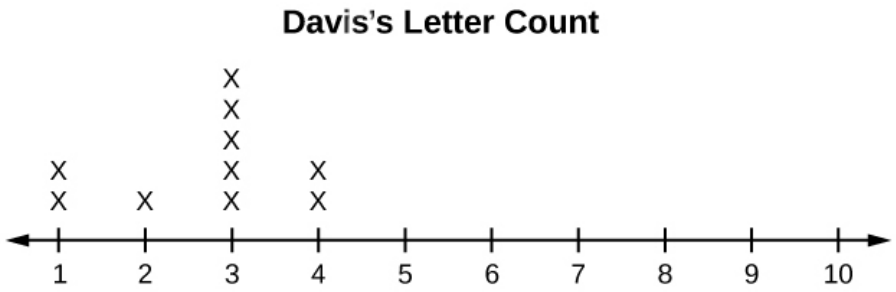


Figure 2.22 Davis's distribution has a left (negative) skew.

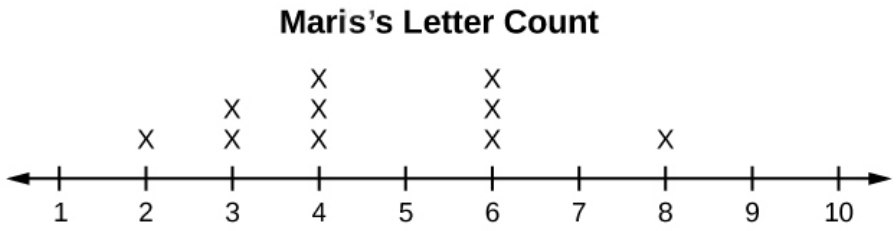


Figure 2.23 Maris's distribution is symmetrically shaped.

b. Terry's mean is 3.7, Davis's mean is 2.7, and Maris's mean is 4.6.
c. Terry's median is 3, Davis's median is 3, and Maris's median is four. It would be helpful to manually calculate these descriptive statistics, using the given data sets and then compare to the graphs.
d. It appears that the median is always closest to the high point (the mode), while the mean tends to be farther out on the tail. In a symmetrical distribution, the mean and the median are both centrally located close to the high point of the distribution.

TRY IT 2.32

Discuss the mean, median, and mode for each of the following problems. Is there a pattern between the shape and measure of the center?

a.

Figure 2.24

b.

The Ages at Which Former U.S. Presidents Died	
4	6 9
5	3 6 7 7 7 8
6	0 0 3 3 4 4 5 6 7 7 7 8
7	0 1 1 2 3 4 7 8 8 9
8	0 1 3 5 8
9	0 0 3 3
Key: 8|0 means 80.	

Table 2.30

c.

Figure 2.25

2.7 Measures of the Spread of the Data

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation. The **standard deviation** is a number that measures how far data values are from their mean.

The standard deviation

- provides a numerical measure of the overall amount of variation in a data set and
- can be used to determine whether a particular data value is close to or far from the mean.

The standard deviation provides a measure of the overall variation in a data set.

The standard deviation is always positive or zero. The standard deviation is small when all the data are concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying the amount of time customers wait in line at the checkout at Supermarket A and Supermarket B. The average wait time at both supermarkets is five minutes. At Supermarket A, the standard deviation for the wait time is two minutes; at Supermarket B, the standard deviation for the wait time is four minutes.

Because Supermarket B has a higher standard deviation, we know that there is more variation in the wait times at Supermarket B. Overall, wait times at Supermarket B are more spread out from the average whereas wait times at Supermarket A are more concentrated near the average.

The standard deviation can be used to determine whether a data value is close to or far from the mean.

Suppose that both Rosa and Binh shop at Supermarket A. Rosa waits at the checkout counter for seven minutes, and Binh waits for one minute. At Supermarket A, the mean waiting time is five minutes, and the standard deviation is two minutes. The standard deviation can be used to determine whether a data value is close to or far from the mean. A *z*-score is a standardized score that lets us compare data sets. It tells us how many standard deviations a data value is from the mean and is calculated as the ratio of the difference in a particular score and the population mean to the population standard deviation.

We can use the given information to create the table below.

Supermarket	Population Standard Deviation, σ	Individual Score, x	Population Mean, μ
Supermarket A	2 minutes	7, 1	5
Supermarket B	4 minutes		5

Table 2.31

Since Rosa and Binh only shop at Supermarket A, we can ignore the row for Supermarket B.

We need the values from the first row to determine the number of standard deviations above or below the mean each individual wait time is; we can do so by calculating two different *z*-scores.

Rosa waited for seven minutes, so the *z*-score representing this deviation from the population mean may be calculated as

$$z = \frac{x - \mu}{\sigma} = \frac{7 - 5}{2} = 1. \quad \textbf{2.2}$$

The *z*-score of one tells us that Rosa's wait time is one standard deviation above the mean wait time of five minutes.

Binh waited for one minute, so the *z*-score representing this deviation from the population mean may be calculated as

$$z = \frac{x - \mu}{\sigma} = \frac{1 - 5}{2} = -2. \quad \textbf{2.3}$$

The *z*-score of −2 tells us that Binh's wait time is two standard deviations below the mean wait time of five minutes.

A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be far from the average. Considering data to be far from the mean if they are more than two standard deviations away is more of an approximate *rule of thumb* than a rigid rule. In general, the shape of the distribution of the data affects how much of the data is farther away than two standard deviations. You will learn more about this in later chapters.

The number line may help you understand standard deviation. If we were to put five and seven on a number line, seven is to the right of five. We say, then, that seven is **one** standard deviation to the **right** of five because 5 + (1)(2) = 7.

If one were also part of the data set, then one is **two** standard deviations to the **left** of five because 5 + (–2)(2) = 1.

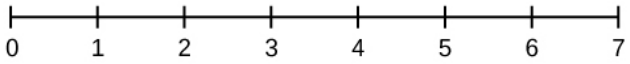


Figure 2.26

- In general, a **value = mean + (#ofSTDEV)(standard deviation)**
- where #ofSTDEVs = the number of standard deviations
- #ofSTDEV does not need to be an integer
- One is **two standard deviations less than the mean** of five because 1 = 5 + (–2)(2).

The equation **value = mean + (#ofSTDEVs)(standard deviation)** can be expressed for a sample and for a population as follows:

- **Sample:** $x = \overline{x} + (\#ofSTDEV)(s)$
- **Population:** $x = \mu + (\#ofSTDEV)(\sigma)$.

The lowercase letter *s* represents the sample standard deviation and the Greek letter σ (lower case) represents the population standard deviation.

The symbol $\overline{x}$ is the sample mean, and the Greek symbol μ is the population mean.

Calculating the Standard Deviation

If *x* is a number, then the difference *x* – *mean* is called its **deviation**. In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation. If the numbers belong to a population, in symbols, a deviation is $x - \mu$. For sample data, in symbols, a deviation is $x - \overline{x}$.

The procedure to calculate the standard deviation depends on whether the numbers are the entire population or are data from a sample. The calculations are similar but not identical. Therefore, the symbol used to represent the standard deviation depends on whether it is calculated from a population or a sample. The lowercase letter *s* represents the sample standard deviation and the Greek letter σ (lowercase sigma) represents the population standard deviation. If the sample has the same characteristics as the population, then *s* should be a good estimate of σ.

To calculate the standard deviation, we need to calculate the variance first. The **variance** is the **average of the squares of the deviations** (the $x - \overline{x}$ values for a sample or the $x - \mu$ values for a population). The symbol σ^2 represents the population variance; the population standard deviation σ is the square root of the population variance. The symbol s^2 represents the sample variance; the sample standard deviation *s* is the square root of the sample variance. You can think of the standard deviation as a special average of the deviations.

If the numbers come from a census of the entire **population** and not a sample, when we calculate the average of the squared deviations to find the variance, we divide by *N*, the number of items in the population. If the data are from a **sample** rather than a population, when we calculate the average of the squared deviations, we divide by ***n* – 1**, one less than the number of items in the sample.

Formulas for the Sample Standard Deviation

- $s = \sqrt{\frac{\Sigma(x-\overline{x})^2}{n-1}}$ or $s = \sqrt{\frac{\Sigma f(x-\overline{x})^2}{n-1}}$
- For the sample standard deviation, the denominator is ***n*–**; that is, the sample size **minus** 1.

Formulas for the Population Standard Deviation

- $\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\Sigma f(x-\mu)^2}{N}}$
- For the population standard deviation, the denominator is *N*, the number of items in the population.

In these formulas, *f* represents the frequency with which a value appears. For example, if a value appears once, *f* is one. If a value appears three times in the data set or population, *f* is three.

Types of Variability in Samples

When researchers study a population, they often use a sample, either for convenience or because it is not possible to access the entire population. *Variability* is the term used to describe the differences that may occur in these outcomes. Common types of variability include the following:

- Observational or measurement variability
- Natural variability

- Induced variability
- Sample variability

Here are some examples to describe each type of variability:

Example 1: Measurement variability

Measurement variability occurs when there are differences in the instruments used to measure or in the people using those instruments. If we are gathering data on how long it takes for a ball to drop from a height by having students measure the time of the drop with a stopwatch, we may experience measurement variability if the two stopwatches used were made by different manufacturers. For example, one stopwatch measures to the nearest second, whereas the other one measures to the nearest tenth of a second. We also may experience measurement variability because two different people are gathering the data. Their reaction times in pressing the button on the stopwatch may differ; thus, the outcomes will vary accordingly. The differences in outcomes may be affected by measurement variability.

Example 2: Natural variability

Natural variability arises from the differences that naturally occur because members of a population differ from each other. For example, if we have two identical corn plants and we expose both plants to the same amount of water and sunlight, they may still grow at different rates simply because they are two different corn plants. The difference in outcomes may be explained by natural variability.

Example 3: Induced variability

Induced variability is the counterpart to natural variability. This occurs because we have artificially induced an element of variation that, by definition, was not present naturally. For example, we assign people to two different groups to study memory, and we induce a variable in one group by limiting the amount of sleep they get. The difference in outcomes may be affected by induced variability.

Example 4: Sample variability

Sample variability occurs when multiple random samples are taken from the same population. For example, if I conduct four surveys of 50 people randomly selected from a given population, the differences in outcomes may be affected by sample variability.

Sampling Variability of a Statistic

The statistic of a sampling distribution was discussed in Descriptive Statistics: Measures of the Center of the Data. How much the statistic varies from one sample to another is known as the **sampling variability of a statistic**. You typically measure the sampling variability of a statistic by its standard error. The **standard error of the mean** is an example of a standard error. The standard error is the standard deviation of the sampling distribution. In other words, it is the average standard deviation that results from repeated sampling. You will cover the standard error of the mean in the chapter The Central Limit Theorem (not now). The notation for the standard error of the mean is $\frac{\sigma}{\sqrt{n}}$, where σ is the standard deviation of the population and n is the size of the sample.

NOTE
In practice, use a calculator or computer software to calculate the standard deviation. If you are using a TI-83, 83+, or 84+ calculator, you need to select the appropriate standard deviation σ_x or s_x from the summary statistics. We will concentrate on using and interpreting the information that the standard deviation gives us. However, you should study the following step-by-step example to help you understand how the standard deviation measures variation from the mean. The calculator instructions appear at the end of this example.

EXAMPLE 2.33

In a fifth-grade class, the teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a SAMPLE of $n = 20$ fifth-grade students. The ages are rounded to the nearest half year.

9, 9.5, 9.5, 10, 10, 10, 10, 10.5, 10.5, 10.5, 10.5, 11, 11, 11, 11, 11, 11, 11.5, 11.5, 11.5

$$\overline{x} = \frac{9 + 9.5(2) + 10(4) + 10.5(4) + 11(6) + 11.5(3)}{20} = 10.525$$

The average age is 10.53 years, rounded to two places.

The variance may be calculated by using a table. Then the standard deviation is calculated by taking the square root of the variance. We will explain the parts of the table after calculating s.

Data	Frequency	Deviations	$Deviations^2$	(Frequency)($Deviations^2$)
x	f	$(x - \overline{x})$	$(x - \overline{x})^2$	$(f)(x - \overline{x})^2$
9	1	9 – 10.525 = –1.525	$(-1.525)^2 = 2.325625$	1 × 2.325625 = 2.325625
9.5	2	9.5 – 10.525 = –1.025	$(-1.025)^2 = 1.050625$	2 × 1.050625 = 2.101250
10	4	10 – 10.525 = –.525	$(-.525)^2 = .275625$	4 × .275625 = 1.1025
10.5	4	10.5 – 10.525 = –.025	$(-.025)^2 = .000625$	4 × .000625 = .0025
11	6	11 – 10.525 = .475	$(.475)^2 = .225625$	6 × .225625 = 1.35375
11.5	3	11.5 – 10.525 = .975	$(.975)^2 = .950625$	3 × .950625 = 2.851875
				The total is 9.7375.

Table 2.32

The last column simply multiplies each squared deviation by the frequency for the corresponding data value.

The sample variance, s^2, is equal to the sum of the last column (9.7375) divided by the total number of data values minus one (20 – 1):

$$s^2 = \frac{9.7375}{20 - 1} = .5125 \quad \textbf{2.4}$$

The **sample standard deviation** s is equal to the square root of the sample variance:

$s = \sqrt{.5125} = .715891$, which is rounded to two decimal places, $s = .72$.

Typically, you do the calculation for the standard deviation on your calculator or computer. The intermediate results are not rounded. This is done for accuracy.

Problem

- For the following problems, recall that **value = mean + (#ofSTDEVs)(standard deviation)**. Verify the mean and standard deviation on a calculator or computer. Note that these formulas are derived by algebraically manipulating the z-score formulas, given either parameters or statistics.
- For a sample: $x = \overline{x} + (\#ofSTDEVs)(s)$
- For a population: $x = \mu + (\#ofSTDEVs)(\sigma)$
- For this example, use $x = \overline{x} + (\#ofSTDEVs)(s)$ because the data is from a sample

a. Verify the mean and standard deviation on your calculator or computer.
b. Find the value that is one standard deviation above the mean. Find $(\overline{x} + 1s)$.
c. Find the value that is two standard deviations below the mean. Find $(\overline{x} - 2s)$.
d. Find the values that are 1.5 standard deviations **from** (below and above) the mean.

Solution

a. **USING THE TI-83, 83+, 84, 84+ CALCULATOR**

- Clear lists L1 and L2. Press STAT 4:ClrList. Enter 2^{nd} 1 for L1, the comma (,), and 2^{nd} 2 for L2.
- Enter data into the list editor. Press STAT 1:EDIT. If necessary, clear the lists by arrowing up into the name. Press CLEAR and arrow down.

- Put the data values (9, 9.5, 10, 10.5, 11, 11.5) into list L1 and the frequencies (1, 2, 4, 4, 6, 3) into list L2. Use the arrow keys to move around.
- Press STAT and arrow to CALC. Press 1:1-VarStats and enter L1 (2^{nd} 1), L2 (2^{nd} 2). Do not forget the comma. Press ENTER.
- $\overline{x}$ = 10.525.
- Use Sx because this is sample data (not a population): Sx=.715891.

b. $(\overline{x} + 1s) = 10.53 + (1)(.72) = 11.25$
c. $(\overline{x} - 2s) = 10.53 - (2)(.72) = 9.09$
d.
   - $(\overline{x} - 1.5s) = 10.53 - (1.5)(.72) = 9.45$
   - $(\overline{x} + 1.5s) = 10.53 + (1.5)(.72) = 11.61$

TRY IT 2.33

On a baseball team, the ages of each of the players are as follows:

21, 21, 22, 23, 24, 24, 25, 25, 28, 29, 29, 31, 32, 33, 33, 34, 35, 36, 36, 36, 36, 38, 38, 38, 40

Use your calculator or computer to find the mean and standard deviation. Then find the value that is two standard deviations above the mean.

Explanation of the standard deviation calculation shown in the table

The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11, which is indicated by the deviations .97 and .47. A positive deviation occurs when the data value is greater than the mean, whereas a negative deviation occurs when the data value is less than the mean. The deviation is –1.525 for the data value nine. **If you add the deviations, the sum is always zero**. We can sum the products of the frequencies and deviations to show that the sum of the deviations is always zero.
$1(-1.525) + 2(-1.025) + 4(-.525) + 4(-.025) + 6(.475) + 3(.975) = 0$ For Example 2.33, there are $n = 20$ deviations. So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation.

The variance is a squared measure and does not have the same units as the data. Taking the square root solves the problem. The standard deviation measures the spread in the same units as the data.

Notice that instead of dividing by $n = 20$, the calculation divided by $n - 1 = 20 - 1 = 19$ because the data is a sample. For the **sample** variance, we divide by the sample size minus one ($n - 1$). Why not divide by n? The answer has to do with the population variance. **The sample variance is an estimate of the population variance.** Based on the theoretical mathematics that lies behind these calculations, dividing by ($n - 1$) gives a better estimate of the population variance.

NOTE
Your concentration should be on what the standard deviation tells us about the data. The standard deviation is a number that measures how far the data are spread from the mean. Let a calculator or computer do the arithmetic.

The standard deviation, s or σ, is either zero or larger than zero. Describing the data with reference to the spread is called *variability*. The variability in data depends on the method by which the outcomes are obtained, for example, by measuring or by random sampling. When the standard deviation is zero, there is no spread; that is, all the data values are equal to each other. The standard deviation is small when all the data are concentrated close to the mean and larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make s or σ very large.

The standard deviation, when first presented, can seem unclear. By graphing your data, you can get a better *feel* for the deviations and the standard deviation. You will find that in symmetrical distributions, the standard deviation can be very helpful, but in skewed distributions, the standard deviation may not be much help. The reason is that the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, **always graph your**

data. Display your data in a histogram or a box plot.

EXAMPLE 2.34

Problem

Use the following data (first exam scores) from Susan Dean's spring precalculus class:

33, 42, 49, 49, 53, 55, 55, 61, 63, 67, 68, 68, 69, 69, 72, 73, 74, 78, 80, 83, 88, 88, 88, 90, 92, 94, 94, 94, 94, 96, 100

a. Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.
b. Calculate the following to one decimal place using a TI-83+ or TI-84 calculator:
   i. The sample mean
   ii. The sample standard deviation
   iii. The median
   iv. The first quartile
   v. The third quartile
   vi. *IQR*
c. Construct a box plot and a histogram on the same set of axes. Make comments about the box plot, the histogram, and the chart.

Solution

a. See Table 2.33.
b. Entering the data values into a list in your graphing calculator and then selecting Stat, Calc, and 1-Var Stats will produce the one-variable statistics you need.
c. The *x*-axis goes from 32.5 to 100.5; the *y*-axis goes from –2.4 to 15 for the histogram. The number of intervals is 5, so the width of an interval is (100.5 – 32.5) divided by 5, equal to 13.6. Endpoints of the intervals are as follows: the starting point is 32.5, 32.5 + 13.6 = 46.1, 46.1 + 13.6 = 59.7, 59.7 + 13.6 = 73.3, 73.3 + 13.6 = 86.9, 86.9 + 13.6 = 100.5 = the ending value; no data values fall on an interval boundary.

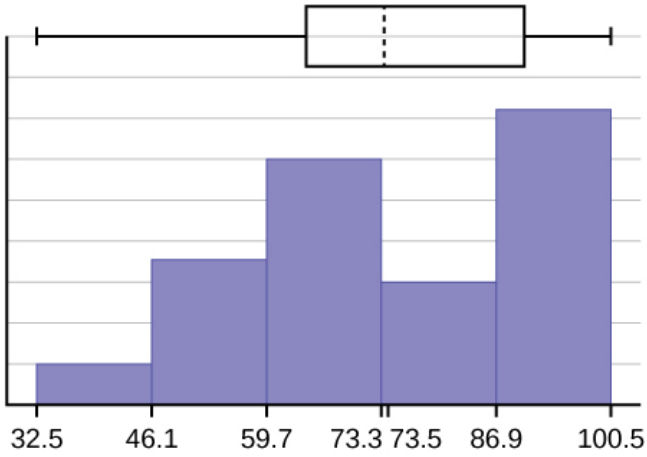


Figure 2.27

The long left whisker in the box plot is reflected in the left side of the histogram. The spread of the exam scores in the lower 50 percent is greater (73 – 33 = 40) than the spread in the upper 50 percent (100 – 73 = 27). The histogram, box plot, and chart all reflect this. There are a substantial number of A and B grades (80s, 90s, and 100). The histogram clearly shows this. The box plot shows us that the middle 50 percent of the exam scores (*IQR* = 29) are Ds, Cs, and Bs. The box plot also shows us that the lower 25 percent of the exam scores are Ds and Fs.

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
33	1	.032	.032
42	1	.032	.064
49	2	.065	.129
53	1	.032	.161
55	2	.065	.226
61	1	.032	.258
63	1	.032	.290
67	1	.032	.322
68	2	.065	.387
69	2	.065	.452
72	1	.032	.484
73	1	.032	.516
74	1	.032	.548
78	1	.032	.580
80	1	.032	.612
83	1	.032	.644
88	3	.097	.741
90	1	.032	.773
92	1	.032	.805
94	4	.129	.934
96	1	.032	.966
100	1	.032	.998 (Why isn't this value 1?)

Table 2.33

TRY IT 2.34

The following data show the different types of pet food that stores in the area carry:
6, 6, 6, 6, 7, 7, 7, 7, 7, 8, 9, 9, 9, 9, 10, 10, 10, 10, 10, 11, 11, 11, 11, 12, 12, 12, 12, 12, 12

Calculate the sample mean and the sample standard deviation to one decimal place using a TI-83+ or TI-84 calculator.

Standard deviation of Grouped Frequency Tables

Recall that for grouped data we do not know individual data values, so we cannot describe the typical value of the data with precision. In other words, we cannot find the exact mean, median, or mode. We can, however, determine the best estimate of the measures of center by finding the mean of the grouped data with the formula

$$Mean\ of\ Frequency\ Table = \frac{\sum fm}{\sum f},$$

where f = interval frequencies and m = interval midpoints.

Just as we could not find the exact mean, neither can we find the exact standard deviation. Remember that standard deviation describes numerically the expected deviation a data value has from the mean. In simple English, the standard deviation allows us to compare how *unusual* individual data are when compared to the mean.

EXAMPLE 2.35

Find the standard deviation for the data in Table 2.34.

Class	Frequency, f	Midpoint, m	m^2	$\overline{x}^2$	fm^2	Standard Deviation
0–2	1	1	1	7.58	1	3.5
3–5	6	4	16	7.58	96	3.5
6–8	10	7	49	7.58	490	3.5
9–11	7	10	100	7.58	700	3.5
12–14	0	13	169	7.58	0	3.5
15–17	2	16	256	7.58	512	3.5

Table 2.34

For this data set, we have the mean, $\overline{x}$ = 7.58, and the standard deviation, s_x = 3.5. This means that a randomly selected data value would be expected to be 3.5 units from the mean. If we look at the first class, we see that the class midpoint is equal to one. This is almost two full standard deviations from the mean since 7.58 – 3.5 – 3.5 = .58. While the formula for calculating the standard deviation is not complicated, $s_x = \sqrt{\frac{f(m-\overline{x})^2}{n-1}}$, where s_x = sample standard deviation, $\overline{x}$ = sample mean; the calculations are tedious. It is usually best to use technology when performing the calculations.

TRY IT 2.35

Find the standard deviation for the data from the previous example:

Class	Frequency, f
0–2	1
3–5	6
6–8	10
9–11	7
12–14	0
15–17	2

Table 2.35

First, press the **STAT** key and select **1:Edit.**

Figure 2.28

Input the midpoint values into **L1** and the frequencies into **L2.**

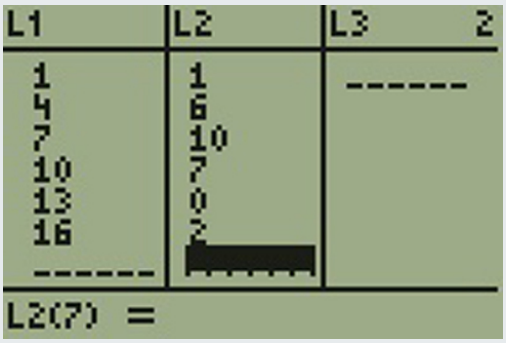


Figure 2.29

Select **STAT**, **CALC**, and **1: 1-Var Stats.**

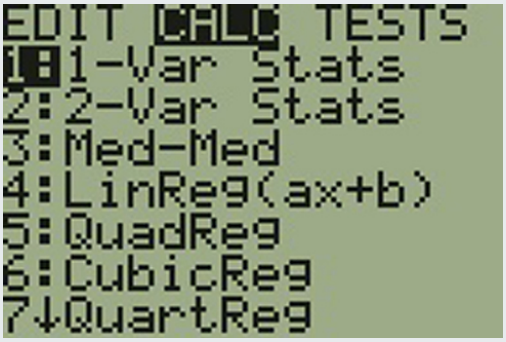


Figure 2.30

Select **2nd**, then **1**, then, **2nd**, then **2 Enter.**

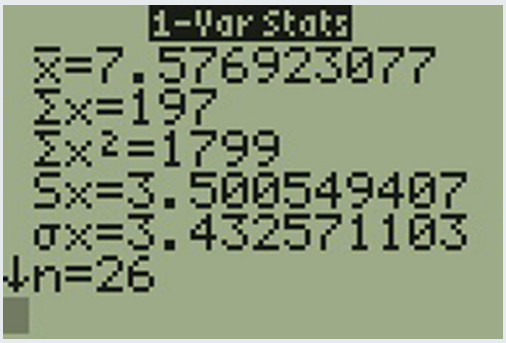


Figure 2.31

You will see displayed both a population standard deviation, σ_x, and the sample standard deviation, s_x.

Comparing Values from Different Data Sets

As explained before, a *z*-score allows us to compare statistics from different data sets. If the data sets have different means and standard deviations, then comparing the data values directly can be misleading.

- For each data value, calculate how many standard deviations away from its mean the value is.
- In symbols, the formulas for calculating *z*-scores become the following.

Sample		$z = \frac{x - \bar{x}}{s}$
Population		$z = \frac{x - \mu}{\sigma}$

Table 2.36

As shown in the table, when only a sample mean and sample standard deviation are given, the top formula is used. When the population mean and population standard deviation are given, the bottom formula is used.

EXAMPLE 2.36

Problem

Two students, John and Ali, from different high schools, wanted to find out who had the highest GPA when compared to his school. Which student had the highest GPA when compared to his school?

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	.7
Ali	77	80	10

Table 2.37

Solution

For each student, determine how many standard deviations (#ofSTDEVs) his GPA is away from the average, for his school. Pay careful attention to signs when comparing and interpreting the answer.

$z = \#\text{ of STDEVs} = \frac{\text{value} - \text{mean}}{\text{standard deviation}} = \frac{x+\mu}{\sigma}$

For John, $z = \#ofSTDEVs = \frac{2.85-3.0}{.7} = -0.21$

For Ali, $z = \#ofSTDEVs = \frac{77-80}{10} = -0.3$

John has the better GPA when compared to his school because his GPA is 0.21 standard deviations **below** his school's mean, while Ali's GPA is .3 standard deviations **below** his school's mean.

John's *z*-score of –.21 is higher than Ali's *z*-score of –.3. For GPA, higher values are better, so we conclude that John has the better GPA when compared to his school. The *z*-score representing John's score does not fall as far below the mean as the *z*-score representing Ali's score.

TRY IT 2.36

Two swimmers, Angie and Beth, from different teams, wanted to find out who had the fastest time for the 50-meter freestyle when compared to her team. Which swimmer had the fastest time when compared to her team?

Swimmer	Time (seconds)	Team Mean Time	Team Standard Deviation
Angie	26.2	27.2	.8
Beth	27.3	30.1	1.4

Table 2.38

The following lists give a few facts that provide a little more insight into what the standard deviation tells us about the distribution of the data.

For *any* data set, no matter what the distribution of the data is, the following are true:

- At least 75 percent of the data is within two standard deviations of the mean.
- At least 89 percent of the data is within three standard deviations of the mean.
- At least 95 percent of the data is within 4.5 standard deviations of the mean.
- This is known as Chebyshev's Rule.

A bell-shaped distribution is one that is normal and symmetric, meaning the curve can be folded along a line of symmetry drawn through the median, and the left and right sides of the curve would fold on each other symmetrically.. With a bell-shaped distribution, the mean, median, and mode are all located at the same place.

For data having a distribution that is *bell-shaped* and *symmetric*, the following are true:

- Approximately 68 percent of the data is within one standard deviation of the mean.
- Approximately 95 percent of the data is within two standard deviations of the mean.
- More than 99 percent of the data is within three standard deviations of the mean.
- This is known as the Empirical Rule.

- It is important to note that this rule applies only when the shape of the distribution of the data is bell-shaped and symmetric; we will learn more about this when studying the *Normal* or *Gaussian* probability distribution in later chapters.

2.8 Descriptive Statistics

Stats Lab

Descriptive Statistics

Student Learning Outcomes

- The student will construct a histogram and a box plot.
- The student will calculate univariate statistics.
- The student will examine the graphs to interpret what the data imply.

Collect the Data

Record the number of pairs of shoes you own.

1. Randomly survey 30 classmates about the number of pairs of shoes they own. Record their values.

____	____	____	____	____
____	____	____	____	____
____	____	____	____	____
____	____	____	____	____
____	____	____	____	____
____	____	____	____	____

Table 2.39 Survey Results

2. Construct a histogram. Make five to six intervals. Sketch the graph using a ruler and pencil and scale the axes.

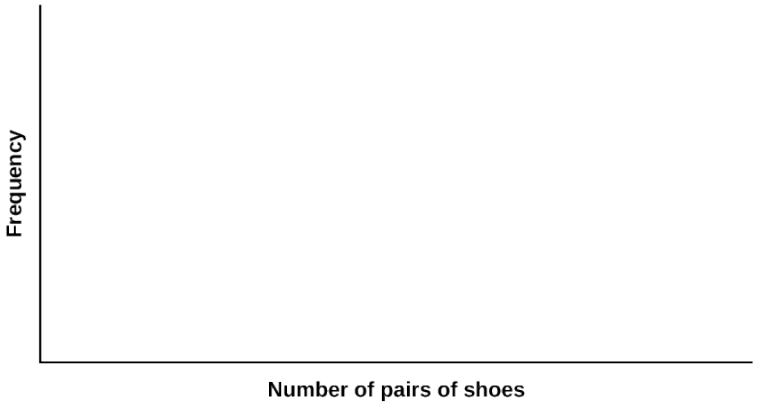


Figure 2.32

3. Calculate the following values:
   a. $\overline{x}$ = _____
   b. s = _____
4. Are the data discrete or continuous? How do you know?
5. In complete sentences, describe the shape of the histogram.
6. Are there any potential outliers? List the value(s) that could be outliers. Use a formula to check the end values to determine if they are potential outliers.

Analyze the Data

1. Determine the following values:
   a. Min = _____
   b. M = _____
   c. Max = _____
   d. Q_1 = _____
   e. Q_3 = _____
   f. IQR = _____
2. Construct a box plot of data.
3. What does the shape of the box plot imply about the concentration of data? Use complete sentences.
4. Using the box plot, how can you determine if there are potential outliers?
5. How does the standard deviation help you to determine concentration of the data and whether there are potential outliers?
6. What does the IQR represent in this problem?
7. Show your work to find the value that is 1.5 standard deviations
   a. above the mean.
   b. below the mean.

Key Terms

box plot a graph that gives a quick picture of the middle 50 percent of the data
first quartile the value that is the median of the lower half of the ordered data set
frequency the number of times a value of the data occurs
frequency polygon a data display that looks like a line graph but uses intervals to display ranges of large amounts of data
frequency table a data representation in which grouped data are displayed along with the corresponding frequencies
histogram a graphical representation in *x-y* form of the distribution of data in a data set; *x* represents the data and *y* represents the frequency, or relative frequency; the graph consists of contiguous rectangles
interquartile range or *IQR*, is the range of the middle 50 percent of the data values; the *IQR* is found by subtracting the first quartile from the third quartile
interval also called a class interval; an interval represents a range of data and is used when displaying large data sets
mean a number that measures the central tendency of the data; a common name for mean is *average*.
The term *mean* is a shortened form of *arithmetic mean*. By definition, the mean for a sample (denoted by $\overline{x}$) is $\overline{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$
median a number that separates ordered data into halves; half the values are the same number or smaller than the median, and half the values are the same number or larger than the median
The median may or may not be part of the data.
midpoint the mean of an interval in a frequency table
mode the value that appears most frequently in a set of data
outlier an observation that does not fit the rest of the data
paired data set two data sets that have a one-to-one relationship so that
- both data sets are the same size, and
- each data point in one data set is matched with exactly one point from the other set

percentile a number that divides ordered data into hundredths; percentiles may or may not be part of the data. The median of the data is the second quartile and the 50^{th} percentile
The first and third quartiles are the 25^{th} and the 75^{th} percentiles, respectively.
quartiles the numbers that separate the data into quarters; quartiles may or may not be part of the data; the second quartile is the median of the data
relative frequency the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes
skewed used to describe data that is not symmetrical; when the right side of a graph looks *chopped off* compared to the left side, we say it is *skewed to the left*.
When the left side of the graph looks *chopped off* compared to the right side, we say the data are *skewed to the right*. Alternatively, when the lower values of the data are more spread out, we say the data are skewed to the left. When the greater values are more spread out, the data are skewed to the right.
standard deviation a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: *s* for sample standard deviation and σ for population standard deviation
variance mean of the squared deviations from the mean, or the square of the standard deviation; for a set of data, a deviation can be represented as $x - \overline{x}$ where *x* is a value of the data and $\overline{x}$ is the sample mean; the sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and 1

Chapter Review

2.1 Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

A **stem-and-leaf plot** is a way to plot data and look at the distribution. In a stem-and-leaf plot, all data values within a class are visible. The advantage in a stem-and-leaf plot is that all values are listed, unlike a histogram, which gives classes of data values. A **line graph** is often used to represent a set of data values in which a quantity varies with time. These graphs are useful for finding trends, that is, finding a general pattern in data sets, including temperature, sales, employment, company profit, or cost, over a period of time. A **bar graph** is a chart that uses either horizontal or vertical bars to show comparisons among categories. One axis of the chart shows the specific categories being compared, and the other axis represents a discrete value. Bar graphs are especially useful when categorical data are being used.

2.2 Histograms, Frequency Polygons, and Time Series Graphs

A **histogram** is a graphic version of a frequency distribution. The graph consists of bars of equal width drawn adjacent to each other. The horizontal scale represents classes of quantitative data values, and the vertical scale represents frequencies. The heights of the bars correspond to frequency values. Histograms are typically used for large, continuous,

quantitative data sets. A frequency polygon can also be used when graphing large data sets with data points that repeat. The data usually go on the *y*-axis with the frequency being graphed on the *x*-axis. Time series graphs can be helpful when looking at large amounts of data for one variable over a period of time.

2.3 Measures of the Location of the Data

The values that divide a rank-ordered set of data into 100 equal parts are called percentiles. Percentiles are used to compare and interpret data. For example, an observation at the 50^{th} percentile would be greater than 50 percent of the other observations in the set. Quartiles divide data into quarters. The first quartile (Q_1) is the 25^{th} percentile, the second quartile (Q_2 or median) is the 50^{th} percentile, and the third quartile (Q_3) is the 75^{th} percentile. The interquartile range, or *IQR*, is the range of the middle 50 percent of the data values. The *IQR* is found by subtracting Q_1 from Q_3 and can help determine outliers by using the following two expressions.

- $Q_3 + IQR(1.5)$
- $Q_1 - IQR(1.5)$

2.4 Box Plots

Box plots are a type of graph that can help visually organize data. Before a box plot can be graphed, the following data points must be calculated: the minimum value, the first quartile, the median, the third quartile, and the maximum value. Once the box plot is graphed, you can display and compare distributions of data.

2.5 Measures of the Center of the Data

The mean and the median can be calculated to help you find the *center* of a data set. The mean is the best estimate for the actual data set, but the median is the best measurement when a data set contains several outliers or extreme values. The mode will tell you the most frequently occurring datum (or data) in your data set. The mean, median, and mode are extremely helpful when you need to analyze your data, but if your data set consists of ranges that lack specific values, the mean may seem impossible to calculate. However, the mean can be approximated if you add the lower boundary with the upper boundary and divide by two to find the midpoint of each interval. Multiply each midpoint by the number of values found in the corresponding range. Divide the sum of these values by the total number of data values in the set.

2.6 Skewness and the Mean, Median, and Mode

Looking at the distribution of data can reveal a lot about the relationship between the mean, the median, and the mode. There are three types of distributions. A **right (or positive) skewed** distribution has a shape like Figure 2.19. A **left (or negative) skewed** distribution has a shape like Figure 2.20. A **symmetrical** distribution looks like Figure 2.18.

2.7 Measures of the Spread of the Data

The standard deviation can help you calculate the spread of data. There are different equations to use if you are calculating the standard deviation of a sample or of a population.

- The standard deviation allows us to compare individual data or classes to the data set mean numerically.
- $s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$ or $s = \sqrt{\frac{\sum f(x-\bar{x})^2}{n-1}}$ is the formula for calculating the standard deviation of a sample. To calculate the standard deviation of a population, we would use the population mean, μ, and the formula $\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\sum f(x-\mu)^2}{N}}$.

Formula Review

2.3 Measures of the Location of the Data

$i = \left(\frac{k}{100}\right)(n+1)$

where i = the ranking or position of a data value,

k = the kth percentile,

n = total number of data.

Expression for finding the percentile of a data value $\left(\frac{x+0.5y}{n}\right)(100)$

where x = the number of values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile,

y = the number of data values equal to the data value for which you want to find the percentile,

n = total number of data.

2.5 Measures of the Center of the Data

$\mu = \frac{\sum fm}{\sum f}$ where f = interval frequencies and m = interval midpoints.

2.7 Measures of the Spread of the Data

$s_x = \sqrt{\frac{\sum fm^2}{n} - \overline{x}^2}$ where

s_x = sample standard deviation

$\overline{x}$ = sample mean

$z = \frac{(x-\mu)}{\theta}$.

$z = \frac{(x-\overline{x})}{s}$ and

Practice

2.1 Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

For each of the following data sets, create a stemplot and identify any outliers.

1. The miles-per-gallon ratings for 30 cars are shown below (lowest to highest):
   19, 19, 19, 20, 21, 21, 25, 25, 25, 26, 26, 28, 29, 31, 31, 32, 32, 33, 34, 35, 36, 37, 37, 38, 38, 38, 38, 41, 43, 43.
2. The height in feet of 25 trees is shown below (lowest to highest):
   25, 27, 33, 34, 34, 34, 35, 37, 37, 38, 39, 39, 39, 40, 41, 45, 46, 47, 49, 50, 50, 53, 53, 54, 54.
3. The data are the prices of different laptops at an electronics store. Round each value to the nearest 10.
   249, 249, 260, 265, 265, 280, 299, 299, 309, 319, 325, 326, 350, 350, 350, 365, 369, 389, 409, 459, 489, 559, 569, 570, 610
4. The following data are daily high temperatures in a town for one month:
   61, 61, 62, 64, 66, 67, 67, 67, 68, 69, 70, 70, 70, 71, 71, 72, 74, 74, 74, 75, 75, 75, 76, 76, 77, 78, 78, 79, 79, 95.

For the next three exercises, use the data to construct a line graph.

5. In a survey, 40 people were asked how many times they visited a store before making a major purchase. The results are shown in Table 2.40.

Number of Times in Store	Frequency
1	4
2	10
3	16
4	6
5	4

Table 2.40

6. In a survey, several people were asked how many years it has been since they purchased a mattress. The results are shown in Table 2.41.

Years Since Last Purchase	Frequency
0	2
1	8
2	13
3	22

Table 2.41

Years Since Last Purchase	Frequency
4	16
5	9

Table 2.41

7. Several children were asked how many TV shows they watch each day. The results of the survey are shown in Table 2.42.

Number of TV Shows	Frequency
0	12
1	18
2	36
3	7
4	2

Table 2.42

8. The students in Ms. Ramirez's math class have birthdays in each of the four seasons. Table 2.43 shows the four seasons, the number of students who have birthdays in each season, and the percentage of students in each group. Construct a bar graph showing the number of students.

Seasons	Number of Students	Proportion of Population
Spring	8	24%
Summer	9	26%
Autumn	11	32%
Winter	6	18%

Table 2.43

9. Using the data from Mrs. Ramirez's math class supplied in Exercise 2.8, construct a bar graph showing the percentages.

10. David County has six high schools. Each school sent students to participate in a county-wide science competition. Table 2.44 shows the percentage breakdown of competitors from each school and the percentage of the entire student population of the county that goes to each school. Construct a bar graph that shows the population percentage of competitors from each school.

High School	Science Competition Population	Overall Student Population
Alabaster	28.9%	8.6%
Concordia	7.6%	23.2%
Genoa	12.1%	15.0%
Mocksville	18.5%	14.3%
Tynneson	24.2%	10.1%
West End	8.7%	28.8%

Table 2.44

11. Use the data from the David County science competition supplied in Exercise 2.10. Construct a bar graph that shows the county-wide population percentage of students at each school.

2.2 Histograms, Frequency Polygons, and Time Series Graphs

12. 65 randomly selected car salespersons were asked the number of cars they generally sell in one week. 14 people answered that they generally sell three cars, 19 generally sell four cars, 12 generally sell five cars, nine generally sell six cars, and 11 generally sell seven cars. Complete the table.

Data Value (Number of Cars)	Frequency	Relative Frequency	Cumulative Relative Frequency

Table 2.45

13. What does the frequency column in Table 2.45 sum to? Why?

14. What does the relative frequency column in Table 2.45 sum to? Why?

15. What is the difference between relative frequency and frequency for each data value in Table 2.45?

16. What is the difference between cumulative relative frequency and relative frequency for each data value?

17. To construct the histogram for the data in Table 2.45, determine appropriate minimum and maximum *x*- and *y*-values and the scaling. Sketch the histogram. Label the horizontal and vertical axes with words. Include numerical scaling.

Figure 2.33

18. Construct a frequency polygon for the following.

a.

Pulse Rates for Women	Frequency
60–69	12
70–79	14
80–89	11
90–99	1
100–109	1
110–119	0
120–129	1

Table 2.46

b.

Actual Speed in a 30-MPH Zone	Frequency
42–45	25
46–49	14
50–53	7
54–57	3
58–61	1

Table 2.47

c.

Tar (mg) in Nonfiltered Cigarettes	Frequency
10–13	1
14–17	0
18–21	15
22–25	7
26–29	2

Table 2.48

19. Construct a frequency polygon from the frequency distribution for the 50 highest-ranked countries for depth of hunger.

Depth of Hunger	Frequency
230–259	21
260–289	13
290–319	5
320–349	7
350–379	1
380–409	1
410–439	1

Table 2.49

20. Use the two frequency tables to compare the life expectancy of men and women from 20 randomly selected countries. Include an overlaid frequency polygon and discuss the shapes of the distributions, the center, the spread, and any outliers. What can we conclude about the life expectancy of women compared to men?

Life Expectancy at Birth – Women	Frequency
49–55	3
56–62	3
63–69	1
70–76	3

Table 2.50

Life Expectancy at Birth – Women	Frequency
77–83	8
84–90	2

Table 2.50

Life Expectancy at Birth – Men	Frequency
49–55	3
56–62	3
63–69	1
70–76	1
77–83	7
84–90	5

Table 2.51

21. Construct a times series graph for (a) the number of male births, (b) the number of female births, and (c) the total number of births.

Sex/Year	1855	1856	1857	1858	1859	1860	1861
Female	45,545	49,582	50,257	50,324	51,915	51,220	52,403
Male	47,804	52,239	53,158	53,694	54,628	54,409	54,606
Total	93,349	101,821	103,415	104,018	106,543	105,629	107,009

Table 2.52

Sex/Year	1862	1863	1864	1865	1866	1867	1868	1869
Female	51,812	53,115	54,959	54,850	55,307	55,527	56,292	55,033
Male	55,257	56,226	57,374	58,220	58,360	58,517	59,222	58,321
Total	107,069	109,341	112,333	113,070	113,667	114,044	115,514	113,354

Table 2.53

Sex/Year	1871	1870	1872	1871	1872	1827	1874	1875
Female	56,099	56,431	57,472	56,099	57,472	58,233	60,109	60,146

Table 2.54

Male	60,029	58,959	61,293	60,029	61,293	61,467	63,602	63,432
Total	116,128	115,390	118,765	116,128	118,765	119,700	123,711	123,578

Table 2.54

22. The following data sets list full-time police per 100,000 citizens along with incidents of a certain crime per 100,000 citizens for the city of Detroit, Michigan, during the period from 1961 to 1973.

Year	1961	1962	1963	1964	1965	1966	1967
Police	260.35	269.8	272.04	272.96	272.51	261.34	268.89
Incidents	8.6	8.9	8.52	8.89	13.07	14.57	21.36

Table 2.55

Year	1968	1969	1970	1971	1972	1973
Police	295.99	319.87	341.43	356.59	376.69	390.19
Incidents	28.03	31.49	37.39	46.26	47.24	52.33

Table 2.56

a. Construct a double time series graph using a common *x*-axis for both sets of data.
b. Which variable increased the fastest? Explain.
c. Did Detroit's increase in police officers have an impact on the incident rate? Explain.

2.3 Measures of the Location of the Data

23. Listed are 29 ages for Academy Award-winning best actors *in order from smallest to largest:*
18, 21, 22, 25, 26, 27, 29, 30, 31, 33, 36, 37, 41, 42, 47, 52, 55, 57, 58, 62, 64, 67, 69, 71, 72, 73, 74, 76, 77
a. Find the 40^{th} percentile.
b. Find the 78^{th} percentile.

24. Listed are 32 ages for Academy Award-winning best actors *in order from smallest to largest:*
18, 18, 21, 22, 25, 26, 27, 29, 30, 31, 31, 33, 36, 37, 37, 41, 42, 47, 52, 55, 57, 58, 62, 64, 67, 69, 71, 72, 73, 74, 76, 77
a. Find the percentile of 37.
b. Find the percentile of 72.

25. Jesse was ranked 37^{th} in his graduating class of 180 students. At what percentile is Jesse's ranking?

26. a. For runners in a race, a low time means a faster run. The winners in a race have the shortest running times. Is it more desirable to have a finish time with a high or a low percentile when running a race?
b. The 20^{th} percentile of run times in a particular race is 5.2 minutes. Write a sentence interpreting the 20^{th} percentile in the context of the situation.
c. A bicyclist in the 90^{th} percentile of a bicycle race completed the race in 1 hour and 12 minutes. Is he among the fastest or slowest cyclists in the race? Write a sentence interpreting the 90^{th} percentile in the context of the situation.

27. a. For runners in a race, a higher speed means a faster run. Is it more desirable to have a speed with a high or a low percentile when running a race?
b. The 40^{th} percentile of speeds in a particular race is 7.5 miles per hour. Write a sentence interpreting the 40^{th} percentile in the context of the situation.

28. On an exam, would it be more desirable to earn a grade with a high or a low percentile? Explain.

29. Mina is waiting in line at the Department of Motor Vehicles. Her wait time of 32 minutes is the 85th percentile of wait times. Is that good or bad? Write a sentence interpreting the 85th percentile in the context of this situation.

30. In a survey collecting data about the salaries earned by recent college graduates, Li found that her salary was in the 78th percentile. Should Li be pleased or upset by this result? Explain.

31. In a study collecting data about the repair costs of damage to automobiles in a certain type of crash tests, a certain model of car had $1,700 in damage and was in the 90th percentile. Should the manufacturer and the consumer be pleased or upset by this result? Explain and write a sentence that interprets the 90th percentile in the context of this problem.

32. The University of California has two criteria used to set admission standards for freshman to be admitted to a college in the UC system:
   a. Students' GPAs and scores on standardized tests (SATs and ACTs) are entered into a formula that calculates an *admissions index* score. The admissions index score is used to set eligibility standards intended to meet the goal of admitting the top 12 percent of high school students in the state. In this context, what percentile does the top 12 percent represent?
   b. Students whose GPAs are at or above the 96th percentile of all students at their high school are eligible, called eligible in the local context, even if they are not in the top 12 percent of all students in the state. What percentage of students from each high school are *eligible in the local context*?

33. Suppose that you are buying a house. You and your real estate agent have determined that the most expensive house you can afford is the 34th percentile. The 34th percentile of housing prices is $240,000 in the town you want to move to. In this town, can you afford 34 percent of the houses or 66 percent of the houses?

Use Exercise 2.25 to calculate the following values.

34. First quartile = _____

35. Second quartile = median = 50th percentile = _____

36. Third quartile = _____

37. Interquartile range (*IQR*) = _____ – _____ = _____

38. 10th percentile = _____

39. 70th percentile = _____

2.4 Box Plots

Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars, 19 generally sell four cars, 12 generally sell five cars, nine generally sell six cars, and 11 generally sell seven cars.

40. Construct a box plot below. Use a ruler to measure and scale accurately.

41. Looking at your box plot, does it appear that the data are concentrated together, spread out evenly, or concentrated in some areas but not in others? How can you tell?

2.5 Measures of the Center of the Data

42. Find the mean for the following frequency tables:

a.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

Table 2.57

b.

Daily Low Temperature	Frequency
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1
89.5–99.5	0

Table 2.58

c.

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

Table 2.59

Use the following information to answer the next three exercises: The following data show the lengths of boats moored in a marina. The data are ordered from smallest to largest: 16, 17, 19, 20, 20, 21, 23, 24, 25, 25, 25, 26, 26, 27, 27, 27, 28, 29, 30, 32, 33, 33, 34, 35, 37, 39, 40

43. Calculate the mean.

44. Identify the median.

45. Identify the mode.

Use the following information to answer the next three exercises: Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars, 19 generally sell four cars, 12 generally sell five cars, nine generally sell six cars, and 11 generally sell seven cars. Calculate the following.

46. sample mean = $\overline{x}$ = _____

47. median = _____

48. mode = _____

2.6 Skewness and the Mean, Median, and Mode

Use the following information to answer the next three exercises. State whether the data are symmetrical, skewed to the left, or skewed to the right.

49. 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5

50. 16, 17, 19, 22, 22, 22, 22, 22, 23

51. 87, 87, 87, 87, 87, 88, 89, 89, 90, 91

52. When the data are skewed left, what is the typical relationship between the mean and median?

53. When the data are symmetrical, what is the typical relationship between the mean and median?

54. What word describes a distribution that has two modes?

55. Describe the shape of this distribution.

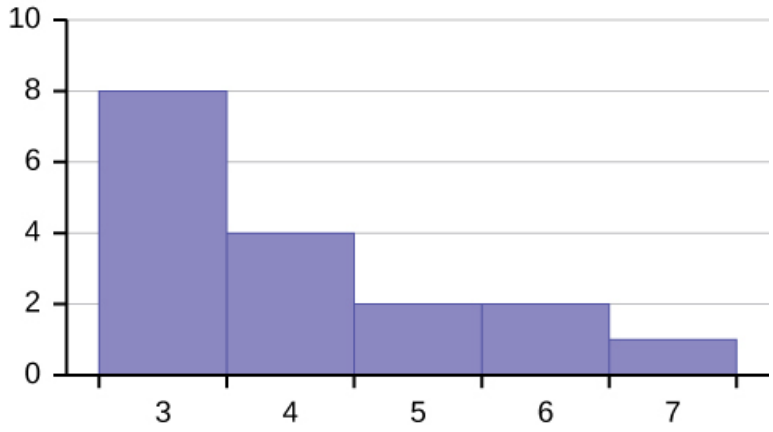


Figure 2.34

56. Describe the relationship between the mode and the median of this distribution.

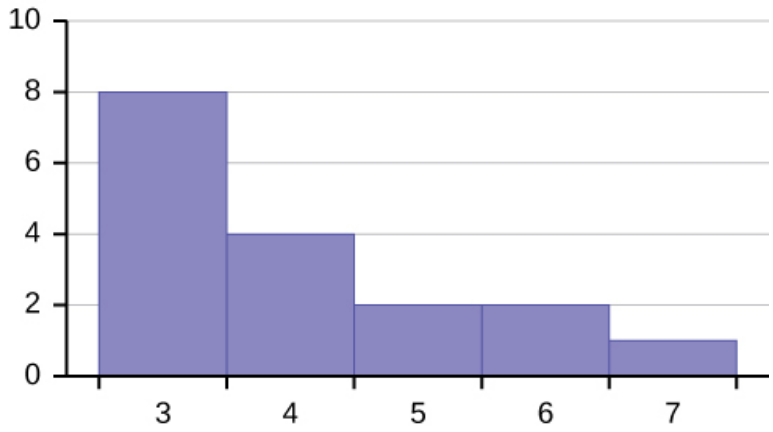


Figure 2.35

57. Describe the relationship between the mean and the median of this distribution.

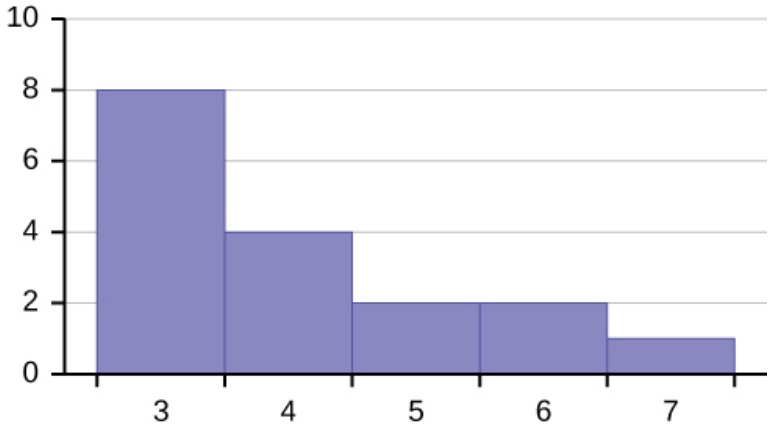


Figure 2.36

58. Describe the shape of this distribution.

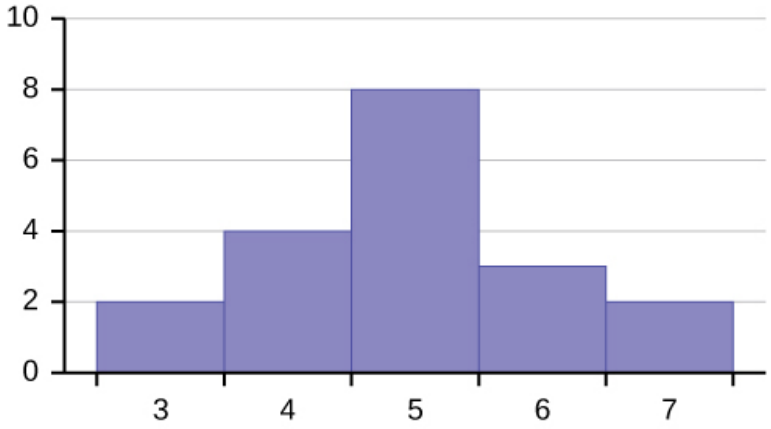


Figure 2.37

59. Describe the relationship between the mode and the median of this distribution.

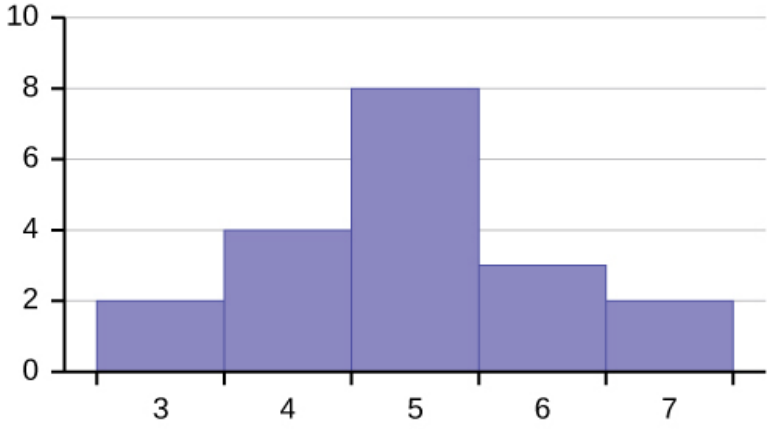


Figure 2.38

60. Are the mean and the median the exact same in this distribution? Why or why not?

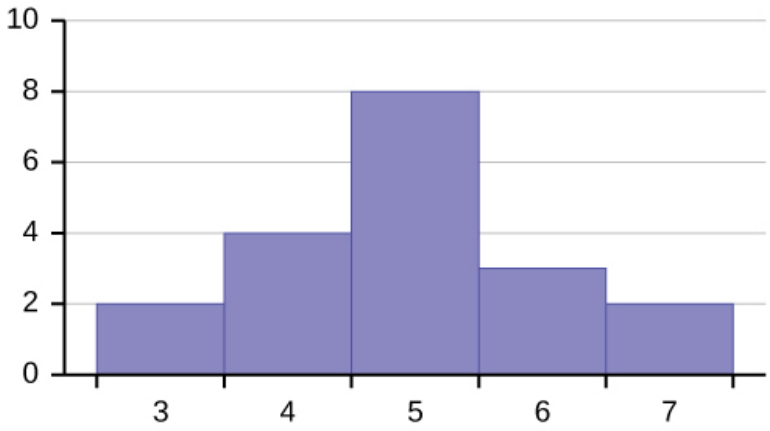


Figure 2.39

61. Describe the shape of this distribution.

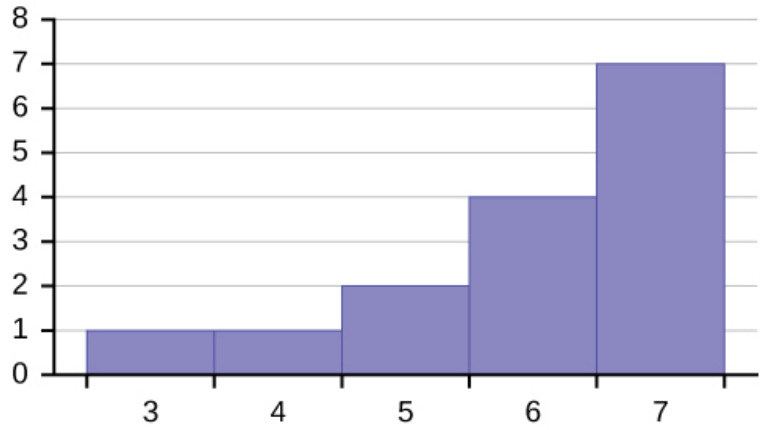


Figure 2.40

62. Describe the relationship between the mode and the median of this distribution.

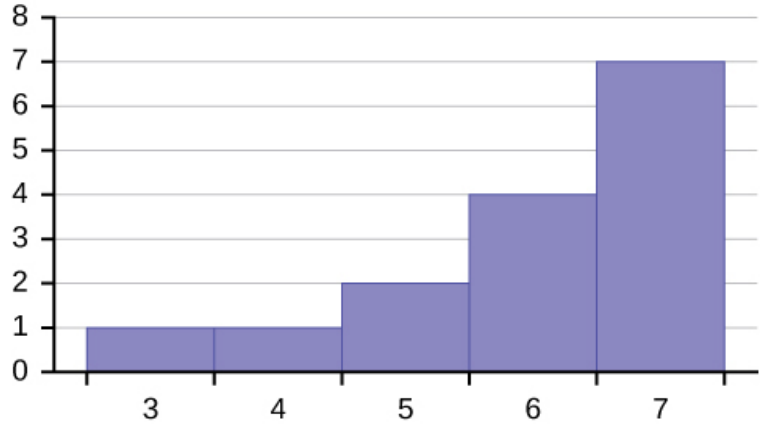


Figure 2.41

63. Describe the relationship between the mean and the median of this distribution.

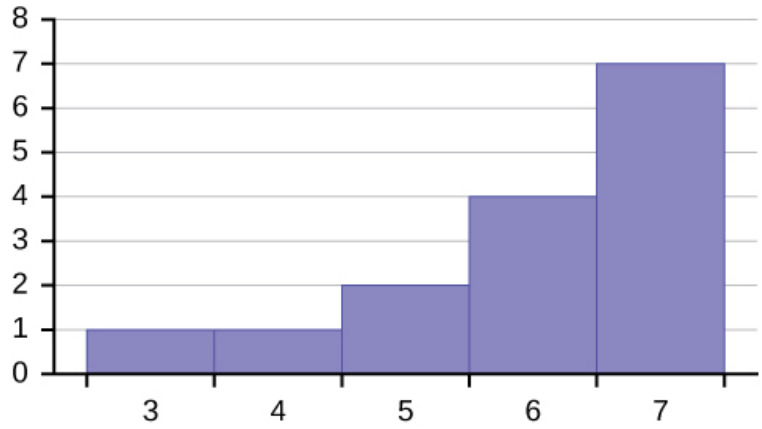


Figure 2.42

64. The mean and median for the data are the same.
3, 4, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7
Is the data perfectly symmetrical? Why or why not?

65. Which is the greatest, the mean, the mode, or the median of the data set?
11, 11, 12, 12, 12, 12, 13, 15, 17, 22, 22, 22

66. Which is the least, the mean, the mode, and the median of the data set?
56, 56, 56, 58, 59, 60, 62, 64, 64, 65, 67

67. Of the three measures, which tends to reflect skewing the most, the mean, the mode, or the median? Why?

68. In a perfectly symmetrical distribution, when would the mode be different from the mean and median?

2.7 Measures of the Spread of the Data

For each of the examples given below, tell whether the differences in outcomes may be explained by measurement variability, natural variability, induced variability, or sampling variability.

69. Scientists randomly select five groups of 10 women from a population of 1,000 women to record their body fat percentage. The scientists compute the mean body fat percentage from each group. The differences in outcomes may be attributed to which type of variability?

70. A pharmaceutical company randomly assigns participants to one of two groups: one is a control group receiving a placebo, and another is a treatment group receiving a new drug to lower blood pressure. The differences in outcomes may be attributed to which type of variability?

71. Jaiqua and Harold are trying to determine how ramp steepness affects the speed of a ball rolling down the ramp. They measure the time it takes for the ball to roll down ramps of differing slopes. When Jaiqua rolls the ball and Harold works the stopwatch, they get different results than when Harold rolls the ball and Jaiqua works the stopwatch. The differences in outcomes may be attributed to which type of variability?

72. Twenty people begin the same workout program on the same day and continue for three months. During that time, all participants worked out for the same amount of time and did the same number of exercises and repetitions. Each person was weighed at both the beginning and the end of the program. The differences in outcomes regarding the amount of weight lost may be attributed to which type of variability?

Use the following information to answer the next two exercises. The following data are the distances between 20 retail stores and a large distribution center. The distances are in miles.
29, 37, 38, 40, 58, 67, 68, 69, 76, 86, 87, 95, 96, 96, 99, 106, 112, 127, 145, 150

73. Use a graphing calculator or computer to find the standard deviation and round to the nearest tenth.

74. Find the value that is one standard deviation below the mean.

75. Two baseball players, Fredo and Karl, on different teams wanted to find out who had the higher batting average when compared to his team. Which baseball player had the higher batting average when compared to his team?

Baseball Player	Batting Average	Team Batting Average	Team Standard Deviation
Fredo	.158	.166	.012
Karl	.177	.189	.015

Table 2.60

76. Use Table 2.60 to find the value that is three standard deviations
above the mean, and
below the mean

77. Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

a.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

Table 2.61

b.

Daily Low Temperature	Frequency
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1
89.5–99.5	0

Table 2.62

c.

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

Table 2.63

Homework

2.1 Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

78. Student grades on a chemistry exam were 77, 78, 76, 81, 86, 51, 79, 82, 84, and 99.
   a. Construct a stem-and-leaf plot of the data.
   b. Are there any potential outliers? If so, which scores are they? Why do you consider them outliers?

79. Table 2.64 contains the 2010 rates for a specific disease in U.S. states and Washington, DC.

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2
Alaska	24.5	Louisiana	31.0	Ohio	29.2
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0
Georgia	29.6	Nebraska	26.9	Utah	22.5
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0

Table 2.64

State	Percent (%)	State	Percent (%)	State	Percent (%)
Illinois	28.2	New Jersey	23.8	Washington	25.5
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

Table 2.64

a. Use a random number generator to randomly pick eight states. Construct a bar graph of the rates of a specific disease of those eight states.
b. Construct a bar graph for all the states beginning with the letter *A*.
c. Construct a bar graph for all the states beginning with the letter *M*.

2.2 Histograms, Frequency Polygons, and Time Series Graphs

80. Suppose that three book publishers were interested in the number of fiction paperbacks adult consumers purchase per month. Each publisher conducted a survey. In the survey, adult consumers were asked the number of fiction paperbacks they had purchased the previous month. The results are as follows:

Number of Books	Frequency	Relative Frequency
0	10	
1	12	
2	16	
3	12	
4	8	
5	6	
6	2	
8	2	

Table 2.65 Publisher A

Number of Books	Frequency	Relative Frequency
0	18	
1	24	
2	24	
3	22	

Table 2.66 Publisher B

Number of Books	Frequency	Relative Frequency
4	15	
5	10	
7	5	
9	1	

Table 2.66 Publisher B

Number of Books	Frequency	Relative Frequency
0–1	20	
2–3	35	
4–5	12	
6–7	2	
8–9	1	

Table 2.67 Publisher C

a. Find the relative frequencies for each survey. Write them in the charts.
b. Using either a graphing calculator or computer or by hand, use the frequency column to construct a histogram for each publisher's survey. For Publishers A and B, make bar widths of 1. For Publisher C, make bar widths of 2.
c. In complete sentences, give two reasons why the graphs for Publishers A and B are not identical.
d. Would you have expected the graph for Publisher C to look like the other two graphs? Why or why not?
e. Make new histograms for Publisher A and Publisher B. This time, make bar widths of 2.
f. Now, compare the graph for Publisher C to the new graphs for Publishers A and B. Are the graphs more similar or more different? Explain your answer.

81. Often, cruise ships conduct all onboard transactions, with the exception of souvenirs, on a cashless basis. At the end of the cruise, guests pay one bill that covers all onboard transactions. Suppose that 60 single travelers and 70 couples were surveyed as to their onboard bills for a seven-day cruise from Los Angeles to the Mexican Riviera. Following is a summary of the bills for each group:

Amount ($)	Frequency	Relative Frequency
51–100	5	
101–150	10	
151–200	15	
201–250	15	

Table 2.68 Singles

Amount ($)	Frequency	Relative Frequency
251–300	10	
301–350	5	

Table 2.68 Singles

Amount ($)	Frequency	Relative Frequency
100–150	5	
201–250	5	
251–300	5	
301–350	5	
351–400	10	
401–450	10	
451–500	10	
501–550	10	
551–600	5	
601–650	5	

Table 2.69 Couples

a. Fill in the relative frequency for each group.
b. Construct a histogram for the singles group. Scale the *x*-axis by $50 widths. Use relative frequency on the *y*-axis.
c. Construct a histogram for the couples group. Scale the *x*-axis by $50 widths. Use relative frequency on the *y*-axis.
d. Compare the two graphs:
    i. List two similarities between the graphs.
    ii. List two differences between the graphs.
    iii. Overall, are the graphs more similar or different?
e. Construct a new graph for the couples by hand. Since each couple is paying for two individuals, instead of scaling the *x*-axis by $50, scale it by $100. Use relative frequency on the *y*-axis.
f. Compare the graph for the singles with the new graph for the couples:
    i. List two similarities between the graphs.
    ii. Overall, are the graphs more similar or different?
g. How did scaling the couples graph differently change the way you compared it to the singles graph?
h. Based on the graphs, do you think that individuals spend the same amount, more or less, as singles as they do person by person as a couple? Explain why in one or two complete sentences.

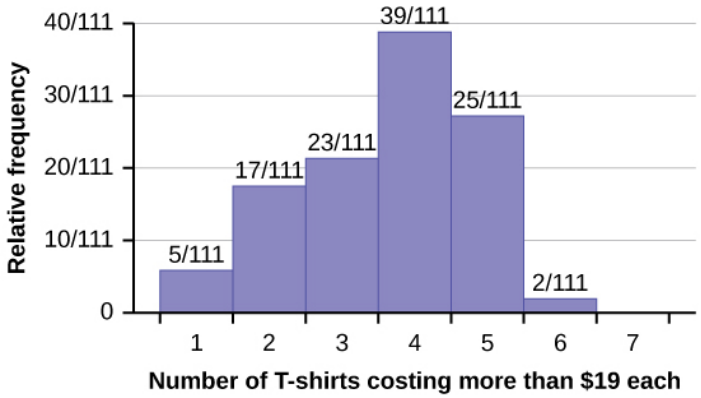
82. 25 randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

Number of Movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0	5		
1	9		
2	6		
3	4		
4	1		

Table 2.70

a. Construct a histogram of the data.
b. Complete the columns of the chart.

Use the following information to answer the next two exercises: Suppose 111 people who shopped in a special T-shirt store were asked the number of T-shirts they own costing more than $19 each.



83. The percentage of people who own at most three T-shirts costing more than $19 each is approximately _____.
a. 21
b. 59
c. 41
d. cannot be determined

84. If the data were collected by asking the first 111 people who entered the store, then the type of sampling is _____.
a. cluster
b. simple random
c. stratified
d. convenience

85. Following are the 2010 obesity rates by U.S. states and Washington, DC.

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2
Alaska	24.5	Louisiana	31.0	Ohio	29.2

Table 2.71

State	Percent (%)	State	Percent (%)	State	Percent (%)
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0
Georgia	29.6	Nebraska	26.9	Utah	22.5
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0
Illinois	28.2	New Jersey	23.8	Washington	25.5
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

Table 2.71

Construct a bar graph of obesity rates of your state and the four states closest to your state. Hint—Label the *x*-axis with the states.

2.3 Measures of the Location of the Data

86. The median age for U.S. ethnicity A currently is 30.9 years; for U.S. ethnicity B, it is 42.3 years.

a. Based on this information, give two reasons why ethnicity A median age could be lower than the ethnicity B median age.
b. Does the lower median age for ethnicity A necessarily mean that ethnicity A die younger than ethnicity B? Why or why not?
c. How might it be possible for ethnicity A and ethnicity B to die at approximately the same age but for the median age for ethnicity B to be higher?

87. Six hundred adult Americans were asked by telephone poll, "What do you think constitutes a middle-class income?" The results are in Table 2.72. Also, include the left endpoint but not the right endpoint.

Salary ($)	Relative Frequency
< 20,000	.02
20,000–25,000	.09
25,000–30,000	.19
30,000–40,000	.26
40,000–50,000	.18
50,000–75,000	.17
75,000–99,999	.02
100,000+	.01

Table 2.72

a. What percentage of the survey answered "not sure"?
b. What percentage think that middle class is from $25,000 to $50,000?
c. Construct a histogram of the data.
   i. Should all bars have the same width, based on the data? Why or why not?
   ii. How should the < 20,000 and the 100,000+ intervals be handled? Why?
d. Find the 40^{th} and 80^{th} percentiles.
e. Construct a bar graph of the data.

88. Given the following box plot, answer the questions.

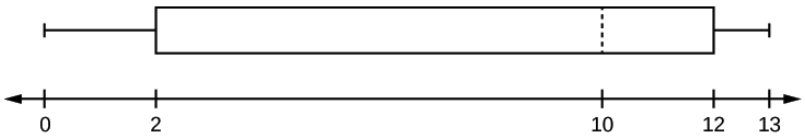


Figure 2.43

a. Which quarter has the smallest spread of data? What is that spread?
b. Which quarter has the largest spread of data? What is that spread?
c. Find the interquartile range (*IQR*).
d. Are there more data in the interval 5–10 or in the interval 10–13? How do you know this?
e. Which interval has the fewest data in it? How do you know this?
   i. 0–2
   ii. 2–4
   iii. 10–12
   iv. 12–13
   v. need more information

89. The following box plot shows the ages of the U.S. population for 1990, the latest available year:

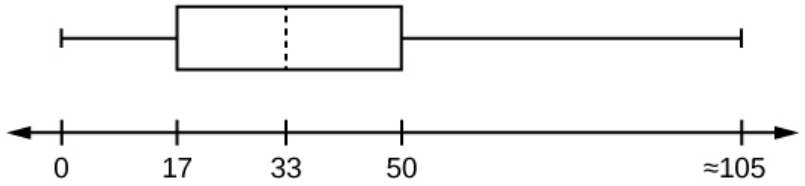


Figure 2.44

a. Are there fewer or more children (age 17 and under) than senior citizens (age 65 and over)? How do you know?
b. 12.6 percent are age 65 and over. Approximately what percentage of the population are working-age adults (above age 17 to age 65)?

2.4 Box Plots

90. In a survey of 20-year-olds in China, Germany, and the United States, people were asked the number of foreign countries they had visited in their lifetime. The following box plots display the results:

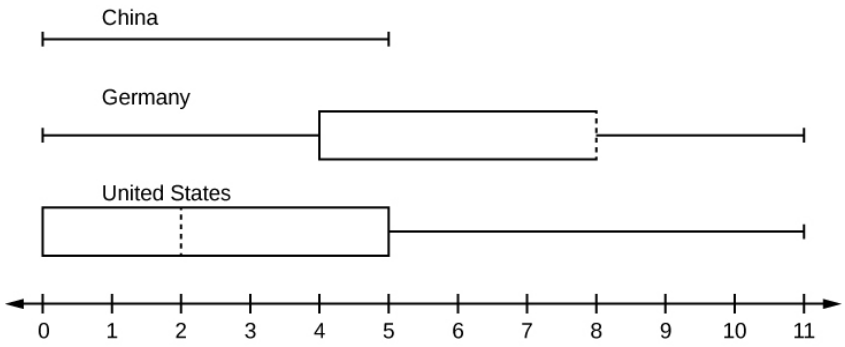


Figure 2.45

a. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected.
b. Have more Americans or more Germans surveyed been to more than eight foreign countries?
c. Compare the three box plots. What do they imply about the foreign travel of 20-year-old residents of the three countries when compared to each other?

91. Given the following box plot, answer the questions.

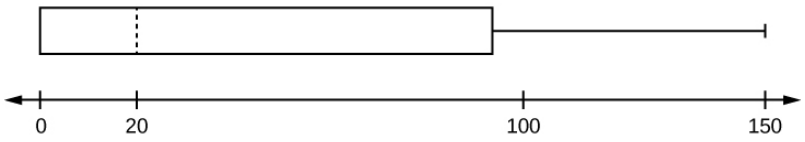


Figure 2.46

a. Think of an example (in words) where the data might fit into the above box plot. In two to five sentences, write down the example.
b. What does it mean to have the first and second quartiles so close together, while the second to third quartiles are far apart?

92. Given the following box plots, answer the questions.

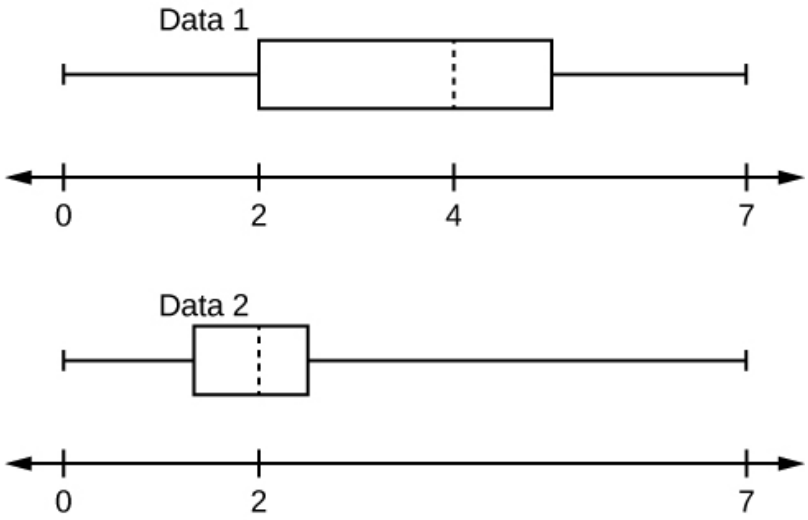


Figure 2.47

a. In complete sentences, explain why each statement is false.
   i. **Data 1** has more data values above two than **Data 2** has above two.
   ii. The data sets cannot have the same mode.
   iii. For **Data 1**, there are more data values below four than there are above four.

b. For which group, Data 1 or Data 2, is the value of 7 more likely to be an outlier? Explain why in complete sentences.

93. A survey was conducted of 130 purchasers of new black sports cars, 130 purchasers of new red sports cars, and 130 purchasers of new white sports cars. In it, people were asked the age they were when they purchased their car. The following box plots display the results:

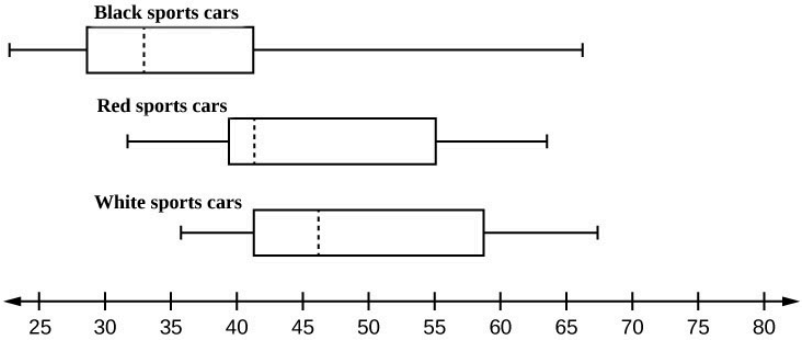


Figure 2.48

a. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected for that car series.
b. Which group is most likely to have an outlier? Explain how you determined that.
c. Compare the three box plots. What do they imply about the age of purchasing a sports car from the series when compared to each other?
d. Look at the red sports cars. Which quarter has the smallest spread of data? What is the spread?
e. Look at the red sports cars. Which quarter has the largest spread of data? What is the spread?
f. Look at the red sports cars. Estimate the interquartile range (*IQR*).
g. Look at the red sports cars. Are there more data in the interval 31–38 or in the interval 45–55? How do you know this?
h. Look at the red sports cars. Which interval has the fewest data in it? How do you know this?
   i. 31–35
   ii. 38–41
   iii. 41–64

94. Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

Number of Movies	Frequency
0	5
1	9
2	6
3	4
4	1

Table 2.73

Construct a box plot of the data.

2.5 Measures of the Center of the Data

95. Scientists are studying a particular disease. They found that countries that have the highest rates of people who have ever been diagnosed with this disease range from 11.4 percent to 74.6 percent.

Percentage of Population Diagnosed	Number of Countries
11.4–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

Table 2.74

a. What is the best estimate of the average percentage affected by the disease for these countries?
b. The United States has an average disease rate of 33.9 percent. Is this rate above average or below?
c. How does the United States compare to other countries?

96. Table 2.75 gives the percentage of children under age five have been diagnosed with a medical condition. What is the best estimate for the mean percentage of children with the condition?

Percentage of Children with the Condition	Number of Countries
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

Table 2.75

2.6 Skewness and the Mean, Median, and Mode

97. The median age of the U.S. population in 1980 was 30.0 years. In 1991, the median age was 33.1 years.
   a. What does it mean for the median age to rise?
   b. Give two reasons why the median age could rise.
   c. For the median age to rise, is the actual number of children less in 1991 than it was in 1980? Why or why not?

2.7 Measures of the Spread of the Data

Use the following information to answer the next nine exercises: The population parameters below describe the full-time equivalent number of students (FTES) each year at Lake Tahoe Community College from 1976–1977 through 2004–2005.

- μ = 1,000 FTES
- median = 1,014 FTES
- σ = 474 FTES
- first quartile = 528.5 FTES
- third quartile = 1,447.5 FTES
- n = 29 years

98. A sample of 11 years is taken. About how many are expected to have an FTES of 1,014 or above? Explain how you determined your answer.

99. Seventy-five percent of all years have an FTES
   a. at or below _____.
   b. at or above _____.

100. The population standard deviation = _____.

101. What percentage of the FTES were from 528.5 to 1,447.5? How do you know?

102. What is the *IQR*? What does the *IQR* represent?

103. How many standard deviations away from the mean is the median?
Additional Information: The population FTES for 2005–2006 through 2010–2011 was given in an updated report. The data are reported here.

Year	2005–2006	2006–2007	2007–2008	2008–2009	2009–2010	2010–2011
Total FTES	1,585	1,690	1,735	1,935	2,021	1,890

Table 2.76

104. Calculate the mean, median, standard deviation, the first quartile, the third quartile, and the *IQR*. Round to one decimal place.

105. Construct a box plot for the FTES for 2005–2006 through 2010–2011 and a box plot for the FTES for 1976–1977 through 2004–2005.

106. Compare the *IQR* for the FTES for 1976–1977 through 2004–2005 with the *IQR* for the FTES for 2005-2006 through 2010–2011. Why do you suppose the *IQR*s are so different?

107. Three students were applying to the same graduate school. They came from schools with different grading systems. Which student had the best GPA when compared to other students at his school? Explain how you determined your answer.

Student	GPA	School Average GPA	School Standard Deviation
Thuy	2.7	3.2	.8
Vichet	87	75	20
Kamala	8.6	8	.4

Table 2.77

108. A music school has budgeted to purchase three musical instruments. The school plans to purchase a piano costing $3,000, a guitar costing $550, and a drum set costing $600. The mean cost for a piano is $4,000 with a standard deviation of $2,500. The mean cost for a guitar is $500 with a standard deviation of $200. The mean cost for drums is $700 with a standard deviation of $100. Which cost is the lowest when compared to other instruments of the same type? Which cost is the highest when compared to other instruments of the same type? Justify your answer.

109. An elementary school class ran one mile with a mean of 11 minutes and a standard deviation of three minutes. Rachel, a student in the class, ran one mile in eight minutes. A junior high school class ran one mile with a mean of nine minutes and a standard deviation of two minutes. Kenji, a student in the class, ran one mile in 8.5 minutes. A high school class ran one mile with a mean of seven minutes and a standard deviation of four minutes. Nedda, a student in the class, ran one mile in eight minutes.

a. Why is Kenji considered a better runner than Nedda even though Nedda ran faster than he?
b. Who is the fastest runner with respect to his or her class? Explain why.

110. Scientists are studying a particular disease. They found that countries that have the highest rates of people who have ever been diagnosed with this disease range from 11.4 percent to 74.6 percent.

Percentage of Population with Disease	Number of Countries
11.4–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2
56.45–65.45	1

Table 2.78

Percentage of Population with Disease	Number of Countries
65.45–74.45	0
74.45–83.45	1

Table 2.78

What is the best estimate of the average percentage of people with the disease for these countries? What is the standard deviation for the listed rates? The United States has an average disease rate of 33.9 percent. Is this rate above average or below? How *unusual* is the U.S. obesity rate compared to the average rate? Explain.

111. Table 2.79 gives the percentage of children under age five diagnosed with a specific medical condition.

Percentage of Children with the Condition	Number of Countries
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

Table 2.79

What is the best estimate for the mean percentage of children with the condition? What is the standard deviation? Which interval(s) could be considered unusual? Explain.

Bringing It Together: Homework

112. Santa Clara County, California, has approximately 27,873 Japanese Americans. Table 2.80 shows their ages by group and each age-group's percentage of the Japanese American community.

Age-Group	Percentage of Community
0–17	18.9
18–24	8.0
25–34	22.8
35–44	15.0
45–54	13.1
55–64	11.9
65+	10.3

Table 2.80

a. Construct a histogram of the Japanese American community in Santa Clara County. The bars will **not** be the same width for this example. Why not? What impact does this have on the reliability of the graph?
b. What percentage of the community is under age 35?
c. Which box plot most resembles the information above?

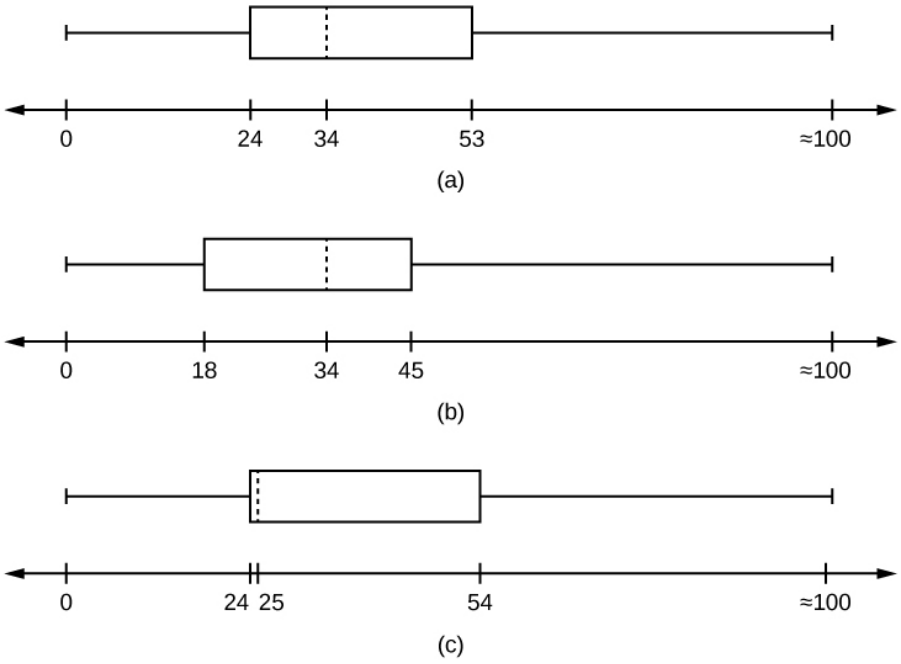


Figure 2.49

113. Javier and Ercilia are supervisors at a shopping mall. Each was given the task of estimating the mean distance that shoppers live from the mall. They each randomly surveyed 100 shoppers. The samples yielded the following information.

	Javier	Ercilia
$\overline{x}$	6.0 miles	6.0 miles
s	4.0 miles	7.0 miles

Table 2.81

a. How can you determine which survey was correct?
b. Explain what the difference in the results of the surveys implies about the data.
c. If the two histograms depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?

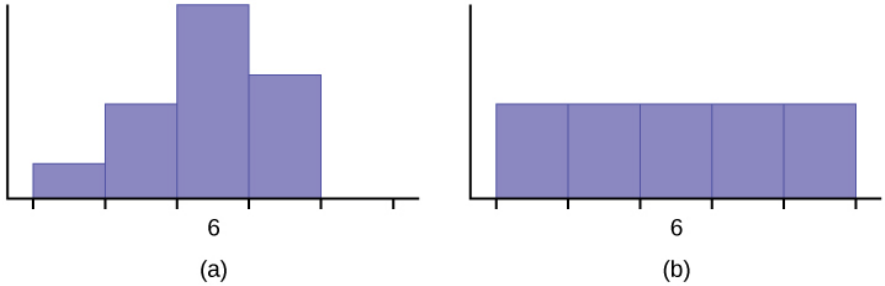


Figure 2.50

d. If the two box plots depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?

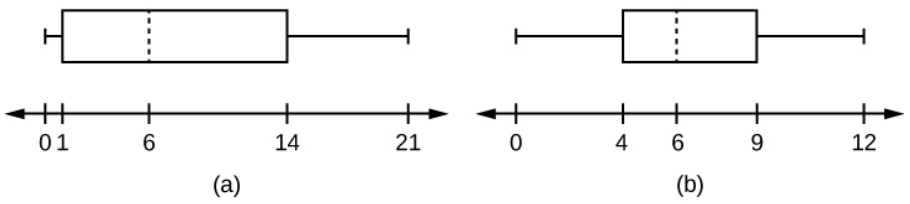


Figure 2.51

Use the following information to answer the next three exercises: We are interested in the number of years students in a particular elementary statistics class have lived in California. The information in the following table is from the entire section.

Number of Years	Frequency	Number of Years	Frequency
7	1	22	1
14	3	23	1
15	1	26	1
18	1	40	2
19	4	42	2
20	3		
			Total = 20

Table 2.82

114. What is the *IQR*?
- a. 8
- b. 11
- c. 15
- d. 35

115. What is the mode?
- a. 19
- b. 19.5
- c. 14 and 20
- d. 22.65

116. Is this a sample or the entire population?
- a. sample
- b. entire population
- c. neither

117. Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

Number of Movies	Frequency
0	5
1	9
2	6
3	4
4	1

Table 2.83

a. Find the sample mean $\overline{x}$.
b. Find the approximate sample standard deviation, *s*.

118. Forty randomly selected students were asked the number of pairs of sneakers they owned. Let *X* = the number of pairs of sneakers owned. The results are as follows:

X	Frequency
1	2
2	5
3	8
4	12
5	12
6	0
7	1

Table 2.84

a. Find the sample mean, $\overline{x}$
b. Find the sample standard deviation, *s*.
c. Construct a histogram of the data.
d. Complete the columns of the chart.
e. Find the first quartile.
f. Find the median.
g. Find the third quartile.
h. Construct a box plot of the data.
i. What percentage of the students owned at least five pairs?
j. Find the 40^{th} percentile.
k. Find the 90^{th} percentile.
l. Construct a line graph of the data.
m. Construct a stemplot of the data.

119. Following are the published weights (in pounds) of all of the football team members of the San Francisco 49ers from a previous year:
177, 205, 210, 210, 232, 205, 185, 185, 178, 210, 206, 212, 184, 174, 185, 242, 188, 212, 215, 247, 241, 223, 220,

260, 245, 259, 278, 270, 280, 295, 275, 285, 290, 272, 273, 280, 285, 286, 200, 215, 185, 230, 250, 241, 190, 260, 250, 302, 265, 290, 276, 228, 265

a. Organize the data from smallest to largest value.
b. Find the median.
c. Find the first quartile.
d. Find the third quartile.
e. Construct a box plot of the data.
f. The middle 50 percent of the weights are from _____ to _____.
g. If our population were all professional football players, would the above data be a sample of weights or the population of weights? Why?
h. If our population included every team member who ever played for a California-based football team, would the above data be a sample of weights or the population of weights? Why?
i. Assume the population was a California-based football team. Find
   i. the population mean, μ,
   ii. the population standard deviation, σ, and
   iii. the weight that is two standard deviations below the mean.
   iv. In addition, when the team's most famous quarterback, played football, he weighed 205 pounds. Also find how many standard deviations above or below the mean was he?
j. That same year, the mean weight for a player from a Texas football team was 240.08 pounds with a standard deviation of 44.38 pounds. One player weighed in at 209 pounds. With respect to his team, who was lighter, the California quarterback or the Texas player? How did you determine your answer?

120. One hundred teachers attended a seminar on mathematical problem solving. The attitudes of a representative sample of 12 of the teachers were measured before and after the seminar. A positive number for change in attitude indicates that a teacher's attitude toward math became more positive. The 12 change scores are as follows:

3, 8, –1, 2, 0, 5, –3, 1, –1, 6, 5, –2

a. What is the mean change score?
b. What is the standard deviation for this population?
c. What is the median change score?
d. Find the change score that is 2.2 standard deviations below the mean.

121. Refer to Figure 2.52 to determine which of the following are true and which are false. Explain your solution to each part in complete sentences.

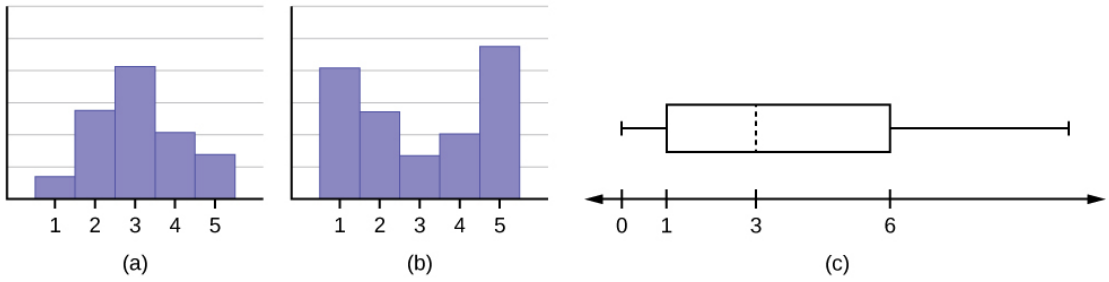


Figure 2.52

a. The medians for all three graphs are the same.
b. We cannot determine if any of the means for the three graphs are different.
c. The standard deviation for Graph b is larger than the standard deviation for Graph a.
d. We cannot determine if any of the third quartiles for the three graphs are different.

122. In a recent issue of the *IEEE Spectrum*, 84 engineering conferences were announced. Four conferences lasted two days. Thirty-six lasted three days. Eighteen lasted four days. Nineteen lasted five days. Four lasted six days. One lasted seven days. One lasted eight days. One lasted nine days. Let X = the length (in days) of an engineering conference.

a. Organize the data in a chart.
b. Find the median, the first quartile, and the third quartile.

c. Find the 65th percentile.
d. Find the 10th percentile.
e. Construct a box plot of the data.
f. The middle 50 percent of the conferences last from _____ days to _____ days.
g. Calculate the sample mean of days of engineering conferences.
h. Calculate the sample standard deviation of days of engineering conferences.
i. Find the mode.
j. If you were planning an engineering conference, which would you choose as the length of the conference, mean, median, or mode? Explain why you made that choice.
k. Give two reasons why you think that three to five days seem to be popular lengths of engineering conferences.

123. A survey of enrollment at 35 community colleges across the United States yielded the following figures: 6,414; 1,550; 2,109; 9,350; 21,828; 4,300; 5,944; 5,722; 2,825; 2,044; 5,481; 5,200; 5,853; 2,750; 10,012; 6,357; 27,000; 9,414; 7,681; 3,200; 17,500; 9,200; 7,380; 18,314; 6,557; 13,713; 17,768; 7,493; 2,771; 2,861; 1,263; 7,285; 28,165; 5,080; 11,622

a. Organize the data into a chart with five intervals of equal width. Label the two columns *Enrollment* and *Frequency*.
b. Construct a histogram of the data.
c. If you were to build a new community college, which piece of information would be more valuable: the mode or the mean?
d. Calculate the sample mean.
e. Calculate the sample standard deviation.
f. A school with an enrollment of 8,000 would be how many standard deviations away from the mean?

Use the following information to answer the next two exercises. X = the number of days per week that 100 clients use a particular exercise facility.

X	Frequency
0	3
1	12
2	33
3	28
4	11
5	9
6	4

Table 2.85

124. The 80th percentile is _____.
a. 5
b. 80
c. 3
d. 4

125. The number that is 1.5 standard deviations **below** the mean is approximately _____.
a. 0.7
b. 4.8
c. –2.8

d. cannot be determined

126. Suppose that a publisher conducted a survey asking adult consumers the number of fiction paperback books they had purchased in the previous month. The results are summarized in Table 2.86.

Number of Books	Frequency	Relative Frequency
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

Table 2.86

a. Are there any outliers in the data? Use an appropriate numerical test involving the *IQR* to identify outliers, if any, and clearly state your conclusion.
b. If a data value is identified as an outlier, what should be done about it?
c. Are any data values farther than two standard deviations away from the mean? In some situations, statisticians may use this criterion to identify data values that are unusual, compared to the other data values. Note that this criterion is most appropriate to use for data that is mound shaped and symmetric rather than for skewed data.
d. Do Parts a and c of this problem give the same answer?
e. Examine the shape of the data. Which part, a or c, of this question gives a more appropriate result for this data?
f. Based on the shape of the data, which is the most appropriate measure of center for this data, mean, median, or mode?

References

2.1 Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

Burbary, K. (2011, March 7). Facebook demographics revisited – 2001 statistics. *Social Media Today*. Retrieved from http://www.kenburbary.com/2011/03/facebook-demographics-revisited-2011-statistics-2/

Centers for Disease Control and Prevention. (n.d.). *Overweight and obesity: Adult obesity facts*. Available online http://www.cdc.gov/obesity/data/adult.html

CollegeBoard. (2013). *The 9th annual AP report to the nation*. Retrieved from http://apreport.collegeboard.org/goals-andfindings/promoting-equity

2.2 Histograms, Frequency Polygons, and Time Series Graphs

Bureau of Labor Statistics, U.S. Department of Labor. (n.d.). *Consumer price index*. Retrieved from https://www.bls.gov/cpi/

CIA World Factbook. (n.d.). *Demographics: Children under the age of 5 years underweight*. Available at http://www.indexmundi.com/g/r.aspx?t=50&v=2224&aml=en

Centers for Disease Control and Prevention. (n.d.). *Overweight and obesity: Adult obesity facts*. Available online http://www.cdc.gov/obesity/data/adult.html

Food and Agriculture Organization of the United Nations. (n.d.). *Food security statistics*. Retrieved from http://www.fao.org/economic/ess/ess-fs/en/

General Register Office for Scotland. *Births time series data*. (2013). Retrieved from http://www.gro-scotland.gov.uk/statistics/theme/vital-events/births/time-series.html

Gunst, R., and Mason, R. (1980). *Regression analysis and its application: A data-oriented approach*. Boca Raton, FL: CRC Press.

Sandbox Networks. (2007). *Presidents*. Available online at http://www.factmonster.com/ipka/A0194030.html

Scholastic. (2013). *Timeline: Guide to the U.S. presidents*. Retrieved from http://www.scholastic.com/teachers/article/timeline-guide-us-presidents

World Bank Group. (2013). *DataBank: CO2 emissions (kt)*. Retrieved from http://databank.worldbank.org/data/home.aspx

2.3 Measures of the Location of the Data

Cauchon, D., and Overberg, P. (2012). Census data shows minorities now a majority of U.S. births. *USA Today*. Retrieved from http://usatoday30.usatoday.com/news/nation/story/2012-05-17/minority-birthscensus/55029100/1

The Mercury News. (n.d.). Retrieved from http://www.mercurynews.com/

Time. (n.d.). Survey by Yankelovich Partners, Inc.

U.S. Census Bureau. (1990). *1990 census*. Retrieved from http://www.census.gov/main/www/cen1990.html

U.S. Census Bureau. (n.d.). *Data*. Retrieved from http://www.census.gov/

2.4 Box Plots

West Magazine. (n.d.). Retrieved from https://westmagazine.net/

2.5 Measures of the Center of the Data

CIA World Factbook. (n.d.). *Obesity – adult prevalence rate*. Available at http://www.indexmundi.com/g/r.aspx?t=50&v=2228&l=en

World Bank Group. (n.d.). Retrieved from http://www.worldbank.org

2.7 Measures of the Spread of the Data

King, B. (2005, Dec.). *Graphically Speaking*. Retrieved from http://www.ltcc.edu/web/about/institutional-research

Microsoft Bookshelf. (n.d.).

Solutions

1.

Stem	Leaf
1	9 9 9
2	0 1 1 5 5 5 6 6 8 9
3	1 1 2 2 3 4 5 6 7 7 8 8 8 8
4	1 3 3

Table 2.87

3.

Stem	Leaf
2	5 5 6 7 7 8
3	0 0 1 2 3 3 5 5 5 7 7 9
4	1 6 9
5	6 7 7
6	1

Table 2.88

5.

Figure 2.53

7.

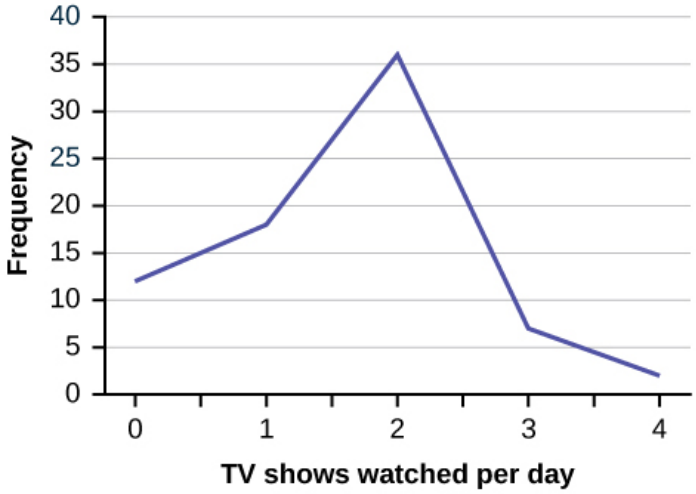


Figure 2.54

9.

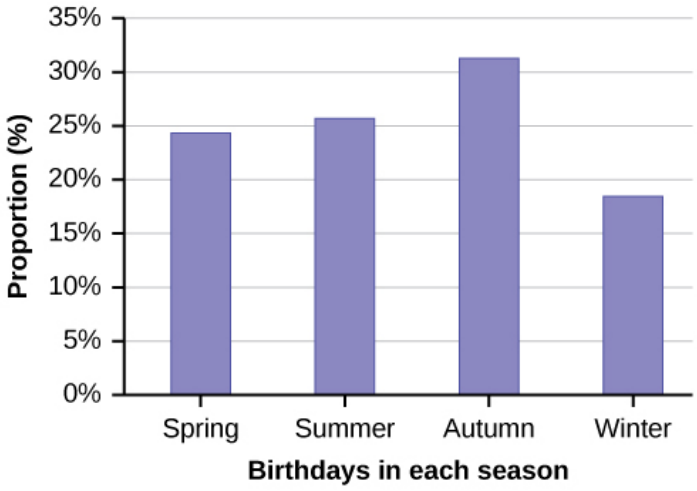


Figure 2.55

11.

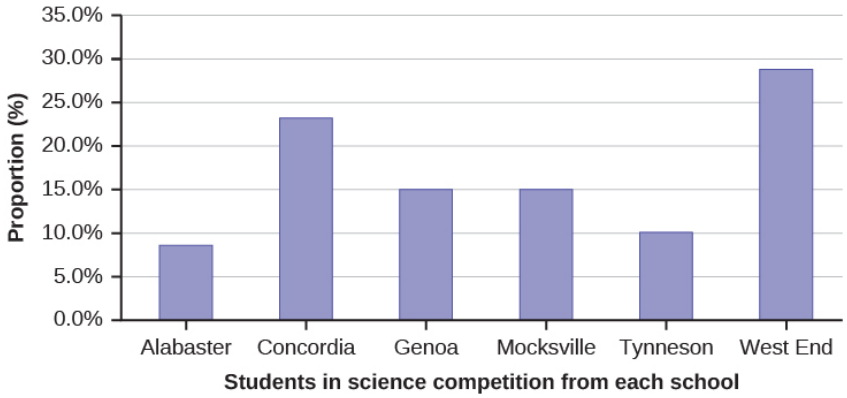


Figure 2.56

13. 65

15. The relative frequency shows the *proportion* of data points that have each value. The frequency tells the *number* of data points that have each value.

17. Answers will vary. One possible histogram is shown below.

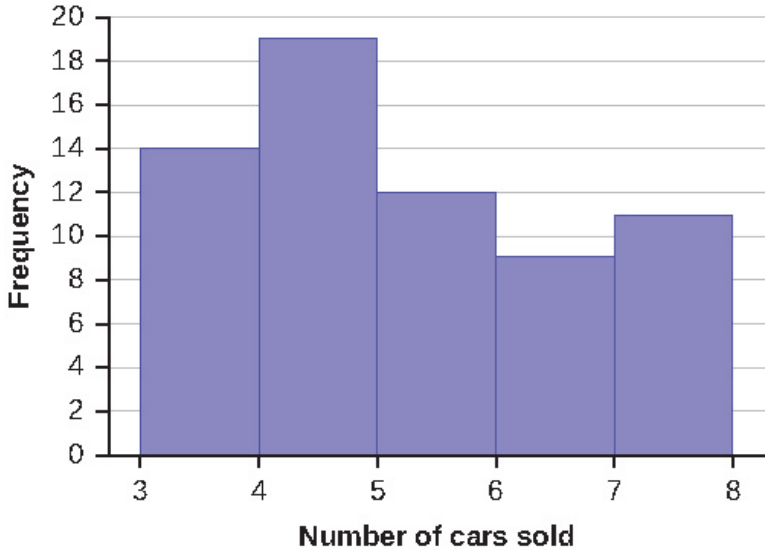


Figure 2.57

19. Find the midpoint for each class. These will be graphed on the *x*-axis. The frequency values will be graphed on the *y*-axis values.

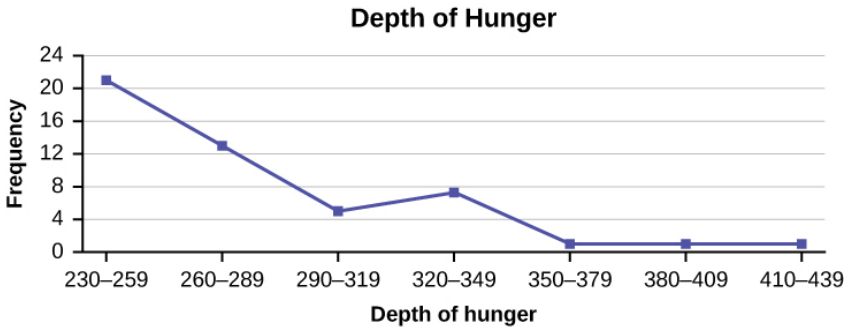


Figure 2.58

21.

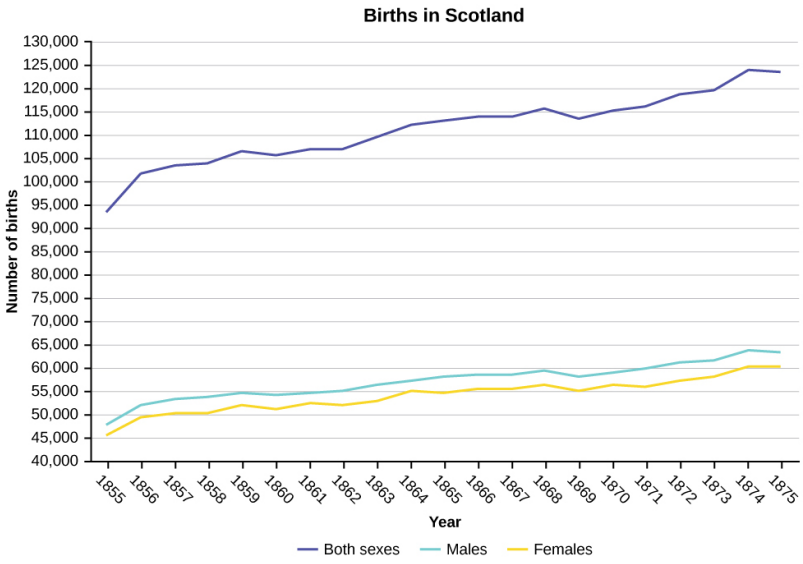


Figure 2.59

23. a. The 40th percentile is 37 years.
b. The 78th percentile is 70 years.

25. Jesse graduated 37th out of a class of 180 students. There are 180 – 37 = 143 students ranked below Jesse. There is one rank of 37.
x = 143 and y = 1. $\frac{x+.5y}{n}(100) = \frac{143+.5(1)}{180}(100) = 79.72$. Jesse's rank of 37 puts him at the 80th percentile.

27. a. For runners in a race, it is more desirable to have a high percentile for speed. A high percentile means a higher speed, which is faster.
b. 40 percent of runners ran at speeds of 7.5 miles per hour or less (slower), and 60 percent of runners ran at speeds of 7.5 miles per hour or more (faster).

29. When waiting in line at the DMV, the 85th percentile would be a long wait time compared to the other people waiting. 85 percent of people had shorter wait times than Mina. In this context, Mina would prefer a wait time corresponding to a lower percentile. 85 percent of people at the DMV waited 32 minutes or less. 15 percent of people at the DMV waited 32 minutes or longer.

31. The manufacturer and the consumer would be upset. This is a large repair cost for the damages, compared to the other cars in the sample. INTERPRETATION: 90 percent of the crash-tested cars had damage repair costs of $1,700 or less; only 10 percent had damage repair costs of $1,700 or more.

33. You can afford 34 percent of houses. 66 percent of the houses are too expensive for your budget. INTERPRETATION: 34 percent of houses cost $240,000 or less; 66 percent of houses cost $240,000 or more.

35. 4

37. 6 – 4 = 2

39. 6

41. More than 25 percent of salespersons sell four cars in a typical week. You can see this concentration in the box plot because the first quartile is equal to the median. The top 25 percent and the bottom 25 percent are spread out evenly; the whiskers have the same length.

43. Mean: 16 + 17 + 19 + 20 + 20 + 21 + 23 + 24 + 25 + 25 + 25 + 26 + 26 + 27 + 27 + 27 + 28 + 29 + 30 + 32 + 33 + 33 + 34 + 35 + 37 + 39 + 40 = 738;
$\frac{738}{27}$ = 27.33

45. The most frequent lengths are 25 and 27, which occur three times. Mode = 25, 27

47. 4

49. The data are symmetrical. The median is 3, and the mean is 2.85. They are close, and the mode lies close to the middle of the data, so the data are symmetrical.

51. The data are skewed right. The median is 87.5, and the mean is 88.2. Even though they are close, the mode lies to the left of the middle of the data, and there are many more instances of 87 than any other number, so the data are skewed right.

53. When the data are symmetrical, the mean and median are close or the same.

55. The distribution is skewed right because it looks pulled out to the right.

57. The mean is 4.1 and is slightly greater than the median, which is 4.

59. The mode and the median are the same. In this case, both 5.

61. The distribution is skewed left because it looks pulled out to the left.

63. Both the mean and the median are 6.

65. The mode is 12, the median is 13.5, and the mean is 15.1. The mean is the largest.

67. The mean tends to reflect skewing the most because it is affected the most by outliers.

69. sampling variability

70. induced variability

71. measurement variability

72. natural variability

73. $s = 34.5$

75. For Fredo: $z = \frac{.158 - .166}{.012} = -0.67$.
For Karl: $z = \frac{.177 - .189}{.015} = -.8$.
Fredo's z score of –.67 is higher than Karl's z score of –.8. For batting average, higher values are better, so Fredo has a better batting average compared to his team.

77. a. $s_x = \sqrt{\frac{\sum fm^2}{n} - \overline{x}^2} = \sqrt{\frac{193{,}157.45}{30} - 79.5^2} = 10.88$

b. $s_x = \sqrt{\frac{\sum fm^2}{n} - \overline{x}^2} = \sqrt{\frac{380{,}945.3}{101} - 60.94^2} = 7.62$

c. $s_x = \sqrt{\frac{\sum fm^2}{n} - \overline{x}^2} = \sqrt{\frac{440{,}051.5}{86} - 70.66^2} = 11.14$

79. a. Example solution for using the random number generator for the TI-84+ to generate a simple random sample of eight states. Instructions are as follows.

Number the entries in the table 1–51 (includes Washington, DC; numbered vertically)
Press MATH
Arrow over to PRB
Press 5:randInt(
Enter 51,1,8)

Eight numbers are generated (use the right arrow key to scroll through the numbers). The numbers correspond to the numbered states (for this example: {47 21 9 23 51 13 25 4}. If any numbers are repeated, generate a different number by using 5:randInt(51,1)). Here, the states (and Washington DC) are {Arkansas, Washington DC, Idaho, Maryland, Michigan, Mississippi, Virginia, Wyoming}.

Corresponding percents are {30.1, 22.2, 26.5, 27.1, 30.9, 34.0, 26.0, 25.1}.

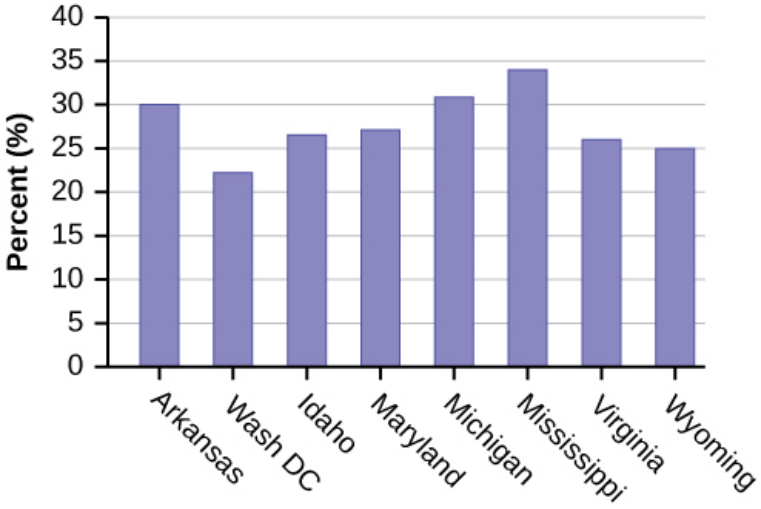


Figure 2.60

b.

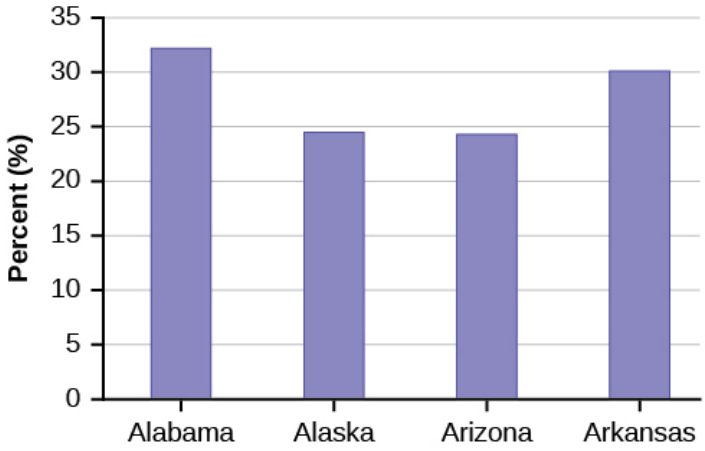


Figure 2.61

c.

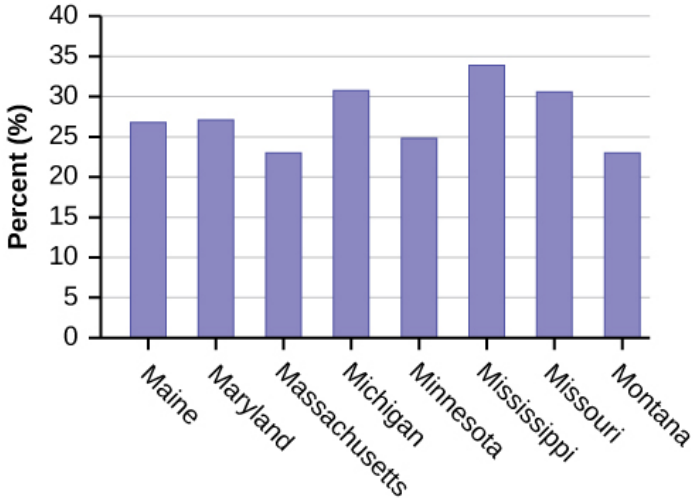


Figure 2.62

81.

Amount($)	Frequency	Relative Frequency
51–100	5	.08
101–150	10	.17
151–200	15	.25
201–250	15	.25
251–300	10	.17
301–350	5	.08

Table 2.89 Singles

Amount ($)	Frequency	Relative Frequency
100–150	5	.07
201–250	5	.07
251–300	5	.07
301–350	5	.07
351–400	10	.14
401–450	10	.14
451–500	10	.14
501–550	10	.14
551–600	5	.07
601–650	5	.07

Table 2.90 Couples

a. See Table 2.89 and Table 2.90.
b. In the following histogram, data values that fall on the right boundary are counted in the class interval, while values that fall on the left boundary are not counted, with the exception of the first interval, where both boundary values are included.

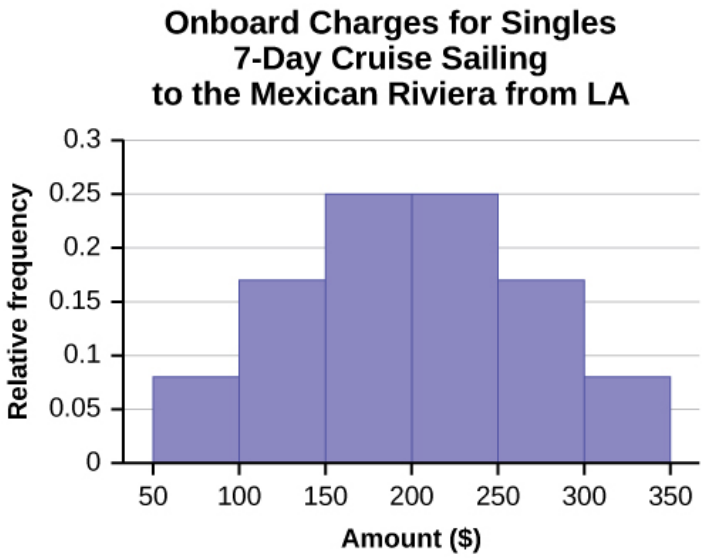


Figure 2.63

c. In the following histogram, the data values that fall on the right boundary are counted in the class interval, while values that fall on the left boundary are not counted, with the exception of the first interval, where values on both boundaries are included.

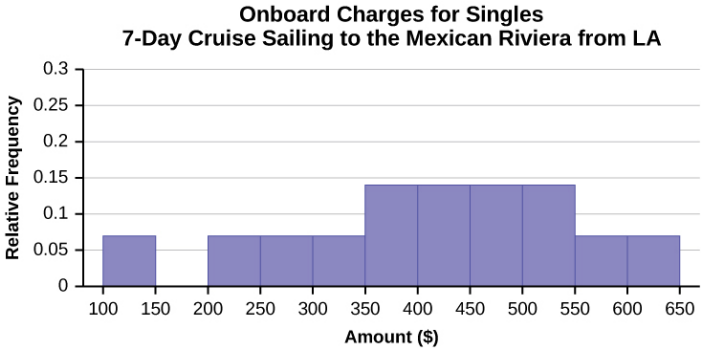


Figure 2.64

d. Compare the two graphs.
   i. Answers may vary. Possible answers include the following:
   - Both graphs have a single peak.
   - Both graphs use class intervals with width equal to $50

   ii. Answers may vary. Possible answers include the following:
   - The couples graph has a class interval with no values
   - It takes almost twice as many class intervals to display the data for couples

   iii. Answers may vary. Possible answers include the following. The graphs are more similar than different because the overall patterns for the graphs are the same.

e. Check student's solution.

f. Compare the graph for the singles with the new graph for the couples:
   - Both graphs have a single peak
   - Both graphs display six class intervals
   - Both graphs show the same general pattern

   ii. Answers may vary. Possible answers include the following. Although the width of the class intervals for couples is double that of the class intervals for singles, the graphs are more similar than they are different.

g. Answers may vary. Possible answers include the following. You are able to compare the graphs interval by interval. It is easier to compare the overall patterns with the new scale on the couples graph. Because a couple

represents two individuals, the new scale leads to a more accurate comparison.

h. Answers may vary. Possible answers include the following. Based on the histograms, it seems that spending does not vary much from singles to individuals who are part of a couple. The overall patterns are the same. The range of spending for couples is approximately double the range for individuals.

83. c

85. Answers will vary.

87. a. 1 – (.02+.09+.19+.26+.18+.17+.02+.01) = .06

b. .19+.26+.18 = .63

c. Check student's solution.

d. 40^{th} percentile will fall between 30,000 and 40,000

80^{th} percentile will fall between 50,000 and 75,000

e. Check student's solution.

89. a. more children; the left whisker shows that 25 percent of the population are children 17 and younger; the right whisker shows that 25 percent of the population are adults 50 and older, so adults 65 and over represent less than 25 percent

b. 62.4 percent

91. a. Answers will vary. Possible answer: State University conducted a survey to see how involved its students are in community service. The box plot shows the number of community service hours logged by participants over the past year.

b. Because the first and second quartiles are close, the data in this quarter is very similar. There is not much variation in the values. The data in the third quarter is much more variable, or spread out. This is clear because the second quartile is so far away from the third quartile.

93. a. Each box plot is spread out more in the greater values. Each plot is skewed to the right, so the ages of the top 50 percent of buyers are more variable than the ages of the lower 50 percent.

b. The black sports car is most likely to have an outlier. It has the longest whisker.

c. Comparing the median ages, younger people tend to buy the black sports car, while older people tend to buy the white sports car. However, this is not a rule, because there is so much variability in each data set.

d. The second quarter has the smallest spread. There seems to be only a three-year difference between the first quartile and the median.

e. The third quarter has the largest spread. There seems to be approximately a 14-year difference between the median and the third quartile.

f. $IQR \sim 17$ years

g. There is not enough information to tell. Each interval lies within a quarter, so we cannot tell exactly where the data in that quarter is are concentrated.

h. The interval from 31 to 35 years has the fewest data values. Twenty-five percent of the values fall in the interval 38 to 41, and 25 percent fall between 41 and 64. Since 25 percent of values fall between 31 and 38, we know that fewer than 25 percent fall between 31 and 35.

96. the mean percentage, $\overline{x} = \frac{1{,}328.65}{50} = 26.75$

98. The median value is the middle value in the ordered list of data values. The median value of a set of 11 will be the sixth number in order. Six years will have totals at or below the median.

100. 474 FTES

102. 919

104.
- mean = 1,809.3
- median = 1,812.5
- standard deviation = 151.2
- first quartile = 1,690
- third quartile = 1,935
- *IQR* = 245

106. Hint: think about the number of years covered by each time period and what happened to higher education during those periods.

108. For pianos, the cost of the piano is .4 standard deviations BELOW the mean. For guitars, the cost of the guitar is 0.25 standard deviations ABOVE the mean. For drums, the cost of the drum set is 1.0 standard deviations BELOW the mean. Of the three, the drums cost the lowest in comparison to the cost of other instruments of the same type. The guitar costs the most in comparison to the cost of other instruments of the same type.

110.
- $\overline{x} = 23.32$
- Using the TI 83/84, we obtain a standard deviation of: $s_x = 12.95$.
- The obesity rate of the United States is 10.58 percent higher than the average obesity rate.
- Since the standard deviation is 12.95, we see that 23.32 + 12.95 = 36.27 is the disease percentage that is one standard deviation from the mean. The U.S. disease rate is slightly less than one standard deviation from the mean. Therefore, we can assume that the United States, although 34 percent have the disease, does not have an unusually high percentage of people with the disease.

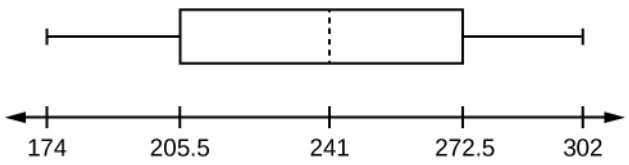
112.
a. For graph, check student's solution.
b. 49.7 percent of the community is under the age of 35
c. Based on the information in the table, graph (a) most closely represents the data.

114. a

116. b

117.
a. 1.48
b. 1.12

119.
a. 174, 177, 178, 184, 185, 185, 185, 185, 188, 190, 200, 205, 205, 206, 210, 210, 210, 212, 212, 215, 215, 220, 223, 228, 230, 232, 241, 241, 242, 245, 247, 250, 250, 259, 260, 260, 265, 265, 270, 272, 273, 275, 276, 278, 280, 280, 285, 285, 286, 290, 290, 295, 302
b. 241
c. 205.5
d. 272.5
e.



f. 205.5, 272.5
g. sample
h. population
i.
   i. 236.34
   ii. 37.50
   iii. 161.34
   iv. .84 standard deviations below the mean
j. young

121. a. true
b. true
c. true
d. false

123. a.

Enrollment	Frequency
1,000–5,000	10
5,000–10,000	16
10,000–15,000	3
15,000–20,000	3
20,000–25,000	1
25,000–30,000	2

Table 2.91

b. Check student's solution.
c. mode
d. 8,628.74
e. 6,943.88
f. –0.09

125. a

3 PROBABILITY TOPICS

Figure 3.1 Meteor showers are rare, but the probability of them occurring can be calculated. (credit: Navicore/flickr)

Chapter Objectives

By the end of this chapter, the student should be able to do the following:

- Understand and use the terminology of probability
- Determine whether two events are mutually exclusive and whether two events are independent
- Calculate probabilities using the addition rules and multiplication rules
- Construct and interpret contingency tables
- Construct and interpret Venn diagrams
- Construct and interpret tree diagrams

Introduction

It is often necessary to *guess* about the outcome of an event in order to make a decision. Politicians study polls to guess their likelihood of winning an election. Teachers choose a particular course of study based on what they think students can comprehend. Doctors choose the treatments needed for various diseases based on their assessment of likely results. You may have visited a casino where people play games chosen because of the belief that the likelihood of winning is good. You may have chosen your course of study based on the probable availability of jobs.

You have, more than likely, used probability. In fact, you probably have an intuitive sense of probability. Probability deals with the chance of an event occurring. Whenever you weigh the odds of whether or not to do your homework or to study for an exam, you are using probability. In this chapter, you will learn how to solve probability problems using a systematic approach.

COLLABORATIVE EXERCISE

How likely is it that a randomly chosen person in your class has change in his or her pocket? Would you say that it is very likely? Somewhat likely? Not likely?

How likely is it that a randomly chosen person in your class has ridden a bus in the past month?

If a person is chosen at random from your classroom and you know that he or she has ridden a bus in the past month, do you think that person is more likely or less likely to have change?

Probability theory allows us to measure how likely—or unlikely—a given result is.

Your instructor will survey your class. Count the number of students in the class today.

- Raise your hand if you have any change in your pocket or purse. Record the number of raised hands.
- Raise your hand if you rode a bus within the past month. Record the number of raised hands.
- Raise your hand if you answered *yes* to BOTH of the first two questions. Record the number of raised hands.

Use the class data as estimates of the following probabilities. *P*(change) means the probability that a randomly chosen person in your class has change in his/her pocket or purse. *P*(bus) means the probability that a randomly chosen person in your class rode a bus within the last month and so on. Discuss your answers.

- Find *P*(change).
- Find *P*(bus).
- Find *P*(change AND bus). Find the probability that a randomly chosen student in your class has change in his/her pocket or purse and rode a bus within the last month.
- Find *P*(change|bus). Find the probability that a randomly chosen student has change given that he or she rode a bus within the last month. Count all the students who rode a bus. From the group of students who rode a bus, count those who have change. The probability is equal to those who have change and rode a bus divided by those who rode a bus.

3.1 Terminology

Probability is a measure that is associated with how certain we are of results, or outcomes, of a particular activity. When the activity is a planned operation carried out under controlled conditions, it is called an **experiment**. If the result is *not* predetermined, then the experiment is said to be a chance experiment. Each time the experiment is attempted is called a trial.

Examples of chance experiments include the following:

- flipping a fair coin,
- spinning a spinner,
- drawing a marble at random from a bag, and
- rolling a pair of dice.

A result of an experiment is called an **outcome**. The **sample space** of an experiment is the set, or collection, of all possible outcomes.

There are four main ways to represent a sample space:

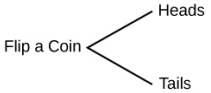
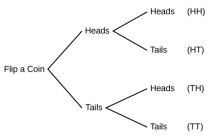
	Flipping a Fair Coin	Flipping Two Fair Coins
Systematic List of Outcomes	heads (H) tails (T)	HH HT TH TT
Tree Diagram*	 **Figure 3.2**	 **Figure 3.3**

Table 3.1

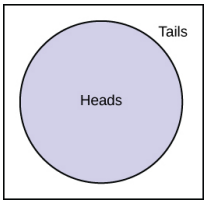
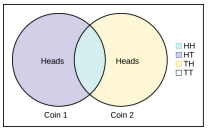
	Flipping a Fair Coin	Flipping Two Fair Coins
Venn Diagram*	 **Figure 3.4**	 **Figure 3.5**
Set Notation	$S = \{H, T\}$	$S = \{HH, HT, TH, TT\}$

Table 3.1

*We will investigate tree diagrams and Venn diagrams in Section 3.5.

Note—when represented as a set, the sample space is denoted with an uppercase *S*.

An event is any combination of outcomes. It is a subset of the sample space, so uppercase letters like *A* and *B* are commonly used to represent events. For example, if the experiment is to flip three fair coins, event *A* might be getting at most one head.

The probability of an event *A* is written *P*(*A*), and $0 \leq P(A) \leq 1$.$P(A) = 0$ means the event *A* can never happen. *P*(*A*) = 1 means the event *A* always happens. $P(A) = 0.5$ means the event *A* is **equally likely** to occur or not to occur.

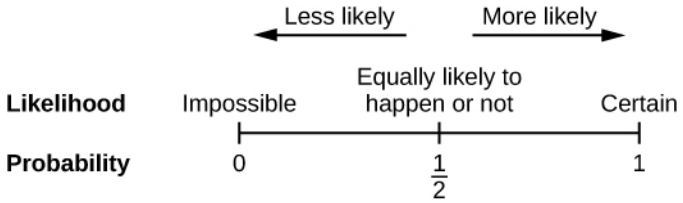


Figure 3.6

If two outcomes or events are equally likely, then they have equal probability. For example, if you toss a fair, six-sided die, each face (1, 2, 3, 4, 5, or 6) is as likely to occur as any other face. If you toss a fair coin, a Head (*H*) and a Tail (*T*) are equally likely to occur. If you randomly guess the answer to a true/false question on an exam, you are equally likely to select a correct answer or an incorrect answer.

To calculate the probability of an event *A* when all outcomes in the sample space are equally likely, count the number of outcomes for event *A* and divide by the total number of outcomes in the sample space. This is known as the theoretical probability of *A*.

Theoretical Probability of Event A

$$P(A) = \frac{\text{Number of outcomes in event A}}{\text{Total number of possible outcomes.}} \quad \textbf{3.1}$$

For example, if you toss a fair dime and a fair nickel, the sample space is {*HH, TH, HT, TT*} where *T* = tails and *H* = heads. The sample space has four outcomes. Let *A* represent the outcome *getting one head.* There are two outcomes that meet this condition {*HT, TH*}, so $P(A) = \frac{2}{4} = \frac{1}{2} = .5$.

Theoretical probability is not sufficient in all situations, however. Suppose we want to calculate the probability that a randomly selected car will run a red light at a given intersection. In this case, we need to look at events that *have* occurred, not theoretical possibilities. We could install a traffic camera and count the number of times that cars failed to stop when the light was red and the total number of cars that passed through the intersection for a period of time. These data will allow us to calculate the experimental, or empirical, probability that a car runs the red light.

Experimental Probability of Event A

$$P(A) = \frac{\text{Number of times event A occurs.}}{\text{Total number of trials}} \quad \textbf{3.2}$$

While theoretical and experimental methods provide two different ways to calculate probability, these methods are closely related. If you flip one fair coin, there is one way to obtain heads and two possible outcomes. So, the theoretical probability of heads is $\frac{1}{2}$. Probability does not predict short-term results, however. If an experiment involves flipping a coin 10 times, you should not expect exactly five heads and five tails. The probability of any outcome measures the long-term relative frequency of that outcome. If you continue to flip the coin (from 20 to 2,000 to 20,000 times) the relative frequency of heads approaches .5 (the probability of heads).This important characteristic of probability experiments is known as the law of large numbers, which states that as the number of repetitions of an experiment is increased, the relative frequency obtained in the experiment tends to become closer and closer to the theoretical probability. Even though the outcomes do not happen according to any set pattern or order, overall, the long-term observed, or empirical, relative frequency will approach the theoretical probability.

Suppose you roll one fair, six-sided die with the numbers {1, 2, 3, 4, 5, 6} on its faces. Let event *E* = rolling a number that is at least five. There are two outcomes {5, 6}. $P(E) = \frac{2}{6}$. If you were to roll the die only a few times, you would not be surprised if your observed results did not match the probability. If you were to roll the die a very large number of times, you would expect that, overall, $\frac{2}{6}$ of the rolls would result in an outcome of *at least five*. You would not expect exactly $\frac{2}{6}$, but the long-term relative frequency of obtaining this result would approach the theoretical probability of $\frac{2}{6}$ as the number of repetitions grows larger and larger.

It is important to realize that in many situations, the outcomes are not equally likely. A coin or die may be **unfair**, or **biased**. Two math professors in Europe had their statistics students test the Belgian one-euro coin and discovered that in 250 trials, a head was obtained 56 percent of the time and a tail was obtained 44 percent of the time. The data seem to show that the coin is not a fair coin; more repetitions would be helpful to draw a more accurate conclusion about such bias. Some dice may be biased. Look at the dice in a game you have at home; the spots on each face are usually small holes carved out and then painted to make the spots visible. Your dice may or may not be biased; it is possible that the outcomes may be affected by the slight weight differences due to the different numbers of holes in the faces. Gambling casinos make a lot of money depending on outcomes from rolling dice, so casino dice are made differently to eliminate bias. Casino dice have flat faces; the holes are completely filled with paint having the same density as the material that the dice are made out of so that each face is equally likely to occur. Later we will learn techniques to use to work with probabilities for events that are not equally likely.

OR Event

An outcome is in the event *A* OR *B* if the outcome is in *A* or is in *B* or is in both *A* and *B*. For example, let *A* = {1, 2, 3, 4, 5} and *B* = {4, 5, 6, 7, 8}. *A* OR *B* = {1, 2, 3, 4, 5, 6, 7, 8}. Notice that 4 and 5 are **not** listed twice.

AND Event

An outcome is in the event *A* AND *B* if the outcome is in both *A* and *B* at the same time. For example, let *A* and *B* be {1, 2, 3, 4, 5} and {4, 5, 6, 7, 8}, respectively. Then *A* AND *B* = {4, 5}.

The **complement** of event *A* is denoted *A′* (read "*A* prime"). *A′* consists of all outcomes that are **not** in *A*. Notice that $P(A) + P(A') = 1$. For example, let *S* = {1, 2, 3, 4, 5, 6} and let *A* = {1, 2, 3, 4}. Then, *A′* = {5, 6}. $P(A) = \frac{4}{6}$, $P(A') = \frac{2}{6}$, and $P(A) + P(A') = \frac{4}{6} + \frac{2}{6} = 1$.

The **conditional probability** of *A* given *B* is written $P(A|B)$, read "the probability of *A*, given *B*." $P(A|B)$ is the probability that event *A* will occur given that the event *B* has already occurred. **A conditional probability reduces the sample space**. We calculate the probability of *A* from the reduced sample space *B*. The formula to calculate $P(A|B)$ is $P(A|B) = \frac{P(A \text{ AND } B)}{P(B)}$ where $P(B)$ is greater than zero.

For example, suppose we toss one fair, six-sided die. The sample space *S* = {1, 2, 3, 4, 5, 6}. Let *A* = {2, 3} and *B* = {2, 4, 6}. $P(A|B)$ represents the probability that a randomly selected outcome is in *A given that* it is in *B*. We know that the outcome must lie in *B*, so there are three possible outcomes. There is only one outcome in *B* that also lies in *A*, so $P(A|B) = \frac{1}{3}$.

We get the same result by using the formula. Remember that *S* has six outcomes.

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{\frac{(\text{the number of outcomes that are 2 or 3 and even in } S)}{6}}{\frac{(\text{the number of outcomes that are even in } S)}{6}} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$$

Understanding Terminology and Symbols

It is important to read each problem carefully to think about and understand what the events are. Understanding the wording is the first very important step in solving probability problems. Reread the problem several times if necessary. Clearly identify the event of interest. Determine whether there is a condition stated in the wording that would indicate that the probability is conditional; carefully identify the condition, if any.

EXAMPLE 3.1

Problem

The sample space S is the whole numbers starting at one and less than 20.

a. S = _____
Let event A = the even numbers and event B = numbers greater than 13.

b. A = _____, B = _____
c. $P(A)$ = _____, $P(B)$ = _____
d. A AND B = _____, A OR B = _____
e. $P(A \text{ AND } B)$ = _____, $P(A \text{ OR } B)$ = _____
f. A' = _____, $P(A')$ = _____
g. $P(A) + P(A')$ = _____
h. $P(A|B)$ = _____, $P(B|A)$ = _____; are the probabilities equal?

Solution

a. $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19\}$
b. $A = \{2, 4, 6, 8, 10, 12, 14, 16, 18\}$, $B = \{14, 15, 16, 17, 18, 19\}$
c. $P(A) = \frac{\text{number of outcomes in } A}{\text{number of outcomes in } S} = \frac{9}{19}$, $P(B) = \frac{\text{number of outcomes in } B}{\text{number of outcomes in } S} = \frac{6}{19}$
d. The set A AND B contains all outcomes that lie in both sets A and B, so A AND B = {14,16,18}, The set A OR B contains all outcomes that lie either of the sets A or B, so A OR B = {2, 4, 6, 8, 10, 12, 14, 15, 16, 17, 18, 19}.
e. $P(A \text{ AND } B) = \frac{3}{19}$, $P(A \text{ OR } B) = \frac{12}{19}$
f. A' consists of all outcomes in the sample space, S, that DO NOT lie in A, so A' = 1, 3, 5, 7, 9, 11, 13, 15, 17, 19; $P(A') = \frac{10}{19}$.
g. $P(A) + P(A') = \frac{9}{19} + \frac{10}{19} = 1$
h. $P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{\frac{3}{19}}{\frac{6}{19}} = \frac{3}{6}$, $P(B|A) = \frac{P(A \text{ AND } B)}{P(A)} = \frac{\frac{3}{19}}{\frac{9}{19}} = \frac{3}{9}$, No, the probabilities are not equal.

TRY IT 3.1

The sample space S is all the ordered pairs of two whole numbers, the first from one to three and the second from one to four (Example: (1, 4)).

a. S = _____

Let event A = the sum is even and event B = the first number is prime.
b. A = _____, B = _____
c. $P(A)$ = _____, $P(B)$ = _____
d. A AND B = _____, A OR B = _____
e. $P(A \text{ AND } B)$ = _____, $P(A \text{ OR } B)$ = _____
f. B' = _____, $P(B')$ = _____
g. $P(A) + P(A')$ = _____
h. $P(A|B)$ = _____, $P(B|A)$ = _____; are the probabilities equal?

EXAMPLE 3.2

Problem

A fair, six-sided die is rolled. The sample space, S, is {1, 2, 3, 4, 5, 6}. Describe each event and calculate its probability.

a. Event T = the outcome is two.
b. Event A = the outcome is an even number.
c. Event B = the outcome is less than four.
d. The complement of A
e. A GIVEN B
f. B GIVEN A
g. A AND B
h. A OR B
i. A OR B'
j. Event N = the outcome is a prime number.
k. Event I = the outcome is seven.

Solution

a. $T = \{2\}$, $P(T) = \frac{\text{number of outcomes in } T}{\text{number of outcomes in } S} = \frac{1}{6}$
b. $A = \{2, 4, 6\}$, $P(A) = \frac{3}{6} = \frac{1}{2}$
c. $B = \{1, 2, 3\}$, $P(B) = \frac{3}{6} = \frac{1}{2}$
d. $A' = \{1, 3, 5\}$, $P(A') = \frac{3}{6} = \frac{1}{2}$
e. $A|B = \{2\}$, There are three outcomes in B, and only 1 of these lies in A, so $P(A|B) = \frac{1}{3}$
f. $B|A = \{2\}$, There are three outcomes in A, and only 1 of these lies in B, so $P(B|A) = \frac{1}{3}$
g. A AND $B = \{2\}$, $P(A \text{ AND } B) = \frac{1}{6}$
h. A OR $B = \{1, 2, 3, 4, 6\}$, $P(A \text{ OR } B) = \frac{5}{6}$
i. A OR $B' = \{2, 4, 5, 6\}$, $P(A \text{ OR } B') = \frac{4}{6} = \frac{2}{3}$
j. $N = \{2, 3, 5\}$, $P(N) = \frac{1}{2}$
k. It is impossible to roll a die and get an outcome of 7, so $P(7) = 0$.

EXAMPLE 3.3

Table 3.2 describes the distribution of a random sample S of 100 individuals, organized by gender and whether they are right or left-handed.

	Right-Handed	Left-Handed
Males	43	9
Females	44	4

Table 3.2

Problem

Let's denote the events M = the subject is male, F = the subject is female, R = the subject is right-handed, L = the subject is left-handed. Compute the following probabilities:

a. $P(M)$
b. $P(F)$
c. $P(R)$
d. $P(L)$
e. $P(M \text{ AND } R)$
f. $P(F \text{ AND } L)$
g. $P(M \text{ OR } F)$

h. $P(M \text{ OR } R)$
i. $P(F \text{ OR } L)$
j. $P(M')$
k. $P(R|M)$
l. $P(F|L)$
m. $P(L|F)$

Solution

a. $P(M) = \frac{\text{number of males}}{\text{total number of subjects}} = \frac{43+9}{43+9+44+4} = \frac{52}{100} = .52$
b. $P(F) = \frac{\text{number of females}}{\text{total number of subjects}} = \frac{44+4}{43+9+44+4} = \frac{48}{100} = .48$
c. $P(R) = \frac{\text{number of right-handed subjects}}{\text{total number of subjects}} = \frac{43+44}{43+9+44+4} = \frac{87}{100} = .87$
d. $P(L) = \frac{\text{number of left-handed subjects}}{\text{total number of subjects}} = \frac{9+4}{43+9+44+4} = \frac{13}{100} = .13$
e. $P(M \text{and} R) = \frac{\text{number of male, right-handed subjects}}{\text{total number of subjects}} = \frac{43}{100} = .43$
f. $P(F \text{and} L) = \frac{\text{number of female, left-handed subjects}}{\text{total number of subjects}} = \frac{4}{100} = .04$
g. $P(M \text{or} F) = \frac{\text{number of subjects that are male or female}}{\text{total number of subjects}} = \frac{52+48}{100} = \frac{100}{100} = 1$
h. $P(M \text{or} R) = \frac{\text{number of subjects that are male or right-handed}}{\text{total number of subjects}} = \frac{43+9+44}{100} = \frac{96}{100} = .96$
i. $P(F \text{or} L) = \frac{\text{number of subjects that are female or left-handed}}{\text{total number of subjects}} = \frac{44+4+9}{100} = \frac{57}{100} = .57$
j. $P(M') = \frac{\text{number of subjects who are not male}}{\text{total number of subjects}} = \frac{44+4}{43+9+44+4} = \frac{48}{100} = .48$
k. $P(R|M) = \frac{P(R \text{and} M)}{P(M)} = \frac{0.43}{0.52} \approx .8269$ (rounded to four decimal places)
l. $P(F|L) = \frac{P(F \text{and} L)}{P(L)} = \frac{0.04}{0.13} \approx .3077$ (rounded to four decimal places)
m. $P(L|F) = \frac{P(L \text{and} F)}{P(F)} = \frac{0.04}{0.48} \approx .0833$ (rounded to four decimal places)

3.2 Independent and Mutually Exclusive Events

Independent and mutually exclusive do **not** mean the same thing.

Independent Events

Two events are independent if the following are true:

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(A \text{ AND } B) = P(A)P(B)$

Two events A and B are **independent events** if the knowledge that one occurred does not affect the chance the other occurs. For example, the outcomes of two roles of a fair die are independent events. The outcome of the first roll does not change the probability for the outcome of the second roll. To show two events are independent, you must show **only one** of the above conditions. If two events are **not** independent, then we say that they are **dependent events**.

Sampling may be done **with replacement** or **without replacement**.

- **With replacement**: If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once. When sampling is done with replacement, then events are considered to be independent, meaning the result of the first pick will not change the probabilities for the second pick.

A bag contains four blue and three white marbles. James draws one marble from the bag at random, records the color, and replaces the marble. The probability of drawing blue is $\frac{4}{7}$. When James draws a marble from the bag a second time, the probability of drawing blue is still $\frac{4}{7}$. James replaced the marble after the first draw, so there are still four blue and three white marbles.

Figure 3.7

- **Without replacement**: When sampling is done without replacement, each member of a population may be chosen only once. In this case, the probabilities for the second pick are affected by the result of the first pick. The events are considered to be dependent or not independent.

The bag still contains four blue and three white marbles. Maria draws one marble from the bag at random, records the color, and sets the marble aside. The probability of drawing blue on the first draw is $\frac{4}{7}$. Suppose Maria draws a blue marble and sets it aside. When she draws a marble from the bag a second time, there are now *three* blue and three white marbles. So, the probability of drawing blue is now $\frac{3}{6} = \frac{1}{2}$. Removing the first marble without replacing it influences the probabilities on the second draw.

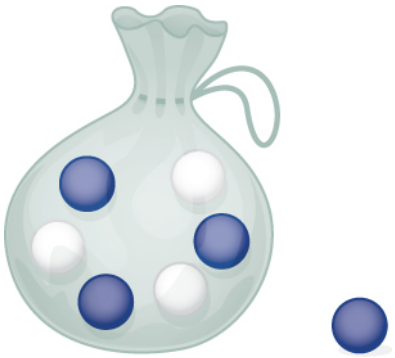


Figure 3.8

If it is not known whether *A* and *B* are independent or dependent, **assume they are dependent until you can show otherwise**.

EXAMPLE 3.4

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. Clubs and spades are black, while diamonds and hearts are red cards. There are 13 cards in each suit consisting of *A* (ace), 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), *Q* (queen), *K* (king) of that suit.

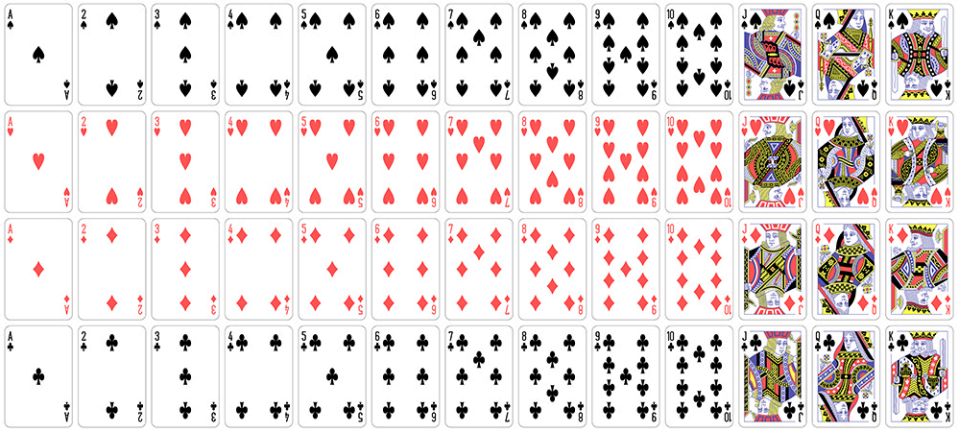


Figure 3.9

a. Sampling with replacement
Suppose you pick three cards with replacement. The first card you pick out of the 52 cards is the *Q* of spades. You put this card back, reshuffle the cards and pick a second card from the 52-card deck. It is the 10 of clubs. You put this card back, reshuffle the cards and pick a third card from the 52-card deck. This time, the card is the *Q* of spades again. Your picks are {*Q* of spades, 10 of clubs, *Q* of spades}. You have picked the *Q* of spades twice. You pick each card from the 52-card deck.

b. Sampling without replacement
Suppose you pick three cards without replacement. The first card you pick out of the 52 cards is the *K* of hearts. You put this card aside and pick the second card from the 51 cards remaining in the deck. It is the three of diamonds. You put this card aside and pick the third card from the remaining 50 cards in the deck. The third card is the *J* of spades. Your picks are {*K* of hearts, three of diamonds, *J* of spades}. Because you have picked the cards without replacement, you cannot pick the same card twice.

TRY IT 3.4

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), *Q* (queen), *K* (king) of that suit. Three cards are picked at random.

a. Suppose you know that the picked cards are *Q* of spades, *K* of hearts and *Q* of spades. Can you decide if the sampling was with or without replacement?
b. Suppose you know that the picked cards are *Q* of spades, *K* of hearts, and *J* of spades. Can you decide if the sampling was with or without replacement?

EXAMPLE 3.5

Problem
You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), *Q* (queen), and *K* (king) of that suit. *S* = spades, *H* = Hearts, *D* = Diamonds, *C* = Clubs.

a. Suppose you pick four cards, but do not put any cards back into the deck. Your cards are *QS*, 1*D*, 1*C*, *QD*.
b. Suppose you pick four cards and put each card back before you pick the next card. Your cards are *KH*, 7*D*, 6*D*, *KH*.

Which of a. or b. did you sample with replacement and which did you sample without replacement?

Solution
a. Because you do not put any cards back, the deck changes after each draw. These events are dependent, and this is

sampling without replacement; b. Because you put each card back before picking the next one, the deck never changes. These events are independent, so this is sampling with replacement.

TRY IT 3.5

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), *Q* (queen), and *K* (king) of that suit. *S* = spades, *H* = Hearts, *D* = Diamonds, *C* = Clubs. Suppose that you sample four cards without replacement. Which of the following outcomes are possible? Answer the same question for sampling with replacement.

a. *QS*, 1*D*, 1*C*, *QD*
b. *KH*, 7*D*, 6*D*, *KH*
c. *QS*, 7*D*, 6*D*, *KS*

Mutually Exclusive Events

A and *B* are **mutually exclusive** events if they cannot occur at the same time. This means that *A* and *B* do not share any outcomes and *P*(*A* AND *B*) = 0.

For example, suppose the sample space *S* = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}. Let *A* = {1, 2, 3, 4, 5}, *B* = {4, 5, 6, 7, 8}, and *C* = {7, 9}. *A* AND *B* = {4, 5}. *P*(*A* AND *B*) = $\frac{2}{10}$ and is not equal to zero. Therefore, *A* and *B* are not mutually exclusive.

A and *C* do not have any numbers in common so *P*(*A* AND *C*) = 0. Therefore, *A* and *C* are mutually exclusive.

If it is not known whether *A* and *B* are mutually exclusive, **assume they are not until you can show otherwise**. The following examples illustrate these definitions and terms.

EXAMPLE 3.6

Flip two fair coins. This is an experiment.

The sample space is {*HH*, *HT*, *TH*, *TT*}, where *T* = tails and *H* = heads. The outcomes are *HH*, *HT*, *TH*, and *TT*. The outcomes *HT* and *TH* are different. The *HT* means that the first coin showed heads and the second coin showed tails. The *TH* means that the first coin showed tails and the second coin showed heads.

- Let *A* = the event of getting **at most one tail**. At most one tail means zero or one tail. Then *A* can be written as {*HH*, *HT*, *TH*}. The outcome *HH* shows zero tails. *HT* and *TH* each show one tail.
- Let *B* = the event of getting all tails. *B* can be written as {*TT*}. *B* is the **complement event** of *A*, so *B* = *A*′. Also, *P*(*A*) + *P*(*B*) = *P*(*A*) + *P*(*A*′) = 1.
- The probabilities for *A* and for *B* are *P*(*A*) = $\frac{3}{4}$ and *P*(*B*) = $\frac{1}{4}$.
- Let *C* = the event of getting all heads. *C* = {*HH*}. Since *B* = {*TT*}, *P*(*B* AND *C*) = 0. *B* and *C* are mutually exclusive. (*B* and *C* have no members in common because you cannot have all tails and all heads at the same time.)
- Let *D* = event of getting **more than one** tail. *D* = {*TT*}. *P*(*D*) = $\frac{1}{4}$.
- Let *E* = event of getting a head on the first roll. This implies you can get either a head or tail on the second roll. *E* = {*HT*, *HH*}. *P*(*E*) = $\frac{2}{4}$.
- Find the probability of getting **at least one** (one or two) tail in two flips. Let *F* = event of getting at least one tail in two flips. *F* = {*HT*, *TH*, *TT*}. *P*(*F*) = $\frac{3}{4}$.

TRY IT 3.6

Draw two cards from a standard 52-card deck with replacement. Find the probability of getting at least one black card.

EXAMPLE 3.7

Problem

Flip two fair coins. Find the probabilities of the events.

a. Let F = the event of getting at most one tail (zero or one tail).
b. Let G = the event of getting two faces that are the same.
c. Let H = the event of getting a head on the first flip followed by a head or tail on the second flip.
d. Are F and G mutually exclusive?
e. Let J = the event of getting all tails. Are J and H mutually exclusive?

Solution

Look at the sample space in Example 3.6.

a. Zero (0) or one (1) tails occur when the outcomes HH, TH, HT show up. $P(F) = \frac{3}{4}$.
b. Two faces are the same if HH or TT show up. $P(G) = \frac{2}{4}$.
c. A head on the first flip followed by a head or tail on the second flip occurs when HH or HT show up. $P(H) = \frac{2}{4}$.
d. F and G share HH so $P(F \text{ AND } G)$ is not equal to zero (0). F and G are not mutually exclusive.
e. Getting all tails occurs when tails shows up on both coins (TT). H's outcomes are HH and HT.

J and H have nothing in common so $P(J \text{ AND } H) = 0$. J and H are mutually exclusive.

TRY IT 3.7

A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Find the probability of the following events:

a. Let F = the event of getting the white ball twice.
b. Let G = the event of getting two balls of different colors.
c. Let H = the event of getting white on the first pick.
d. Are F and G mutually exclusive?
e. Are G and H mutually exclusive?

EXAMPLE 3.8

Roll one fair, six-sided die. The sample space is {1, 2, 3, 4, 5, 6}. Let event A = a face is odd. Then A = {1, 3, 5}. Let event B = a face is even. Then B = {2, 4, 6}.

- Find the complement of A, A'. The complement of A, A', is B because A and B together make up the sample space. $P(A) + P(B) = P(A) + P(A') = 1$. Also, $P(A) = \frac{3}{6}$ and $P(B) = \frac{3}{6}$.
- Let event C = odd faces larger than two. Then C = {3, 5}. Let event D = all even faces smaller than five. Then D = {2, 4}. $P(C \text{ AND } D) = 0$ because you cannot have an odd and even face at the same time. Therefore, C and D are mutually exclusive events.
- Let event E = all faces less than five. E = {1, 2, 3, 4}.

Problem

Are C and E mutually exclusive events? Answer yes or no. Why or why not?

Solution

No. C = {3, 5} and E = {1, 2, 3, 4}. $P(C \text{ AND } E) = \frac{1}{6}$. To be mutually exclusive, $P(C \text{ AND } E)$ must be zero.

- Find $P(C|A)$. This is a conditional probability. Recall that event C is {3, 5} and event A is {1, 3, 5}. To find $P(C|A)$, find the probability of C using the sample space A. You have reduced the sample space from the original sample space {1, 2, 3, 4, 5, 6} to {1, 3, 5}. So, $P(C|A) = \frac{2}{3}$.

TRY IT 3.8

Let event A = learning Spanish. Let event B = learning German. Then A AND B = learning Spanish and German. Suppose $P(A) = 0.4$ and $P(B) = .2$. $P(A \text{ AND } B) = .08$. Are events A and B independent? Hint—You must show **one** of the following:

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(A \text{ AND } B) = P(A)P(B)$

EXAMPLE 3.9

Let event G = taking a math class. Let event H = taking a science class. Then, G AND H = taking a math class and a science class. Suppose $P(G) = .6$, $P(H) = .5$, and $P(G \text{ AND } H) = .3$. Are G and H independent?

If G and H are independent, then you must show **ONE** of the following:

- $P(G|H) = P(G)$
- $P(H|G) = P(H)$
- $P(G \text{ AND } H) = P(G)P(H)$

NOTE

The choice you make depends on the information you have. You could choose any of the methods here because you have the necessary information.

Problem

a. Show that $P(G|H) = P(G)$.

Solution

$P(G|H) = \frac{P(G \text{ AND } H)}{P(H)} = \frac{.3}{.5} = .6 = P(G)$

Problem

b. Show $P(G \text{ AND } H) = P(G)P(H)$.

Solution

$P(G)P(H) = (.6)(.5) = .3 = P(G \text{ AND } H)$

Since G and H are independent, knowing that a person is taking a science class does not change the chance that he or she is taking a math class. If the two events had not been independent, that is, they are dependent, then knowing that a person is taking a science class would change the chance he or she is taking math. For practice, show that $P(H|G) = P(H)$ to show that G and H are independent events.

TRY IT 3.9

In a bag, there are six red marbles and four green marbles. The red marbles are marked with the numbers 1, 2, 3, 4, 5, and 6. The green marbles are marked with the numbers 1, 2, 3, and 4.

- R = a red marble
- G = a green marble
- O = an odd-numbered marble
- The sample space is $S = \{R1, R2, R3, R4, R5, R6, G1, G2, G3, G4\}$.

S has 10 outcomes. What is $P(G \text{ AND } O)$?

EXAMPLE 3.10

Problem

Let event C = taking an English class. Let event D = taking a speech class.

Suppose $P(C) = .75$, $P(D) = .3$, $P(C|D) = .75$ and $P(C \text{ AND } D) = .225$.

Justify your answers to the following questions numerically.

a. Are C and D independent?
b. Are C and D mutually exclusive?
c. What is $P(D|C)$?

Solution

a. Yes, because $P(C|D) = .75 = P(C)$.
b. No, because $P(C \text{ AND } D)$ is not equal to zero.
c. $P(D|C) = \frac{P(C \text{ AND } D)}{P(C)} = \frac{0.225}{.75} = .3$

TRY IT 3.10

A student goes to the library. Let events B = the student checks out a book and D = the student checks out a DVD. Suppose that $P(B) = .40$, $P(D) = .30$ and $P(B \text{ AND } D) = .20$.

a. Find $P(B|D)$.
b. Find $P(D|B)$.
c. Are B and D independent?
d. Are B and D mutually exclusive?

EXAMPLE 3.11

In a box there are three red cards and five blue cards. The red cards are marked with the numbers 1, 2, and 3, and the blue cards are marked with the numbers 1, 2, 3, 4, and 5. The cards are well-shuffled. You reach into the box (you cannot see into it) and draw one card.

Let R = red card is drawn, B = blue card is drawn, E = even-numbered card is drawn.

The sample space $S = R1, R2, R3, B1, B2, B3, B4, B5$. S has eight outcomes.

- $P(R) = \frac{3}{8}$. $P(B) = \frac{5}{8}$. $P(R \text{ AND } B) = 0$. You cannot draw one card that is both red and blue.
- $P(E) = \frac{3}{8}$. There are three even-numbered cards, $R2$, $B2$, and $B4$.
- $P(E|B) = \frac{2}{5}$. There are five blue cards: $B1$, $B2$, $B3$, $B4$, and $B5$. Out of the blue cards, there are two even cards; $B2$ and $B4$.
- $P(B|E) = \frac{2}{3}$. There are three even-numbered cards: $R2$, $B2$, and $B4$. Out of the even-numbered cards, two are blue; $B2$ and $B4$.
- The events R and B are mutually exclusive because $P(R \text{ AND } B) = 0$.
- Let G = card with a number greater than 3. $G = \{B4, B5\}$. $P(G) = \frac{2}{8}$. Let H = blue card numbered between one and four, inclusive. $H = \{B1, B2, B3, B4\}$. $P(G|H) = \frac{1}{4}$. The only card in H that has a number greater than three is $B4$. Since $\frac{2}{8} = \frac{1}{4}$, $P(G) = P(G|H)$, which means that G and H are independent.

TRY IT 3.11

In a basketball arena,

- 70 percent of the fans are rooting for the home team,
- 25 percent of the fans are wearing blue,
- 20 percent of the fans are wearing blue and are rooting for the away team, and
- Of the fans rooting for the away team, 67 percent are wearing blue.

Let A be the event that a fan is rooting for the away team.
Let B be the event that a fan is wearing blue.
Are the events of rooting for the away team and wearing blue independent? Are they mutually exclusive?

EXAMPLE 3.12

In a particular class, 60 percent of the students are female. Fifty percent of all students in the class have long hair. Forty-five percent of the students are female and have long hair. Of the female students, 75 percent have long hair. Let F be the event that a student is female. Let L be the event that a student has long hair. One student is picked randomly. Are the events of being female and having long hair independent?

The following probabilities are given in this example:

- $P(F) = 0.60$; $P(L) = 0.50$
- $P(F \text{ AND } L) = 0.45$
- $P(L \mid F) = 0.75$

NOTE

The choice you make depends on the information you have. You could use the first or last condition on the list for this example. You do not know $P(F \mid L)$ yet, so you cannot use the second condition.

Solution 1

Check whether $P(F \text{ AND } L) = P(F)P(L)$. We are given that $P(F \text{ AND } L) = 0.45$, but $P(F)P(L) = (.60)(.50) = .30$. The events of being female and having long hair are not independent because $P(F \text{ AND } L)$ does not equal $P(F)P(L)$.

Solution 2

Check whether $P(L \mid F)$ equals $P(L)$. We are given that $P(L \mid F) = .75$, but $P(L) = .50$; they are not equal. The events of being female and having long hair are not independent.

Interpretation of Results

The events of being female and having long hair are not independent; knowing that a student is female changes the probability that a student has long hair.

TRY IT 3.12

Mark is deciding which route to take to work. His choices are I = the Interstate and F = Fifth Street.

- $P(I) = .44$ and $P(F) = .55$
- $P(I \text{ AND } F) = 0$ because Mark will take only one route to work.

What is the probability of $P(I \text{ OR } F)$?

EXAMPLE 3.13

Problem

a. Toss one fair coin (the coin has two sides, H and T). The outcomes are _____. Count the outcomes. There are _____ outcomes.

b. Toss one fair, six-sided die (the die has 1, 2, 3, 4, 5, or 6 dots on a side). The outcomes are _____. Count the outcomes. There are _____ outcomes.
c. Multiply the two numbers of outcomes. The answer is _____.
d. If you flip one fair coin and follow it with the toss of one fair, six-sided die, the answer in Part c is the number of outcomes (size of the sample space). List the outcomes. Hint—Two of the outcomes are *H*1 and *T*6.
e. Event *A* = heads (*H*) on the coin followed by an even number (2, 4, 6) on the die. *A* = {_____}. Find *P*(*A*).
f. Event *B* = heads on the coin followed by a three on the die. *B* = {_____}. Find *P*(*B*).
g. Are *A* and *B* mutually exclusive? Hint—What is *P*(*A* AND *B*)? If *P*(*A* AND *B*) = 0, then *A* and *B* are mutually exclusive.
h. Are *A* and *B* independent? Hint—Is *P*(*A* AND *B*) = *P*(*A*)*P*(*B*)? If *P*(*A* AND *B*) = *P*(*A*)*P*(*B*), then *A* and *B* are independent. If not, then they are dependent.

Solution

a. *H* and *T*; 2
b. 1, 2, 3, 4, 5, 6; 6
c. 2(6) = 12
d. Make a systematic list of possible outcomes. Start by listing all possible outcomes when the coin shows tails (*T*). Then list the outcomes that are possible when the coin shows heads (*H*): *T*1, *T*2, *T*3, *T*4, *T*5, *T*6, *H*1, *H*2, *H*3, *H*4, *H*5, *H*6
e. *A* = {*H*2, *H*4, *H*6}; $P(A) = \frac{\text{number of outcomes in } A}{\text{number of possible outcomes}} = \frac{3}{12}$
f. *B* = {*H*3}; $P(B) = \frac{1}{12}$
g. Yes, because *P*(*A* AND *B*) = 0
h. *P*(*A* AND *B*) = 0. $P(A)P(B) = \left(\frac{3}{12}\right)\left(\frac{1}{12}\right)$. *P*(*A* AND *B*) does not equal *P*(*A*)*P*(*B*), so *A* and *B* are dependent.

TRY IT 3.13

A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Let *T* be the event of getting the white ball twice, *F* the event of picking the white ball first, and *S* the event of picking the white ball in the second drawing.

a. Compute *P*(*T*).
b. Compute *P*(*T* | *F*).
c. Are *T* and *F* independent?
d. Are *F* and *S* mutually exclusive?
e. Are *F* and *S* independent?

3.3 Two Basic Rules of Probability

In calculating probability, there are two rules to consider when you are determining if two events are independent or dependent and if they are mutually exclusive or not.

The Multiplication Rule

If *A* and *B* are two events defined on a **sample space**, then *P*(*A* AND *B*) = *P*(*B*)*P*(*A* | *B*).

This equation can be rewritten as *P*(*A* AND *B*) = *P*(*B*)*P*(*A* | *B*), the multiplication rule.

If *A* and *B* are **independent**, then *P*(*A* | *B*) = *P*(*A*). In this special case, *P*(*A* AND *B*) = *P*(*A* | *B*)*P*(*B*) becomes *P*(*A* AND *B*) = *P*(*A*)*P*(*B*).

A bag contains four green marbles, three red marbles, and two yellow marbles. Mark draws two marbles from the bag without replacement. The probability that he draws a yellow marble and then a green marble is

$$P(\text{yellow and green}) = P(\text{yellow}) \cdot P(\text{green} \mid \text{yellow})$$

$$= \frac{2}{9} \cdot \frac{4}{8}$$

$$= \frac{1}{9}$$

Notice that $P(\text{green} \mid \text{yellow}) = \frac{4}{8}$. After the yellow marble is drawn, there are four green marbles in the bag and eight marbles in all.

The Addition Rule

If *A* and *B* are defined on a sample space, then *P*(*A* OR *B*) = *P*(*A*) + *P*(*B*) − *P*(*A* AND *B*).

Draw one card from a standard deck of playing cards. Let *H* = the card is a heart, and let *J* = the card is a jack. These events are not mutually exclusive because a card can be both a heart and a jack.

$$P(H \text{ or } J) = P(H) + P(J) - P(H \text{ and } J)$$

$$\begin{aligned} &= \tfrac{13}{52} + \tfrac{4}{52} - \tfrac{1}{52} \\ &= \tfrac{16}{52} \\ &= \tfrac{4}{13} \\ &\approx .3077 \end{aligned} \qquad 3.3$$

If *A* and *B* are **mutually exclusive**, then *P*(*A* AND *B*) = 0. Then *P*(*A* OR *B*) = *P*(*A*) + *P*(*B*) − *P*(*A* AND *B*) becomes *P*(*A* OR *B*) = *P*(*A*) + *P*(*B*).

Draw one card from a standard deck of playing cards. Let *H* = the card is a heart and *S* = the card is a spade. These events are mutually exclusive because a card cannot be a heart and a spade at the same time. The probability that the card is a heart or a spade is

$$P(H \text{ or } S) = P(H) + P(S)$$

$$\begin{aligned} &= \tfrac{13}{52} + \tfrac{13}{52} \\ &= \tfrac{26}{52} \\ &= \tfrac{1}{2} \\ &= .5 \end{aligned} \qquad 3.4$$

EXAMPLE 3.14

Klaus is trying to choose where to go on vacation. His two choices are: *A* = New Zealand and *B* = Alaska.

- Klaus can only afford one vacation. The probability that he chooses *A* is *P*(*A*) = .6 and the probability that he chooses *B* is *P*(*B*) = .35.
- *P*(*A* AND *B*) = 0 because Klaus can only afford to take one vacation.
- Therefore, the probability that he chooses either New Zealand or Alaska is *P*(*A* OR *B*) = *P*(*A*) + *P*(*B*) = .6 + .35 = .95. Note that the probability that he does not choose to go anywhere on vacation must be .05.

EXAMPLE 3.15

Carlos plays college soccer. He makes a goal 65 percent of the time he shoots. Carlos is going to attempt two goals in a row in the next game. *A* = the event Carlos is successful on his first attempt. *P*(*A*) = .65. *B* = the event Carlos is successful on his second attempt. *P*(*B*) = .65. Carlos tends to shoot in streaks. The probability that he makes the second goal **given** that he made the first goal is .90.

Problem

a. What is the probability that he makes both goals?

Solution

a. The problem is asking you to find *P*(*A* AND *B*) = *P*(*B* AND *A*). Since *P*(*B*|*A*) = .90: *P*(*B* AND *A*) = *P*(*B*|*A*) *P*(*A*) = (.90)(.65) = .585.

Carlos makes the first and second goals with probability .585.

Problem

b. What is the probability that Carlos makes either the first goal or the second goal?

Solution

b. The problem is asking you to find $P(A \text{ OR } B)$.

$P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B) = .65 + .65 - .585 = .715$

Carlos makes either the first goal or the second goal with probability .715.

Problem

c. Are A and B independent?

Solution

c. No, they are not, because $P(B \text{ AND } A) = .585$.

$P(B)P(A) = (.65)(.65) = .423$

$.423 \neq .585 = P(B \text{ AND } A)$

So, $P(B \text{ AND } A)$ is **not** equal to $P(B)P(A)$.

Problem

d. Are A and B mutually exclusive?

Solution

d. No, they are not because $P(A \text{ and } B) = .585$.

To be mutually exclusive, $P(A \text{ AND } B)$ must equal zero.

TRY IT 3.15

Helen plays basketball. For free throws, she makes the shot 75 percent of the time. Helen must now attempt two free throws. C = the event that Helen makes the first shot.
$P(C) = .75$. D = the event Helen makes the second shot. $P(D) = .75$. The probability that Helen makes the second free throw given that she made the first is .85. What is the probability that Helen makes both free throws?

EXAMPLE 3.16

A community swim team has **150** members. **Seventy-five** of the members are advanced swimmers. **Forty-seven** of the members are intermediate swimmers. The remainder are novice swimmers. **Forty** of the advanced swimmers practice four times a week. **Thirty** of the intermediate swimmers practice four times a week. **Ten** of the novice swimmers practice four times a week. Suppose one member of the swim team is chosen randomly.

Problem

a. What is the probability that the member is a novice swimmer?

Solution

a. There are **150** members; 75 of these are advanced, and 47 of these are intermediate swimmers. So there are 150 − 75 − 47 = **28** novice swimmers. The probability that a randomly selected swimmer is a novice is $\frac{28}{150}$.

Problem

b. What is the probability that the member practices four times a week?

Solution

b. $\frac{40 + 30 + 10}{150} = \frac{80}{150}$

Problem

c. What is the probability that the member is an advanced swimmer and practices four times a week?

Solution

c. There are 40 advanced swimmers who practice four times per week, so the probability is $\frac{40}{150}$.

Problem

d. What is the probability that a member is an advanced swimmer and an intermediate swimmer? Are being an advanced swimmer and being an intermediate swimmer mutually exclusive? Why or why not?

Solution

d. P(advanced AND intermediate) = 0, so these are mutually exclusive events. A swimmer cannot be an advanced swimmer and an intermediate swimmer at the same time.

Problem

e. Are being a novice swimmer and practicing four times a week independent events? Why or why not?

Solution

e. No, these are not independent events.
P(novice AND practices four times per week) = .0667
P(novice)P(practices four times per week) = .0996
.0667 ≠ .0996

TRY IT 3.16

A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors going to college are on their school's sports teams. Thirty of the seniors going directly to work are on their school's sports teams. Five of the seniors taking a gap year are on their schools sports teams. What is the probability that a senior is taking a gap year?

EXAMPLE 3.17

Felicity attends a school in Modesto, CA. The probability that Felicity enrolls in a math class is .2 and the probability that she enrolls in a speech class is .65. The probability that she enrolls in a math class GIVEN that she enrolls in speech class is .25.

Let M = math class, S = speech class, and $M|S$ = math given speech.

Problem

a. What is the probability that Felicity enrolls in math and speech?
Find $P(M \text{ AND } S) = P(M|S)P(S)$.
b. What is the probability that Felicity enrolls in math or speech classes?
Find $P(M \text{ OR } S) = P(M) + P(S) - P(M \text{ AND } S)$.
c. Are M and S independent? Is $P(M|S) = P(M)$?
d. Are M and S mutually exclusive? Is $P(M \text{ AND } S) = 0$?

Solution

a. $P(M \text{ AND } S) = P(M|S)P(S) = .25(.65) = .1625$

b. $P(M \text{ OR } S) = P(M) + P(S) - P(M \text{ AND } S) = .2 + .65 - .1625 = .6875$

c. No, $P(M|S) = .25$ and $P(M) = .2$.

d. No, $P(M \text{ AND } S) = .1625$.

TRY IT 3.17

A student goes to the library. Let events B = the student checks out a book and D = the student checks out a DVD. Suppose that $P(B) = .40$, $P(D) = .30$, and $P(D|B) = .5$.

a. Find $P(B \text{ AND } D)$.
b. Find $P(B \text{ OR } D)$.

EXAMPLE 3.18

Researchers are studying one particular type of disease that affects women more often than men. Studies show that about one woman in seven (approximately 14.3 percent) who live to be 90 will develop the disease. Suppose that of those women who develop this disease, a test is negative 2 percent of the time. Also suppose that in the general population of women, the test for the disease is negative about 85 percent of the time. Let B = woman develops the disease and let N = tests negative. Suppose one woman is selected at random.

Problem

a. What is the probability that the woman develops the disease? What is the probability that woman tests negative?

Solution

a. $P(B) = .143$; $P(N) = .85$

Problem

b. Given that the woman develops the disease, what is the probability that she tests negative?

Solution

b. Among women who develop the disease, the test is negative 2 percent of the time, so $P(N|B) = .02$

Problem

c. What is the probability that the woman has the disease AND tests negative?

Solution

c. $P(B \text{ AND } N) = P(B)P(N|B) = (.143)(.02) = .0029$

Problem

d. What is the probability that the woman has the disease OR tests negative?

Solution

d. $P(B \text{ OR } N) = P(B) + P(N) - P(B \text{ AND } N) = .143 + .85 - .0029 = .9901$

Problem

e. Are having the disease and testing negative independent events?

Solution

e. No. $P(N) = .85$; $P(N|B) = .02$. So, $P(N|B)$ does not equal $P(N)$.

Problem

f. Are having the disease and testing negative mutually exclusive?

Solution

f. No. $P(B \text{ AND } N) = .0029$. For B and N to be mutually exclusive, $P(B \text{ AND } N)$ must be zero.

TRY IT 3.18

A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors going to college are on their school's sports teams. Thirty of the seniors going directly to work are on their school's sports teams. Five of the seniors taking a gap year are on their school's sports teams. What is the probability that a senior is going to college and plays sports?

EXAMPLE 3.19

Problem

Refer to the information in Example 3.18. P = tests positive.

a. Given that a woman develops the disease, what is the probability that she tests positive? Find $P(P|B) = 1 - P(N|B)$.
b. What is the probability that a woman develops the disease and tests positive? Find $P(B \text{ AND } P) = P(P|B)P(B)$.
c. What is the probability that a woman does not develop the disease? Find $P(B') = 1 - P(B)$.
d. What is the probability that a woman tests positive for the disease? Find $P(P) = 1 - P(N)$.

Solution

a. $P(P|B) = 1 - P(N|B) = 1 - .02 = .98$

b. $P(B \text{ AND } P) = P(P|B)P(B) = .98(.143) = .1401$

c. $P(B') = 1 - P(B) = 1 - .143 = .857$

d. $P(P) = 1 - P(N) = 1 - .85 = .15$

TRY IT 3.19

A student goes to the library. Let events B = the student checks out a book and D = the student checks out a DVD. Suppose that $P(B) = .40$, $P(D) = .30$, and $P(D|B) = .5$.

a. Find $P(B')$.
b. Find $P(D \text{ AND } B)$.
c. Find $P(B|D)$.
d. Find $P(D \text{ AND } B')$.
e. Find $P(D|B')$.

3.4 Contingency Tables

A **two-way table** provides a way of portraying data that can facilitate calculating probabilities. When used to calculate probabilities, a two-way table is often called a **contingency table**. The table helps in determining conditional probabilities quite easily. The table displays sample values in relation to two different variables that may be dependent or contingent on one another. We used two-way tables in Chapters 1 and 2 to calculate marginal and conditional distributions. These tables organize data in a way that supports the calculation of relative frequency and, therefore, experimental (empirical) probability. Later on, we will use contingency tables again, but in another manner.

EXAMPLE 3.20

Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

	Speeding Violation in the Last Year	No Speeding Violation in the Last Year	Total
Uses a cell phone while driving	25	280	305
Does not use a cell phone while driving	45	405	450
Total	70	685	755

Table 3.3

The total number of people in the sample is 755. The row totals are 305 and 450. The column totals are 70 and 685. Notice that 305 + 450 = 755 and 70 + 685 = 755.

Using the table, calculate the following probabilities:

Problem

a. Find *P*(Person uses a cell phone while driving).
b. Find *P*(Person had no violation in the last year).
c. Find *P*(Person had no violation in the last year *and* uses a cell phone while driving).
d. Find *P*(Person uses a cell phone while driving *or* person had no violation in the last year).
e. Find *P*(Person uses a cell phone while driving *given* person had a violation in the last year).
f. Find *P*(Person had no violation last year *given* person does not use a cell phone while driving).

Solution

a. This is the same as the marginal distribution (Section 1.2).

$$P(\text{Person uses a cell phone while driving}) = \frac{\text{number who use cell phones while driving}}{\text{number in study}} = \frac{305}{755} \approx .4040 \quad \textbf{3.5}$$

b. The marginal distribution is

$$P(\text{Person had no violation in the last year}) = \frac{\text{number who had no violation}}{\text{number in study}} = \frac{685}{755} \approx .9073. \quad \textbf{3.6}$$

c. Find the number of participants who satisfy *both* conditions.

$$P(\text{Person had no violation in the last year AND uses a cell phone while driving}) = \frac{\text{number who had no violation AND uses cell phone while driving}}{\text{number in study}}$$
$$= \frac{280}{755}$$
$$\approx .3709 \quad \textbf{3.7}$$

d. To find this probability, you need to identify how many participants use a cell phone while driving OR have no violation in the past year OR both.

$$P(\text{Person uses a cell phone while driving OR had no violation in the last year}) = \frac{25+405+280}{755}$$
$$= \frac{710}{755}$$
$$\approx .9404 \quad \textbf{3.8}$$

e. This is a conditional probability. You are *given* that the person had no violation in the last year, so you need only consider the values in that column of data.

$$(\text{Person uses a cell phone while driving GIVEN the person had a violation in the last year}) = \frac{\text{number who used cell phone AND had a violation}}{\text{number in study who had a violation in the last year}}$$
$$= \frac{25}{70}$$
$$\approx .3571 \quad \textbf{3.9}$$

f. For this conditional probability, consider only values in the row labeled "Does not use a cell phone while driving."

$$P(\text{Person had no violation last year GIVEN person does not use cell phone while driving}) = \frac{405}{450} = .9$$

TRY IT 3.20

Table 3.4 shows the number of athletes who stretch before exercising and how many had injuries within the past year.

	Injury in Past Year	No Injury in Past Year	Total
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

Table 3.4

a. What is *P*(Athlete stretches before exercising)?
b. What is *P*(Athlete stretches before exercising | no injury in the last year)?

EXAMPLE 3.21

Table 3.5 shows a random sample of 100 hikers and the areas of hiking they prefer.

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	___	45
Male	___	___	14	55
Total	___	41	___	___

Table 3.5 Hiking Area Preference

Problem

a. Complete the table.

Solution

a. There are 45 females in the sample; 18 prefer the coastline and 16 prefer hiking near lakes and streams. So, we know there are 45 − 18 − 16 = 11 female students who prefer hiking on mountain peaks.

Continue reasoning in this way to complete the table.

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	**11**	45
Male	**16**	**25**	14	55
Total	**34**	41	**25**	**100**

Table 3.6 Hiking Area Preference

Problem

b. Are the events *being female* and *preferring the coastline* independent events?

Let F = being female and let C = preferring the coastline.

1. Find $P(F \text{ AND } C)$.
2. Find $P(F)P(C)$.

Are these two numbers the same? If they are, then F and C are independent. If they are not, then F and C are not independent.

Solution

b.

1. $P(F \text{ AND } C) = \frac{18}{100} = .18$
2. $P(F)P(C) = \left(\frac{45}{100}\right)\left(\frac{34}{100}\right) = (.45)(.34) = .153$

$P(F \text{ AND } C) \neq P(F)P(C)$, so the events F and C are not independent.

Problem

c. Find the probability that a person is male given that the person prefers hiking near lakes and streams. Let M = being male, and let L = prefers hiking near lakes and streams.

1. What word tells you this is a conditional?
2. Is the sample space for this problem all 100 hikers? If not, what is it?
3. Fill in the blanks and calculate the probability: $P(_____|_____) = _____$.

Solution

c.

1. The word *given* tells you that this is a conditional.
2. No, the sample space for this problem is the 41 hikers who prefer lakes and streams.
3. Find the conditional probability $P(M|L)$. Because it is given that the person prefers hiking near lakes and streams, you need only consider the values in the column labeled "Near Lakes and Streams." $P(M|L) = \frac{25}{41}$

Problem

d. Find the probability that a person is female or prefers hiking on mountain peaks. Let F = being female, and let P = prefers mountain peaks.

1. Find $P(F)$.
2. Find $P(P)$.
3. Find $P(F \text{ AND } P)$.
4. Find $P(F \text{ OR } P)$.

Solution

d.

1. $P(F) = \frac{45}{100}$

2. $P(P) = \frac{25}{100}$
3. $P(F \text{ AND } P) = \frac{\text{number of hikers that are both female AND prefers mountain peaks}}{\text{number of hikers in study}} = \frac{11}{100}$
4. $P(F \text{ OR } P) = P(F) + P(P) - P(F \text{ AND } P) = \frac{45}{100} + \frac{25}{100} - \frac{11}{100} = \frac{59}{100}$

TRY IT 3.21

Table 3.7 shows a random sample of 200 cyclists and the routes they prefer. Let M = males and H = hilly path.

Gender	Lake Path	Hilly Path	Wooded Path	Total
Female	45	38	27	110
Male	26	52	12	90
Total	71	90	39	200

Table 3.7

a. Out of the males, what is the probability that the cyclist prefers a hilly path?
b. Are the events *being male* and *preferring the hilly path* independent events?

EXAMPLE 3.22

Muddy Mouse lives in a cage with three doors. If Muddy goes out the first door, the probability that he gets caught by Alissa the cat is $\frac{1}{5}$ and the probability he is not caught is $\frac{4}{5}$. If he goes out the second door, the probability he gets caught by Alissa is $\frac{1}{4}$ and the probability he is not caught is $\frac{3}{4}$. The probability that Alissa catches Muddy coming out of the third door is $\frac{1}{2}$ and the probability she does not catch Muddy is $\frac{1}{2}$. It is equally likely that Muddy will choose any of the three doors, so the probability of choosing each door is $\frac{1}{3}$.

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	___
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	___
Total	___	___	___	1

Table 3.8 Door Choice

- The first entry $\frac{1}{15} = \left(\frac{1}{5}\right)\left(\frac{1}{3}\right)$ is P(Door One AND Caught).
- The entry $\frac{4}{15} = \left(\frac{4}{5}\right)\left(\frac{1}{3}\right)$ is P(Door One AND Not Caught).

Verify the remaining entries.

Problem

a. Complete the probability contingency table. Calculate the entries for the totals. Verify that the lower-right corner entry is 1.

Solution

a.

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{19}{60}$
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	$\frac{41}{60}$
Total	$\frac{5}{15}$	$\frac{4}{12}$	$\frac{2}{6}$	1

Table 3.9 Door Choice

Problem

b. What is the probability that Alissa does not catch Muddy?

Solution

b. $\frac{41}{60}$

Problem

c. What is the probability that Muddy chooses Door One OR Door Two given that Muddy is caught by Alissa?

Solution

c. This is a conditional probability, so consider only probabilities in the row labeled "Caught." Choosing Door One and choosing Door Two are mutually exclusive, so

$$P(\text{Choosing Door One OR Choosing Door Two AND Caught}) = \frac{1}{15} + \frac{1}{12} = \frac{9}{60}.$$

Use the formula for conditional probability $P(A|B) = \frac{P(A \text{AND} B)}{P(B)}$.

$$P(\text{Door One OR Door Two}|\text{Caught}) = \frac{P(\text{Door One OR Door Two AND Caught})}{P(\text{Caught})} = \frac{\frac{9}{60}}{\frac{19}{60}} = \frac{9}{19.}$$

EXAMPLE 3.23

Table 3.10 contains the number of crimes per 100,000 inhabitants from 2008 to 2011 in the United States.

Year	Crime A	Crime B	Crime C	Crime D	Total
2008	145.7	732.1	29.7	314.7	
2009	133.1	717.7	29.1	259.2	
2010	119.3	701	27.7	239.1	
2011	113.7	702.2	26.8	229.6	
Total					

Table 3.10 U.S. Crime Index Rates Per 100,000 Inhabitants 2008–2011

Problem

TOTAL each column and each row. Total data = 4,520.7.

a. Find P(2009 AND Crime A).
b. Find P(2010 AND Crime B).
c. Find P(2010 OR Crime B).
d. Find P(2011 | Crime A).
e. Find P(Crime D | 2008).

Solution

a. $\frac{133.1}{4,520.7} = .0294$, b. $\frac{701}{4,520.7} = .1551$, c. P(2010 OR Crime B) = P(2010) + P(Crime B) – P(2010 AND Crime B) = $\frac{1,087.1}{4,520.7}$ + $\frac{2,852.9}{4,520.7} - \frac{701}{4,520.7} = .7165$, d. $\frac{113.7}{511.8} = .2222$, e. $\frac{314.7}{1,222.2} = .2575$

TRY IT 3.23

Table 3.11 relates the weights and heights of a group of individuals participating in an observational study.

Ages	Tall	Medium	Short	Totals
Under 18	18	28	14	
18–50	20	51	28	
51+	12	25	9	
Totals				

Table 3.11

a. Find the total for each row and column.
b. Find the probability that a randomly chosen individual from this group is tall.
c. Find the probability that a randomly chosen individual from this group is Under 18 and tall.
d. Find the probability that a randomly chosen individual from this group is tall given that the individual is Under 18.
e. Find the probability that a randomly chosen individual from this group is Under 18 given that the individual is tall.
f. Find the probability a randomly chosen individual from this group is tall and age 51+.
g. Are the events under 18 and tall independent?

3.5 Tree and Venn Diagrams

Sometimes, when the probability problems are complex, it can be helpful to graph the situation. Tree diagrams and Venn diagrams are two tools that can be used to visualize and solve conditional probabilities.

Tree Diagrams

A **tree diagram** is a special type of graph used to determine the outcomes of an experiment. It consists of *branches* that are labeled with either frequencies or probabilities. Tree diagrams can make some probability problems easier to visualize and solve. The following example illustrates how to use a tree diagram:

EXAMPLE 3.24

In an urn, there are 11 balls. Three balls are red (R) and eight balls are blue (B). Draw two balls, one at a time, **with replacement**. *With replacement* means that you put the first ball back in the urn before you select the second ball. Therefore, you are selecting from exactly the same group each time, so each draw is independent. The tree diagram shows all the possible outcomes.

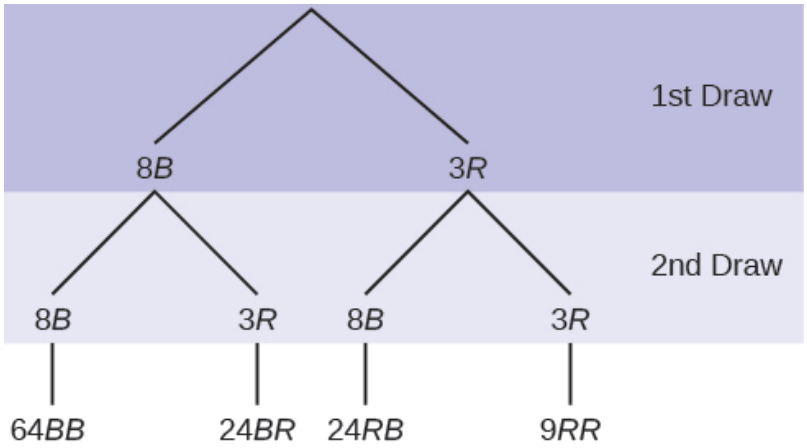


Figure 3.10 Total = 64 + 24 + 24 + 9 = 121.

The first set of branches represents the first draw. There are 8 ways to draw a blue marble and 3 ways to draw a red one. The second set of branches represents the second draw. Regardless of the choice on the first draw, there are again eight ways to draw a blue marble and 3 ways to draw a red one. Read down each branch to see the total number of possible outcomes. For example, there are 8 ways to get a blue marble on the first draw, and eight ways to get one on the second draw, so there are 8 × 8 = 64 different ways to draw two blue marbles in succession. Each of the outcomes is distinct. In fact, we can list each red ball as *R*1, *R*2, and *R*3 and each blue ball as *B*1, *B*2, *B*3, *B*4, *B*5, *B*6, *B*7, and *B*8. Then the nine *RR* outcomes can be written as follows:

*R*1*R*1, *R*1*R*2, *R*1*R*3, *R*2*R*1, *R*2*R*2, *R*2*R*3, *R*3*R*1, *R*3*R*2, *R*3*R*3.

The other outcomes are similar.

There are a total of 11 balls in the urn. Draw two balls, one at a time, with replacement. There are 11(11) = 121 outcomes, the size of the **sample space**.

Problem

a. List the 24 *BR* outcomes: *B*1*R*1, *B*1*R*2, *B*1*R*3, . . .

Solution

a. We know that there will be 24 different possible outcomes because there are eight ways to draw blue and three ways to draw red. Make a systematic list of possible outcomes that consist of a blue marble on the first draw and a red marble on the second draw.

*B*1*R*1, *B*1*R*2, *B*1*R*3
*B*2*R*1, *B*2*R*2, *B*2*R*3
*B*3*R*1, *B*3*R*2, *B*3*R*3
*B*4*R*1, *B*4*R*2, *B*4*R*3
*B*5*R*1, *B*5*R*2, *B*5*R*3
*B*6*R*1, *B*6*R*2, *B*6*R*3
*B*7*R*1, *B*7*R*2, *B*7*R*3
*B*8*R*1, *B*8*R*2, *B*8*R*3

Problem

b. Calculate *P*(*RR*).

Solution

b. You can use the tree diagram. There are nine ways to draw two reds and 121 possible outcomes. So, $P(RR) = \frac{9}{121.}$.

Each draw is independent, so you can also use the formula: $P(RR) = P(R)P(R) = \left(\frac{3}{11}\right)\left(\frac{3}{11}\right) = \frac{9}{121.}$

Problem

c. Calculate *P*(*RB* OR *BR*).

Solution

c. The tree diagram shows that there are 24 ways to draw *RB* and 24 ways to draw *BR*. There are 121 possible outcomes, so $P(RB \text{ or } BR) = \frac{24+24}{121} = \frac{48}{121}$.

The events *RB* and *BR* are mutually exclusive, so $P(RB \text{ OR } BR) = P(RB) + P(BR) = P(R)P(B) + P(B)P(R) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{11}\right) = \frac{48}{121.}$

Problem

d. Using the tree diagram, calculate $P(R$ on 1st draw AND B on 2nd draw).

Solution

d. Follow the path on the tree. There are three ways to get a red marble on the first draw and eight ways to get a blue on the second draw. There are 3 × 8 = 24 ways to draw red then blue, so $P(RB) = \frac{24}{121}$.

Can you think of another way to find this probability? $P(R$ on 1st draw AND B on 2nd draw$) = P(RB) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) = \frac{24}{121}$

Problem

e. Using the tree diagram, calculate $P(R$ on 2nd draw GIVEN B on 1st draw).

Solution

e. Given that a blue marble is selected first, we need only follow the left set of branches on the tree diagram. In this case, there are three ways to obtain red on the second draw and 11 possible outcomes.

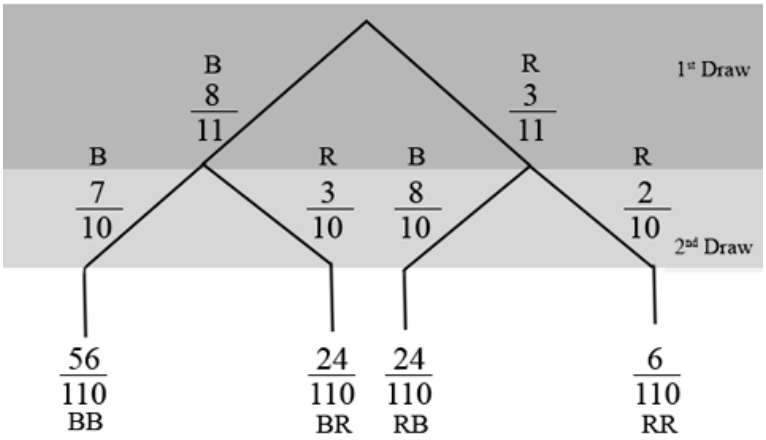


Figure 3.11

$$P(R \text{ on 2nd draw GIVEN } B \text{ on 1st}) = P(R \text{ on 2nd} \mid B \text{ on 1st}) = \frac{3}{11}$$

You can also use the formula

$$P(R \text{ on 2nd} \mid B \text{ on 1st}) = \frac{P(R \text{ on 2nd AND } B \text{ on 1st})}{P(B \text{ on 1st})} = \frac{\frac{24}{121}}{\frac{64+24}{121}} = \frac{24}{88} = \frac{3}{11.}$$

Problem

f. Using the tree diagram, calculate $P(BB)$.

Solution

f. $P(BB) = \frac{64}{121}$

Problem

g. Using the tree diagram, calculate $P(B$ on the 2nd draw GIVEN R on the first draw).

Solution

g. $P(B \text{ on 2nd draw} \mid R \text{ on 1st draw}) = \frac{8}{11}$

There are 9 + 24 outcomes that have R on the first draw (9 RR and 24 RB). The sample space is then 9 + 24 = 33. Twenty-four of the 33 outcomes have B on the second draw. The probability is then $\frac{24}{33}$.

TRY IT 3.24

In a standard deck, there are 52 cards. Twelve cards are face cards (event F) and 40 cards are not face cards (event N). Draw two cards, one at a time, with replacement. All possible outcomes are shown in the tree diagram as frequencies. Using the tree diagram, calculate $P(FF)$.

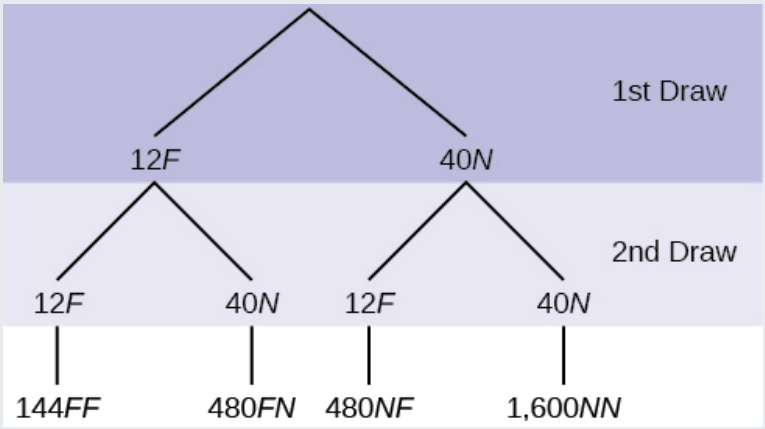


Figure 3.12

EXAMPLE 3.25

An urn has three red marbles and eight blue marbles in it. Draw two marbles, one at a time, this time without replacement, from the urn. ***Without replacement*** means that you do not put the first ball back before you select the second marble. Following is a tree diagram for this situation. The branches are labeled with probabilities instead of frequencies. The numbers at the ends of the branches are calculated by multiplying the numbers on the two corresponding branches, for example, $P(RR) = \left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$.

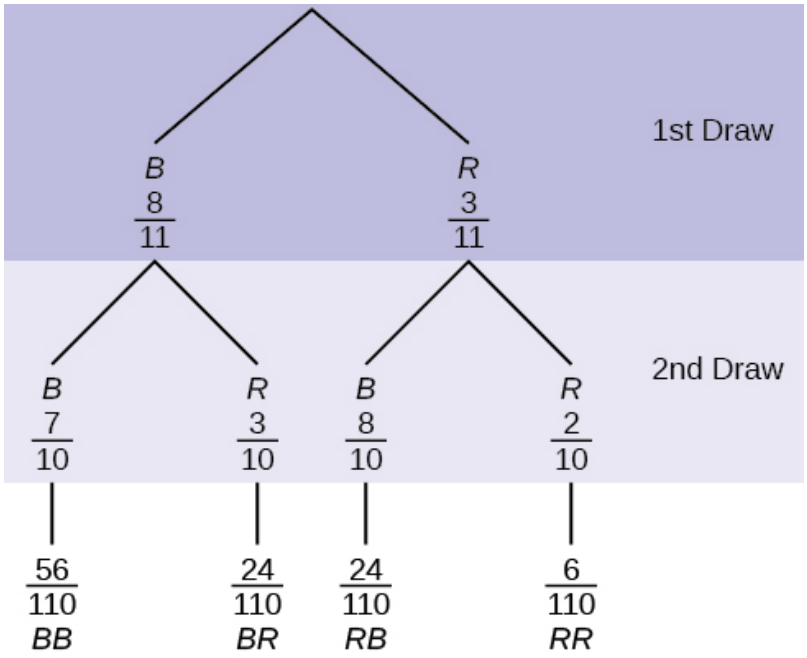


Figure 3.13 Total = $\frac{56+24+24+6}{110} = \frac{110}{110} = 1$.

NOTE

If you draw a red on the first draw from the three red possibilities, there are two red marbles left to draw on the second draw. You do not put back or replace the first marble after you have drawn it. You draw **without replacement**, so that on the second draw there are 10 marbles left in the urn.

Calculate the following probabilities using the tree diagram:

Problem

a. $P(RR)$ = _____

Solution

a. $P(RR) = \left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$

Problem

b. Fill in the blanks.

$P(RB \text{ OR } BR) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) + (_____)(_____) = \frac{48}{110}$

Solution

b. $P(RB \text{ OR } BR) = P(RB) + P(BR) = P(R \text{ on 1st})\ P(B \text{ on 2nd}) + P(B \text{ on 1st})\ P(R \text{ on 2nd}) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{10}\right) = \frac{48}{110}$

Problem

c. Because this is a conditional probability, we restrict the sample space to consider only those outcomes that have a blue marble in the first draw. Look at the second level of the tree to see that $P(R \text{ on 2nd} \mid B \text{ on 1st}) = \frac{3}{10}$

Solution

c. $P(R \text{ on 2nd} \mid B \text{ on 1st}) = \frac{3}{10}$

Problem

d. Fill in the blanks.

$P(R \text{ on 1st AND } B \text{ on 2nd}) = P(RB) = (_____)(_____) = \frac{24}{110}$

Solution

d. $P(R \text{ on 1st AND } B \text{ on 2nd}) = P(RB) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) = \frac{24}{110}$

Problem

e. Find $P(BB)$.

Solution

e. $P(BB) = \left(\frac{8}{11}\right)\left(\frac{7}{10}\right)$

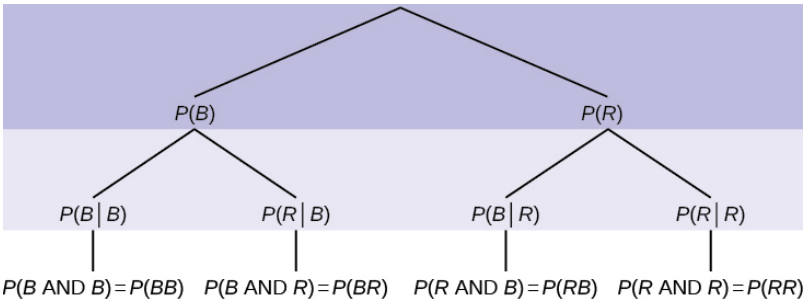
Problem

f. Find $P(B \text{ on 2nd} \mid R \text{ on 1st})$.

Solution

f. Using the tree diagram, $P(B \text{ on 2nd} \mid R \text{ on 1st}) = P(R \mid B) = \frac{8}{10}$.

If we are using probabilities, we can label the tree in the following general way:



- $P(R \mid R)$ here means $P(R \text{ on 2nd} \mid R \text{ on 1st})$
- $P(B \mid R)$ here means $P(B \text{ on 2nd} \mid R \text{ on 1st})$
- $P(R \mid B)$ here means $P(R \text{ on 2nd} \mid B \text{ on 1st})$
- $P(B \mid B)$ here means $P(B \text{ on 2nd} \mid B \text{ on 1st})$

TRY IT 3.25

In a standard deck, there are 52 cards. Twelve cards are face cards (F) and 40 cards are not face cards (N). Draw two cards, one at a time, without replacement. The tree diagram is labeled with all possible probabilities.

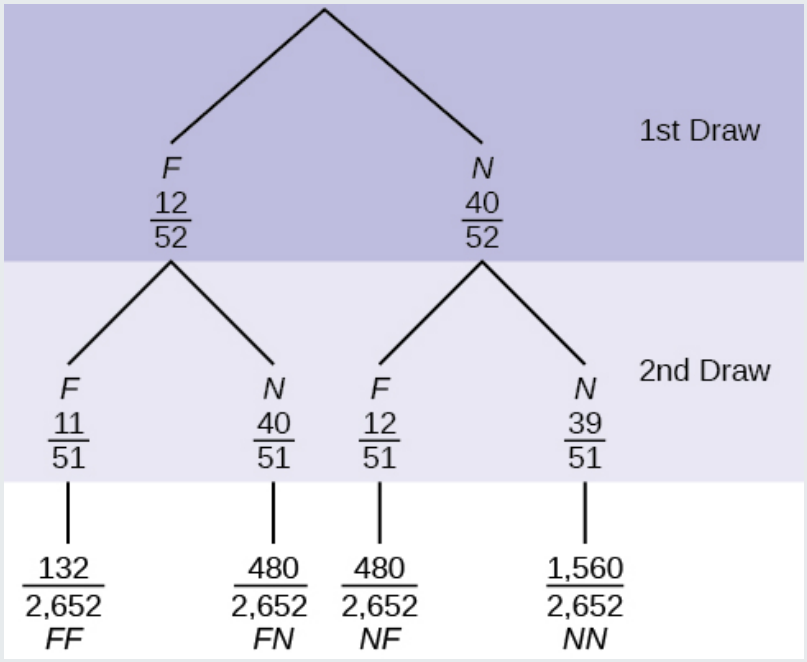
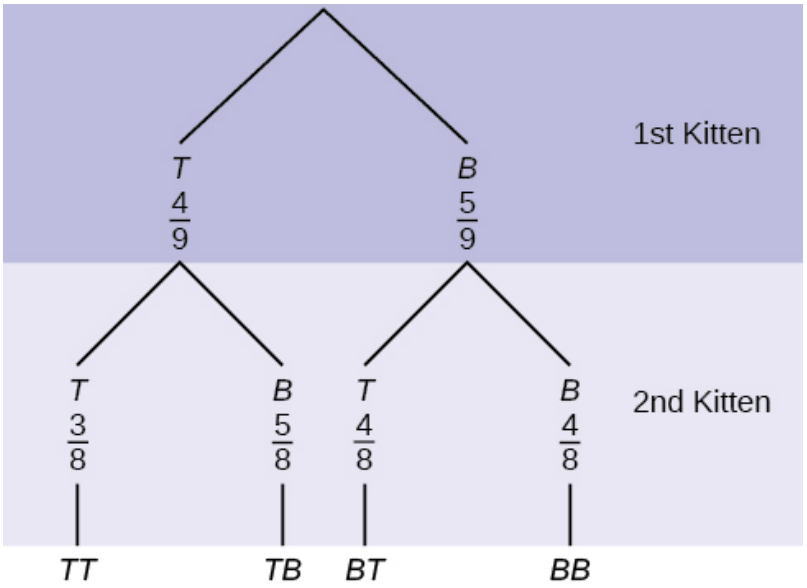


Figure 3.14

a. Find $P(FN \text{ OR } NF)$.
b. Find $P(N \mid F)$.
c. Find P(at most one face card).
   Hint: *At most one face card* means zero or one face card.
d. Find P(at least one face card).
   Hint: *At least one face card* means one or two face cards.

EXAMPLE 3.26

A litter of kittens available for adoption at the Humane Society has four tabby kittens and five black kittens. A family comes in and randomly selects two kittens (without replacement) for adoption.



Problem

a. Which shows the probability that both kittens are tabby?

$$\text{a.}\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\ \text{b.}\left(\frac{4}{9}\right)\left(\frac{4}{9}\right)\ \text{c.}\left(\frac{4}{9}\right)\left(\frac{3}{8}\right)\ \text{d.}\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$$

b. What is the probability that one kitten of each coloring is selected?

$$\text{a.}\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)\ \text{b.}\left(\frac{4}{9}\right)\left(\frac{5}{8}\right)\ \text{c.}\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)+\left(\frac{5}{9}\right)\left(\frac{4}{9}\right)\ \text{d.}\left(\frac{4}{9}\right)\left(\frac{5}{8}\right)+\left(\frac{5}{9}\right)\left(\frac{4}{8}\right)$$

c. What is the probability that a tabby is chosen as the second kitten when a black kitten was chosen as the first?

d. What is the probability of choosing two kittens of the same color?

Solution

a. $\left(\frac{4}{9}\right)\left(\frac{3}{8}\right)$, b. $\left(\frac{4}{9}\right)\left(\frac{5}{8}\right)+\left(\frac{5}{9}\right)\left(\frac{4}{8}\right)$, c. $\frac{4}{8}$, d. $\frac{32}{72}$

TRY IT 3.26

Suppose there are four red balls and three yellow balls in a box. Three balls are drawn from the box without replacement. What is the probability that one ball of each coloring is selected?

Venn Diagram

A **Venn diagram** is a picture that represents the outcomes of an experiment. It generally consists of a box that represents the sample space *S* together with circles or ovals. The circles or ovals represent events.

EXAMPLE 3.27

Suppose an experiment has the outcomes 1, 2, 3, . . . , 12 where each outcome has an equal chance of occurring. Let event *A* = {1, 2, 3, 4, 5, 6} and event *B* = {6, 7, 8, 9}. Then *A* AND *B* = {6} and *A* OR *B* = {1, 2, 3, 4, 5, 6, 7, 8, 9}. The Venn diagram is as follows:

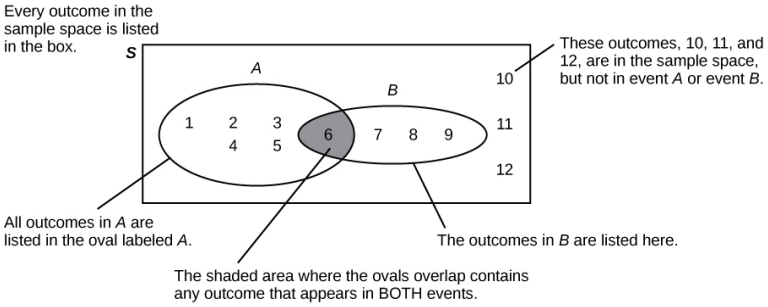


Figure 3.15

TRY IT 3.27

Suppose an experiment has outcomes black, white, red, orange, yellow, green, blue, and purple, where each outcome has an equal chance of occurring. Let event *C* = {green, blue, purple} and event *P* = {red, yellow, blue}. Then *C* AND *P* = {blue} and *C* OR *P* = {green, blue, purple, red, yellow}. Draw a Venn diagram representing this situation.

EXAMPLE 3.28

Flip two fair coins. Let *A* = tails on the first coin. Let *B* = tails on the second coin. Then *A* = {*TT*, *TH*} and *B* = {*TT*, *HT*}.

Therefore, A AND $B = \{TT\}$. A OR $B = \{TH, TT, HT\}$.

The sample space when you flip two fair coins is $X = \{HH, HT, TH, TT\}$. The outcome HH is in NEITHER A NOR B. The Venn diagram is as follows:

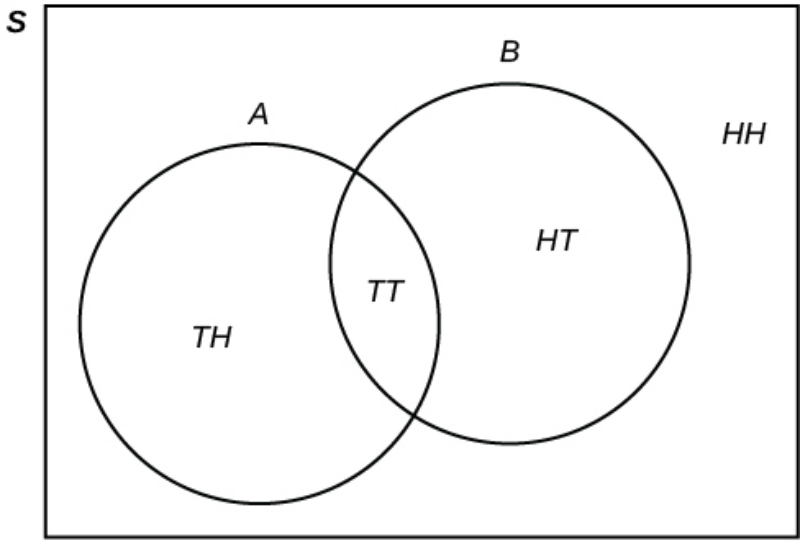


Figure 3.16

TRY IT 3.28

Roll a fair, six-sided die. Let A = a prime number of dots is rolled. Let B = an odd number of dots is rolled. Then $A = \{2, 3, 5\}$ and $B = \{1, 3, 5\}$. Therefore, A AND $B = \{3, 5\}$. A OR $B = \{1, 2, 3, 5\}$. The sample space for rolling a fair die is $S = \{1, 2, 3, 4, 5, 6\}$. Draw a Venn diagram representing this situation.

EXAMPLE 3.29

Problem

Forty percent of the students at a local college belong to a club and **50 percent** work part time. **Five percent** of the students work part time and belong to a club. Draw a Venn diagram showing the relationships. Let C = student belongs to a club and PT = student works part time.

Start by drawing a rectangle to represent the sample space. Then draw two circles or ovals inside the rectangle to represent the events of interest: belonging to a club (C) and working part time (PT). Always draw overlapping shapes to represent outcomes that are in both events.

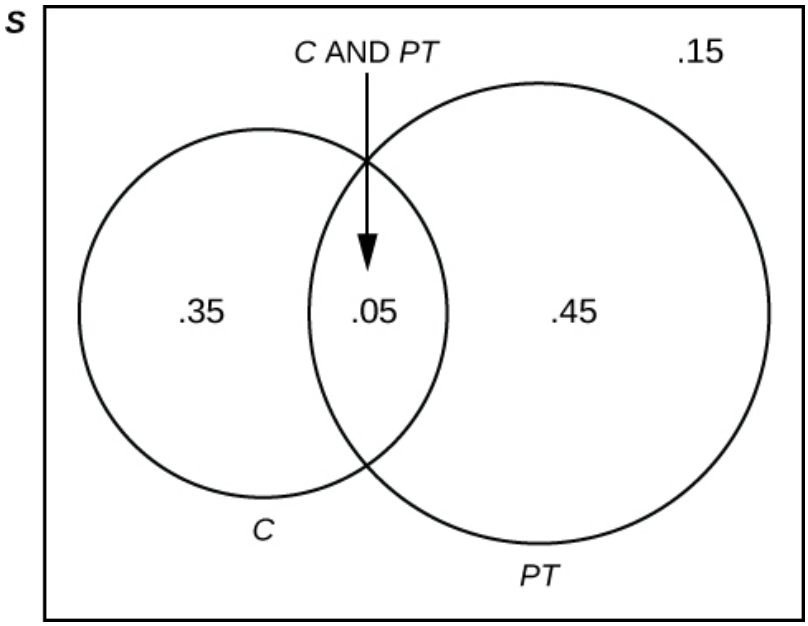


Figure 3.17

Label each piece of the diagram clearly and note the probability or frequency of each part. Start by labeling the overlapping section first. Note that the probabilities in *C* total 0.40 and the sum of the probabilities in *PT* is 0.50. The total of all probabilities displayed must be 1, representing 100 percent of the sample space.

If a student is selected at random, find the following:

a. the probability that the student belongs to a club.
b. the probability that the student works part time.
c. the probability that the student belongs to a club AND works part time.
d. the probability that the student belongs to a club **given** that the student works part time.
e. the probability that the student belongs to a club **OR** works part time.

Solution

$P(C) = .40$

$P(PT) = .50$

$P(C \text{ AND } PT) = .05$

$$P(C|PT) = \frac{P(C \text{ AND } PT)}{P(PT)} = \frac{.05}{.50} = .1$$

$P(C \text{ OR } PT) = P(C) + P(PT) - P(C \text{ AND } PT) = .40 + .50 - .05 = .85$

TRY IT 3.29

Fifty percent of the workers at a factory work a second job, 25 percent have a spouse who also works, and 5 percent work a second job and have a spouse who also works. Draw a Venn diagram showing the relationships. Let *W* = works a second job and *S* = spouse also works.

EXAMPLE 3.30

Problem

A person with type O blood and a negative Rh factor (Rh–) can donate blood to any person with any blood type. Four percent of African Americans have type O blood and a negative Rh factor, 5–10 percent of African Americans have the

Rh– factor, and 51 percent have type O blood.

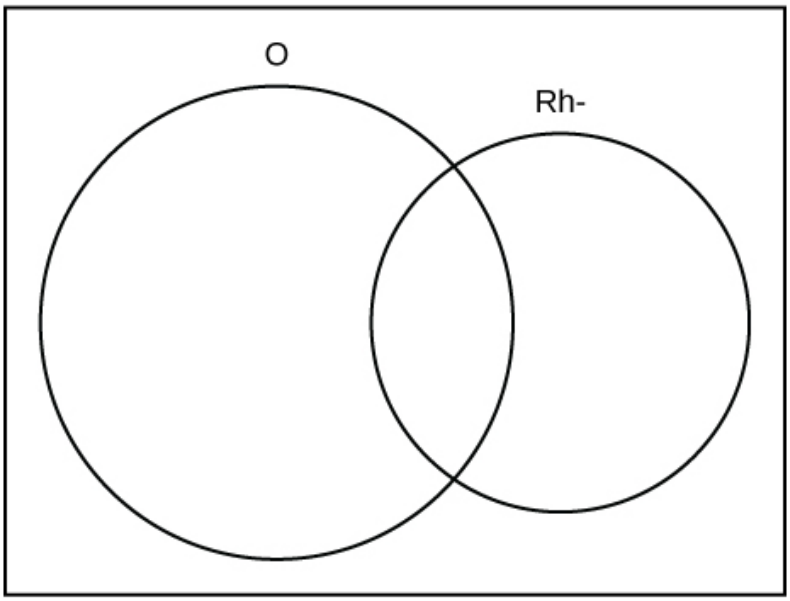


Figure 3.18

The "O" circle represents the African Americans with type O blood. The "Rh––" oval represents the African Americans with the Rh– –factor.

We will use the average of 5 percent and 10 percent, 7.5 percent, as the percentage of African Americans who have the Rh–– factor. Let O = African American with Type O blood and R = African American with Rh– –factor.

a. $P(O)$ = _____
b. $P(R)$ = _____
c. $P(O \text{ AND } R)$ = _____
d. $P(O \text{ OR } R)$ = _____
e. In the Venn Diagram, describe the overlapping area using a complete sentence.
f. In the Venn Diagram, describe the area in the rectangle but outside both the circle and the oval using a complete sentence.

Solution

a. $P(O) = .51$

b. $P(R) = .075$ because an average of 7.5 percent of African Americans have the Rh– –factor.

c. $P(O \text{ AND } R) = 0.04$ because 4 percent of African Americans have both Type O blood and the Rh– –factor.

d. $P(O \text{ OR } R) = P(O) + P(R) - P(O \text{ AND } R) = .51 + .075 - .04 = .545$

e. The area represents the African Americans that have type O blood and the Rh–– factor.

f. The area represents the African Americans that have neither type O blood nor the Rh–– factor.

TRY IT 3.30

In a bookstore, the probability that the customer buys a novel is .6, and the probability that the customer buys a nonfiction book is .4. Suppose that the probability that the customer buys both is .2.

a. Draw a Venn diagram representing the situation.
b. Find the probability that the customer buys either a novel or a nonfiction book.
c. In the Venn diagram, describe the overlapping area using a complete sentence.
d. Suppose that some customers buy only compact disks. Draw an oval in your Venn diagram representing this

event.

3.6 Probability Topics

Stats Lab

Probability Topics

Student Learning Outcomes

- The student will use theoretical and empirical methods to estimate probabilities.
- The student will appraise the differences between the two estimates.
- The student will demonstrate an understanding of long-term relative frequencies.

Do the Experiment

Count out 40 mixed-color candies, which is approximately one small bag's worth. Record the number of each color in Table 3.12. Use the information from this table to complete Table 3.13. Next, put the candies in a cup. The experiment is to pick two candies, one at a time. Do **not** look into the cup as you pick them. The first time through, replace the first candy before picking the second one. Record the results in the With Replacement column of Table 3.14. Do this 24 times. The second time through, after picking the first candy, do **not** replace it before picking the second one. Then, pick the second one. Record the results in the Without Replacement column section of Table 3.15. After you record the pick, put **both** candies back. Do this a total of 24 times, also. Use the data from Table 3.15 to calculate the empirical probability questions. Leave your answers in unreduced fractional form. Do **not** multiply out any fractions.

Color	Quantity
Yellow (Y)	
Green (G)	
Blue (BL)	
Brown (B)	
Orange (O)	
Red (R)	

Table 3.12 Population

	With Replacement	Without Replacement
P(2 reds)		
$P(R_1B_2$ OR $B_1R_2)$		
$P(R_1$ AND $G_2)$		
$P(G_2 \mid R_1)$		
P(no yellows)		

Table 3.13 Theoretical Probabilities

	With Replacement	Without Replacement
P(doubles)		
P(no doubles)		

Table 3.13 Theoretical Probabilities

Note

G_2 = green on second pick, R_1 = red on first pick, B_1 = brown on first pick, B_2 = brown on second pick, doubles = both picks are the same color.

With Replacement	Without Replacement
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)

Table 3.14 Empirical Results

	With Replacement	Without Replacement
P(2 reds)		
$P(R_1B_2$ OR $B_1R_2)$		
$P(R_1$ AND $G_2)$		

Table 3.15 Empirical Probabilities

	With Replacement	Without Replacement
$P(G_2 \mid R_1)$		
P(no yellows)		
P(doubles)		
P(no doubles)		

Table 3.15 Empirical Probabilities

Discussion Questions

1. Why are the With Replacement and Without Replacement probabilities different?
2. Convert P(no yellows) to decimal format for both Theoretical With Replacement and for Empirical With Replacement. Round to four decimal places.
   a. Theoretical *With Replacement*: P(no yellows) = _____
   b. Empirical With Replacement: P(no yellows) = _____
   c. Are the decimal values *close*? Did you expect them to be closer together or farther apart? Why?
3. If you increased the number of times you picked two candies to 240 times, why would empirical probability values change?
4. Would this change (see Question 3) cause the empirical probabilities and theoretical probabilities to be closer together or farther apart? How do you know?
5. Explain the differences in what $P(G_1$ AND $R_2)$ and $P(R_1 \mid G_2)$ represent. Hint: Think about the sample space for each probability.

Key Terms

conditional probability the likelihood that an event will occur given that another event has already occurred

contingency table the method of displaying a frequency distribution as a table with rows and columns to show how two variables may be dependent (contingent) upon each other; the table provides an easy way to calculate conditional probabilities

dependent events if two events are NOT independent, then we say that they are dependent

equally likely each outcome of an experiment has the same probability

event a subset of the set of all outcomes of an experiment; the set of all outcomes of an experiment is called a **sample space** and is usually denoted by S.
An event is an arbitrary subset in S. It can contain one outcome, two outcomes, no outcomes (empty subset), the entire sample space, and the like. Standard notations for events are capital letters such as A, B, C, and so on

experiment a planned activity carried out under controlled conditions

independent events The occurrence of one event has no effect on the probability of the occurrence of another event; events A and B are independent if one of the following is true:

1. $P(A|B) = P(A)$
2. $P(B|A) = P(B)$
3. $P(A \text{ AND } B) = P(A)P(B)$

mutually exclusive two events are mutually exclusive if the probability that they both happen at the same time is zero; if events A and B are mutually exclusive, then $P(A \text{ AND } B) = 0$

outcome a particular result of an experiment

probability a number between zero and one, inclusive, that gives the likelihood that a specific event will occur; the foundation of statistics is given by the following three axioms (by A.N. Kolmogorov, 1930s): Let S denote the sample space and A and B are two events in S; then

- $0 \leq P(A) \leq 1$,
- If A and B are any two mutually exclusive events, then $P(A \text{ OR } B) = P(A) + P(\text{B})$, and
- $P(S) = 1$

sample space the set of all possible outcomes of an experiment

sampling with replacement if each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once

sampling without replacement when sampling is done without replacement, each member of a population may be chosen only once

the AND event an outcome is in the event A AND B if the outcome is in both A AND B at the same time

the complement event the complement of event A consists of all outcomes that are NOT in A

the conditional probability of one event GIVEN another event $P(A|B)$ is the probability that event A will occur given that the event B has already occurred

the OR event an outcome is in the event A OR B if the outcome is in A or is in B or is in both A and B

the OR of two events an outcome is in the event A OR B if the outcome is in A, is in B, or is in both A and B

tree diagram the useful visual representation of a sample space and events in the form of a *tree* with branches marked by possible outcomes together with associated probabilities (frequencies, relative frequencies)

Venn diagram the visual representation of a sample space and events in the form of circles or ovals showing their intersections

Chapter Review

3.1 Terminology

In this module we learned the basic terminology of probability. The set of all possible outcomes of an experiment is called the sample space. Events are subsets of the sample space, and they are assigned a probability that is a number between zero and one, inclusive.

3.2 Independent and Mutually Exclusive Events

Two events A and B are independent if the knowledge that one occurred does not affect the chance the other occurs. If two events are not independent, then we say that they are dependent

In sampling with replacement, each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered not to be independent. When events do not share outcomes, they are mutually exclusive of each other.

3.3 Two Basic Rules of Probability

The multiplication rule and the addition rule are used for computing the probability of A and B, as well as the probability of A or B for two given events A, B defined on the sample space. In sampling with replacement, each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered to be not independent. The events A and B are mutually exclusive events when they do not have any outcomes in common.

3.4 Contingency Tables

There are several tools you can use to help organize and sort data when calculating probabilities. Contingency tables, also known as two-way tables, help display data and are particularly useful when calculating probabilites that have multiple dependent variables.

3.5 Tree and Venn Diagrams

A tree diagram uses branches to show the different outcomes of experiments and makes complex probability questions easy to visualize.

A Venn diagram is a picture that represents the outcomes of an experiment. It generally consists of a box that represents the sample space S together with circles or ovals. The circles or ovals represent events. A Venn diagram is especially helpful for visualizing the OR event, the AND event, and the complement of an event and for understanding conditional probabilities.

Formula Review

3.1 Terminology

A and B are events

$P(S) = 1$ where S is the sample space

$0 \leq P(A) \leq 1$

$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)}$

3.2 Independent and Mutually Exclusive Events

If A and B are independent, $P(A \text{ AND } B) = P(A)P(B)$, $P(A|B) = P(A)$, and $P(B|A) = P(B)$.

If A and B are mutually exclusive, $P(A \text{ OR } B) = P(A) + P(B)$ and $P(A \text{ AND } B) = 0$.

3.3 Two Basic Rules of Probability

The multiplication rule—$P(A \text{ AND } B) = P(A|B)P(B)$

The addition rule—$P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$

Practice

3.1 Terminology

1. In a particular college class, there are male and female students. Some students have long hair and some students have short hair. Write the **symbols** for the probabilities of the events for parts A through J of this question. Note that you cannot find numerical answers here. You were not given enough information to find any probability values yet; concentrate on understanding the symbols.

- Let F be the event that a student is female.
- Let M be the event that a student is male.
- Let S be the event that a student has short hair.
- Let L be the event that a student has long hair.

a. The probability that a student does not have long hair.
b. The probability that a student is male or has short hair.
c. The probability that a student is female and has long hair.
d. The probability that a student is male, given that the student has long hair.
e. The probability that a student has long hair, given that the student is male.
f. Of all female students, the probability that a student has short hair.
g. Of all students with long hair, the probability that a student is female.
h. The probability that a student is female or has long hair.
i. The probability that a randomly selected student is a male student with short hair.
j. The probability that a student is female.

Use the following information to answer the next four exercises. A box is filled with several party favors. It contains 12 hats, 15 noisemakers, 10 finger traps, and five bags of confetti.
Let H = the event of getting a hat.
Let N = the event of getting a noisemaker.
Let F = the event of getting a finger trap.
Let C = the event of getting a bag of confetti.

2. Find $P(H)$.

3. Find $P(N)$.

4. Find $P(F)$.

5. Find $P(C)$.

Use the following information to answer the next six exercises. A jar of 150 jelly beans contains 22 red jelly beans, 38 yellow, 20 green, 28 purple, 26 blue, and the rest are orange.
Let B = the event of getting a blue jelly bean
Let G = the event of getting a green jelly bean.
Let O = the event of getting an orange jelly bean.
Let P = the event of getting a purple jelly bean.
Let R = the event of getting a red jelly bean.
Let Y = the event of getting a yellow jelly bean.

6. Find $P(B)$.

7. Find $P(G)$.

8. Find $P(P)$.

9. Find $P(R)$.

10. Find $P(Y)$.

11. Find $P(O)$.

Use the following information to answer the next six exercises. There are 23 countries in North America, 12 countries in South America, 47 countries in Europe, 44 countries in Asia, 54 countries in Africa, and 14 countries in Oceania (Pacific Ocean region).
Let A = the event that a country is in Asia.
Let E = the event that a country is in Europe.
Let F = the event that a country is in Africa.
Let N = the event that a country is in North America.
Let O = the event that a country is in Oceania.
Let S = the event that a country is in South America.

12. Find $P(A)$.

13. Find $P(E)$.

14. Find $P(F)$.

15. Find $P(N)$.

16. Find $P(O)$.

17. Find $P(S)$.

18. What is the probability of drawing a red card in a standard deck of 52 cards?

19. What is the probability of drawing a club in a standard deck of 52 cards?

20. What is the probability of rolling an even number of dots with a fair, six-sided die numbered one through six?

21. What is the probability of rolling a prime number of dots with a fair, six-sided die numbered one through six?

Use the following information to answer the next two exercises. You see a game at a local fair. You have to throw a dart at a color wheel. Each section on the color wheel is equal in area.

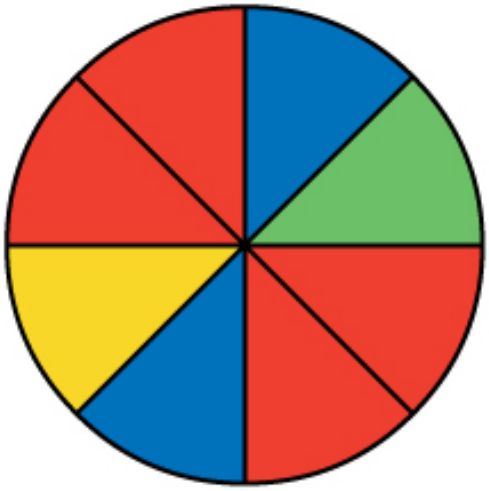


Figure 3.19

Let B = the event of landing on blue.
Let R = the event of landing on red.
Let G = the event of landing on green.
Let Y = the event of landing on yellow.

22. If you land on Y, you get the biggest prize. Find $P(Y)$.

23. If you land on red, you don't get a prize. What is $P(R)$?

Use the following information to answer the next 10 exercises. On a baseball team, there are infielders and outfielders. Some players are great hitters, and some players are not great hitters.
Let I = the event that a player in an infielder.
Let O = the event that a player is an outfielder.
Let H = the event that a player is a great hitter.
Let N = the event that a player is not a great hitter.

24. Write the symbols for the probability that a player is not an outfielder.

25. Write the symbols for the probability that a player is an outfielder or is a great hitter.

26. Write the symbols for the probability that a player is an infielder and is not a great hitter.

27. Write the symbols for the probability that a player is a great hitter, given that the player is an infielder.

28. Write the symbols for the probability that a player is an infielder, given that the player is a great hitter.

29. Write the symbols for the probability that of all the outfielders, a player is not a great hitter.

30. Write the symbols for the probability that of all the great hitters, a player is an outfielder.

31. Write the symbols for the probability that a player is an infielder or is not a great hitter.

32. Write the symbols for the probability that a player is an outfielder and is a great hitter.

33. Write the symbols for the probability that a player is an infielder.

34. What is the word for the set of all possible outcomes?

35. What is conditional probability?

36. A shelf holds 12 books. Eight are fiction and the rest are nonfiction. Each is a different book with a unique title. The fiction books are numbered one to eight. The nonfiction books are numbered one to four. Randomly select one book

Let F = event that book is fiction
Let N = event that book is nonfiction
What is the sample space?

37. What is the sum of the probabilities of an event and its complement?

Use the following information to answer the next two exercises. You are rolling a fair, six-sided number cube. Let E = the event that it lands on an even number. Let M = the event that it lands on a multiple of three.

38. What does $P(E|M)$ mean in words?

39. What does $P(E \text{ OR } M)$ mean in words?

3.2 Independent and Mutually Exclusive Events

40. E and F are mutually exclusive events. $P(E) = .4$; $P(F) = .5$. Find $P(E|F)$.

41. J and K are independent events. $P(J|K) = .3$. Find $P(J)$.

42. U and V are mutually exclusive events. $P(U) = .26$; $P(V) = .37$. Find the following:
a. $P(U \text{ AND } V) =$
b. $P(U|V) =$
c. $P(U \text{ OR } V) =$

43. Q and R are independent events. $P(Q) = .4$ and $P(Q \text{ AND } R) = .1$. Find $P(R)$.

3.3 Two Basic Rules of Probability

Use the following information to answer the next 10 exercises. Forty-eight percent of all voters of a certain state prefer life in prison without parole over the death penalty for a person convicted of first-degree murder. Among Latino registered voters in this state, 55 percent prefer life in prison without parole over the death penalty for a person convicted of first-degree murder. Of all citizens in this state, 37.6 percent are Latino.

In this problem, let

- C = citizens of a certain state (registered voters) preferring life in prison without parole over the death penalty for a person convicted of first-degree murder.
- L = registered voters of the state who are Latino.

Suppose that one citizen is randomly selected.

44. Find $P(C)$.

45. Find $P(L)$.

46. Find $P(C|L)$.

47. In words, what is $C|L$?

48. Find $P(L \text{ AND } C)$.

49. In words, what is L AND C?

50. Are L and C independent events? Show why or why not.

51. Find $P(L \text{ OR } C)$.

52. In words, what is L OR C?

53. Are L and C mutually exclusive events? Show why or why not.

3.4 Contingency Tables

Use the following information to answer the next four exercises. Table 3.16 shows a random sample of musicians and how they learned to play their instruments.

Gender	Self-Taught	Studied in School	Private Instruction	Total
Female	12	38	22	72
Male	19	24	15	58
Total	31	62	37	130

Table 3.16

54. Find *P*(musician is a female).

55. Find *P*(musician is a male AND had private instruction).

56. Find *P*(musician is a female OR is self taught).

57. Are the events *being a female musician* and *learning music in school* mutually exclusive events?

3.5 Tree and Venn Diagrams

58. The probability that a man develops some form of cancer in his lifetime is 0.4567. The probability that a man has at least one false-positive test result, meaning the test comes back for cancer when the man does not have it, is .51. Let *C* = a man develops cancer in his lifetime; *P* = a man has at least one false-positive test. Construct a tree diagram of the situation.

Bringing It Together: Practice

Use the following information to answer the next seven exercises. An article in the *New England Journal of Medicine*, reported about a study of people who use a product in California and Hawaii. In one part of the report, the self-reported ethnicity and using the product levels per day were given. Of the people using the product at most 10 times a day, there were 9,886 African Americans, 2,745 Native Hawaiians, 12,831 Latinos, 8,378 Japanese Americans, and 7,650 whites. Of the people using the product 11 to 20 times per day, there were 6,514 African Americans, 3,062 Native Hawaiians, 4,932 Latinos, 10,680 Japanese Americans, and 9,877 whites. Of the people using the product 21 to 30 times per day, there were 1,671 African Americans, 1,419 Native Hawaiians, 1,406 Latinos, 4,715 Japanese Americans, and 6,062 Whites. Of the people using the product at least 31 times per day, there were 759 African Americans, 788 Native Hawaiians, 800 Latinos, 2,305 Japanese Americans, and 3,970 whites.

59. Complete the table using the data provided. Suppose that one person from the study is randomly selected. Find the probability that person used the product 11 to 20 times a day.

Product Use (times per day)	African Americans	Native Hawaiians	Latinos	Japanese Americans	Whites	TOTALS
1–10						
11–20						
21–30						
31+						
TOTALS						

Table 3.17 Product Use by Ethnicity

60. Suppose that one person from the study is randomly selected. Find the probability that the person used the

product 11 to 20 times per day.

61. Find the probability that the person was Latino.

62. In words, explain what it means to pick one person from the study who is Japanese American **AND** uses the product 21 to 30 times per day. Also, find the probability.

63. In words, explain what it means to pick one person from the study who is Japanese American **OR** uses the product 21 to 30 times per day. Also, find the probability.

64. In words, explain what it means to pick one person from the study who is Japanese American **GIVEN** that the person uses the product 21 to 30 times per day. Also, find the probability.

65. Prove that product use/day and ethnicity are dependent events.

Homework

3.1 Terminology

66.

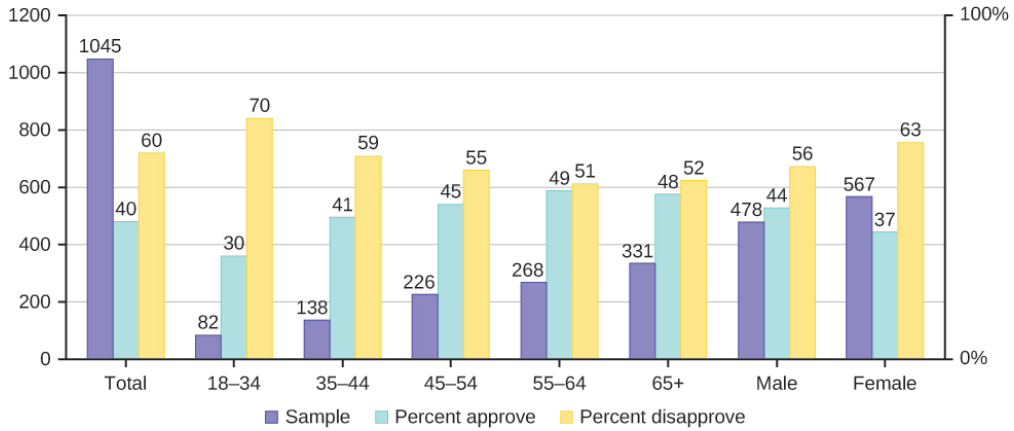


Figure 3.20

The graph in Figure 3.20 displays the sample sizes and percentages of people in different age and gender groups who were polled concerning their approval of Mayor Ford's actions in office. The total number in the sample of all the age groups is 1,045.

a. Define three events in the graph.
b. Describe in words what the entry 40 means.
c. Describe in words the complement of the entry in the previous question.
d. Describe in words what the entry 30 means.
e. Out of the males and females, what percent are males?
f. Out of the females, what percent disapprove of Mayor Ford?
g. Out of all the age groups, what percent approve of Mayor Ford?
h. Find $P(\text{Approve} \mid \text{Male})$.
i. Out of the age groups, what percent are more than 44 years old?
j. Find $P(\text{Approve} \mid \text{Age} < 35)$.

67. Explain what is wrong with the following statements. Use complete sentences.

a. If there is a 60 percent chance of rain on Saturday and a 70 percent chance of rain on Sunday, then there is a 130 percent chance of rain over the weekend.
b. The probability that a baseball player hits a home run is greater than the probability that he gets a successful hit.

3.2 Independent and Mutually Exclusive Events

Use the following information to answer the next 12 exercises. The graph shown is based on more than 170,000 interviews that took place from January through December 2012. The sample consists of employed Americans 18 years of age or older. The Health Index Scores are the sample space. We randomly sample one type of Health Index Score, the emotional well-being score.

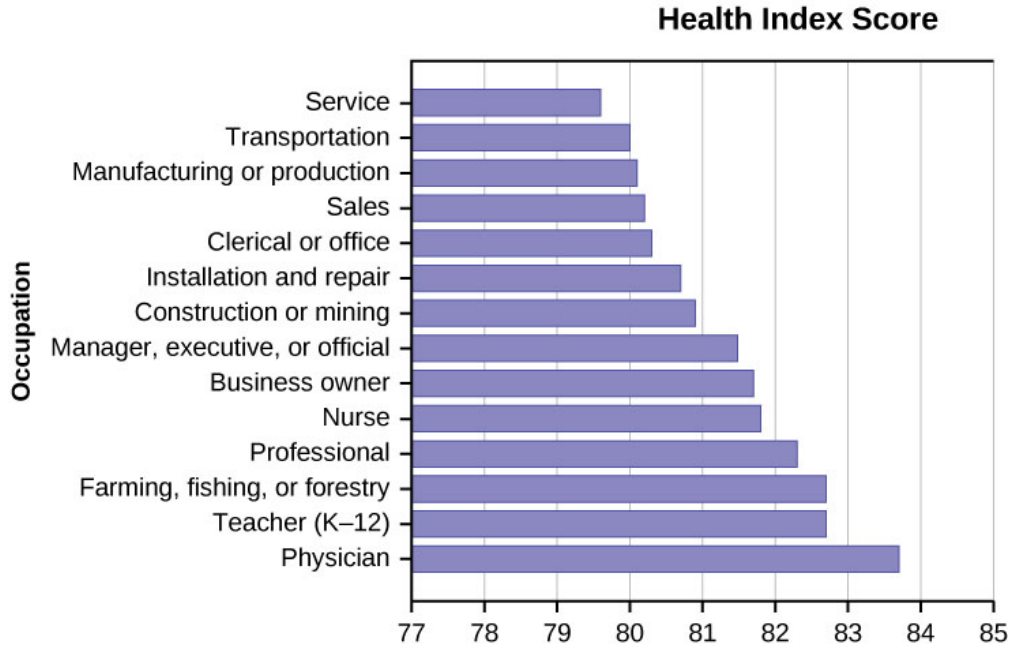


Figure 3.21

68. Find the probability that a Health Index Score is 82.7.

69. Find the probability that a Health Index Score is 81.0.

70. Find the probability that a Health Index Score is more than 81.

71. Find the probability that a Health Index Score is between 80.5 and 82.

72. If we know a Health Index Score is 81.5 or more, what is the probability that it is 82.7?

73. What is the probability that a Health Index Score is 80.7 or 82.7?

74. What is the probability that a Health Index Score is less than 80.2 given that it is already less than 81?

75. What occupation has the highest Health Index Score?

76. What occupation has the lowest emotional index score?

77. What is the range of the data?

78. Compute the average Health Index Score.

79. If all occupations are equally likely for a certain individual, what is the probability that he or she will have an occupation with lower than average Health Index Score?

3.3 Two Basic Rules of Probability

80. On February 28, 2013, a Field Poll Survey reported that 61 percent of California registered voters approved of a law that was about to be passed. Among 18- to 39-year olds (California registered voters), the approval rating was 78 percent. Six in 10 California registered voters said that the upcoming Supreme Court's ruling about the constitutionality of the law was either very or somewhat important to them. Out of those registered voters who supported the law, 75 percent say the ruling is important to them.
In this problem, let

- C = California registered voters who supported the law,
- B = California registered voters who say the Supreme Court's ruling about the law is very or somewhat

important to them, and
- A = California registered voters who are 18 to 39 years old.

a. Find $P(C)$.
b. Find $P(B)$.
c. Find $P(C|A)$.
d. Find $P(B|C)$.
e. In words, what is $C|A$?
f. In words, what is $B|C$?
g. Find $P(C \text{ AND } B)$.
h. In words, what is C AND B?
i. Find $P(C \text{ OR } B)$.
j. Are C and B mutually exclusive events? Show why or why not.

81. After a mayor of a major Canadian city announced his plans to cut budget costs in late 2011, researchers polled 1,046 people to measure the mayor's popularity. Everyone polled expressed either approval or disapproval. These are the results their poll produced:
- In early 2011, 60 percent of the population approved of the mayor's actions in office.
- In mid-2011, 57 percent of the population approved of his actions.
- In late 2011, the percentage of popular approval was measured at 42 percent.

a. What is the sample size for this study?
b. What proportion in the poll disapproved of the mayor, according to the results from late 2011?
c. How many people polled responded that they approved of the mayor in late 2011?
d. What is the probability that a person supported the mayor, based on the data collected in mid-2011?
e. What is the probability that a person supported the mayor, based on the data collected in early 2011?

Use the following information to answer the next three exercises. A local restaurant sells pork chops and chicken breasts. The given values below are the weights (in ounces) of pork chops and chicken breasts listed on the menu. Your server will randomly select one piece of meat (pork chop or chicken breast) that you will be served.

Pork Chops	17	20	21	18	20	20	20	18	19	19
	20	19	21	20	18	20	20	19	18	19
Chicken Breasts	17	19	17	21	17	21	18	21	19	21
	20	17	20	18	19	20	20	17	21	20

Table 3.18

82. a. List the sample space of the possible items that are on the menu.
b. Find P(you will get a 17-oz. piece of meat).
c. Find P(you will get a pork chop).
d. Find P(you will get a 17-oz. pork chop).
e. Is getting a pork chop the complement of getting a chicken breast? Why?
f. Find two mutually exclusive events.
g. Are the events *getting 17 oz. of meat* and *getting a pork chop* independent?

83. Compute the probabilities.
a. P(you will get a chicken breast)
b. P(you will get a 17-oz. chicken breast)
c. P(you will get a chicken breast **or** you will not get a 17-oz. pork chop)
d. P(you will not get a chicken breast **and** you will get an 18-oz. pork chop)
e. P(you will get a piece of meat that is not 21 oz.)
f. P(you will get a piece of chicken that is not 21 oz.)
g. P(you will not get a chicken breast **and** you will not get a pork chop)

84. Compute the probabilities:

a. P(you will not get a pork chop)
b. P(you will get a 20-oz. pork chop)
c. P(you will not get a chicken breast **or** you will not get an 18-oz. pork chop)
d. P(you will not get a chicken breast **and** you will not get an 18-oz. pork chop)
e. P(you will get a pork chop that is not 21 oz.)
f. P(you will not get a chicken breast **or** you will not get a pork chop)

85. Suppose that you have eight cards. Five are green and three are yellow. The five green cards are numbered 1, 2, 3, 4, and 5. The three yellow cards are numbered 1, 2, and 3. The cards are well shuffled. You randomly draw one card.

- G = card drawn is green
- E = card drawn is even-numbered

a. List the sample space.
b. $P(G)$ = _____
c. $P(G \mid E)$ = _____
d. $P(G \text{ AND } E)$ = _____
e. $P(G \text{ OR } E)$ = _____
f. Are G and E mutually exclusive? Justify your answer numerically.

86. Roll two fair dice separately. Each die has six faces.

a. List the sample space.
b. Let A be the event that either a three or four is rolled first, followed by an even number. Find $P(A)$.
c. Let B be the event that the sum of the two rolls is at most seven. Find $P(B)$.
d. In words, explain what $P(A \mid B)$ represents. Find $P(A \mid B)$.
e. Are A and B mutually exclusive events? Explain your answer in one to three complete sentences, including numerical justification.
f. Are A and B independent events? Explain your answer in one to three complete sentences, including numerical justification.

87. A special deck of cards has 10 cards. Four are green, three are blue, and three are red. When a card is picked, its color is recorded. An experiment consists of first picking a card and then tossing a coin.

a. List the sample space.
b. Let A be the event that a blue card is picked first, followed by landing a head on the coin toss. Find $P(A)$.
c. Let B be the event that a red or green is picked, followed by landing a head on the coin toss. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.
d. Let C be the event that a red or blue is picked, followed by landing a head on the coin toss. Are the events A and C mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

88. An experiment consists of first rolling a die and then tossing a coin.

a. List the sample space.
b. Let A be the event that either a three or a four is rolled first, followed by landing a head on the coin toss. Find $P(A)$.
c. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

89. An experiment consists of tossing a nickel, a dime, and a quarter. Of interest is the side the coin lands on.

a. List the sample space.
b. Let A be the event that there are at least two tails. Find $P(A)$.
c. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including justification.

90. Consider the following scenario:
Let $P(C) = .4$.
Let $P(D) = .5$.
Let $P(C \mid D) = .6$.

a. Find $P(C \text{ AND } D)$.

b. Are C and D mutually exclusive? Why or why not?
c. Are C and D independent events? Why or why not?
d. Find $P(C \text{ OR } D)$.
e. Find $P(D \mid C)$.

91. Y and Z are independent events.
a. Rewrite the basic Addition Rule $P(Y \text{ OR } Z) = P(Y) + P(Z) - P(Y \text{ AND } Z)$ using the information that Y and Z are independent events.
b. Use the rewritten rule to find $P(Z)$ if $P(Y \text{ OR } Z) = .71$ and $P(Y) = .42$.

92. G and H are mutually exclusive events. $P(G) = .5$ $P(H) = .3$
a. Explain why the following statement MUST be false: $P(H \mid G) = .4$.
b. Find $P(H \text{ OR } G)$.
c. Are G and H independent or dependent events? Explain in a complete sentence.

93. Approximately 281,000,000 people over age five live in the United States. Of these people, 55,000,000 speak a language other than English at home. Of those who speak another language at home, 62.3 percent speak Spanish.
Let E = speaks English at home; E' = speaks another language at home; and S = speaks Spanish.
Finish each probability statement by matching the correct answer.

Probability Statements	Answers
a. $P(E')$ =	i. .8043
b. $P(E)$ =	ii. .623
c. $P(S \text{ and } E')$ =	iii. .1957
d. $P(S \mid E')$ =	iv. .1219

Table 3.19

94. In 1994, the U.S. government held a lottery to issue 55,000 licenses of a certain type. Renate Deutsch, from Germany, was one of approximately 6.5 million people who entered this lottery. Let G = won license.
a. What was Renate's chance of winning one of the licenses? Write your answer as a probability statement.
b. In the summer of 1994, Renate received a letter stating she was one of 110,000 finalists chosen. Once the finalists were chosen, assuming that each finalist had an equal chance to win, what was Renate's chance of winning one of the licenses? Write your answer as a conditional probability statement. Let F = was a finalist.
c. Are G and F independent or dependent events? Justify your answer numerically and also explain why.
d. Are G and F mutually exclusive events? Justify your answer numerically and explain why.

95. Three professors at George Washington University did an experiment to determine if economists are more likely to return found money than other people. They dropped 64 stamped, addressed envelopes with $10 cash in different classrooms on the George Washington campus. Forty-four percent were returned overall. From the economics classes 56 percent of the envelopes were returned. From the business, psychology, and history classes 31 percent were returned.
Let R = money returned; E = economics classes; and O = other classes.
a. Write a probability statement for the overall percentage of money returned.
b. Write a probability statement for the percentage of money returned out of the economics classes.
c. Write a probability statement for the percentage of money returned out of the other classes.
d. Is money being returned independent of the class? Justify your answer numerically and explain it.
e. Based upon this study, do you think that economists are more selfish than other people? Explain why or why not. Include numbers to justify your answer.

96. The following table of data obtained from www.baseball-almanac.com shows hit information for four players. Suppose that one hit from the table is randomly selected.

Name	Single	Double	Triple	Home Run	Total Hits
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
Total	8,471	1,577	583	1,720	12,351

Table 3.20

Are *the hit being made by Hank Aaron* and *the hit being a double* independent events?

a. Yes, because $P(\text{hit by Hank Aaron} \mid \text{hit is a double}) = P(\text{hit by Hank Aaron})$
b. No, because $P(\text{hit by Hank Aaron} \mid \text{hit is a double}) \neq P(\text{hit is a double})$
c. No, because $P(\text{hit is by Hank Aaron} \mid \text{hit is a double}) \neq P(\text{hit by Hank Aaron})$
d. Yes, because $P(\text{hit is by Hank Aaron} \mid \text{hit is a double}) = P(\text{hit is a double})$

97. United Blood Services is a blood bank that serves more than 500 hospitals in 18 states. According to their website, a person with type O blood and a negative Rh factor (Rh–) can donate blood to any person with any bloodtype. Their data show that 43 percent of people have type O blood and 15 percent of people have Rh– factor; 52 percent of people have type O or Rh– factor.

a. Find the probability that a person has both type O blood and the Rh– factor.
b. Find the probability that a person does **not** have both type O blood and the Rh– factor.

98. At a college, 72 percent of courses have final exams and 46 percent of courses require research papers. Suppose that 32 percent of courses have a research paper and a final exam. Let F be the event that a course has a final exam. Let R be the event that a course requires a research paper.

a. Find the probability that a course has a final exam or a research project.
b. Find the probability that a course has neither of these two requirements.

99. In a box of assorted cookies, 36 percent contain chocolate and 12 percent contain nuts. Of those, 8 percent contain both chocolate and nuts. Sean is allergic to both chocolate and nuts.

a. Find the probability that a cookie contains chocolate or nuts (he can't eat it).
b. Find the probability that a cookie does not contain chocolate or nuts (he can eat it).

100. A college finds that 10 percent of students have taken a distance learning class and that 40 percent of students are part-time students. Of the part-time students, 20 percent have taken a distance learning class. Let D = event that a student takes a distance learning class and E = event that a student is a part-time student.

a. Find $P(D \text{ AND } E)$.
b. Find $P(E \mid D)$.
c. Find $P(D \text{ OR } E)$.
d. Using an appropriate test, show whether D and E are independent.
e. Using an appropriate test, show whether D and E are mutually exclusive.

3.4 Contingency Tables

Use the information in the Table 3.21 to answer the next eight exercises. The table shows the political party affiliation of each of 67 members of the U.S. Senate in June 2012, and when they would next be up for reelection.

Up for Reelection:	Democratic Party	Republican Party	Other	Total
November 2014	20	13	0	
November 2016	10	24	0	
Total				

Table 3.21

101. What is the probability that a randomly selected senator had an *Other* affiliation?

102. What is the probability that a randomly selected senator would be up for reelection in November 2016?

103. What is the probability that a randomly selected senator was a Democrat and was up for reelection in November 2016?

104. What is the probability that a randomly selected senator was a Republican or was up for reelection in November 2014?

105. Suppose that a member of the U.S. Senate is randomly selected. Given that the randomly selected senator was up for reelection in November 2016, what is the probability that this senator was a Democrat?

106. Suppose that a member of the U.S. Senate is randomly selected. What is the probability that the senator was up for reelection in November 2014, knowing that this senator was a Republican?

107. The events *Republican* and *Up for reelection in 2016* are _____.
a. mutually exclusive
b. independent
c. both mutually exclusive and independent
d. neither mutually exclusive nor independent

108. The events *Other* and *Up for reelection in November 2016* are _____.
a. mutually exclusive
b. independent
c. both mutually exclusive and independent
d. neither mutually exclusive nor independent

Use the following information to answer the next two exercises. The table of data obtained from www.baseball-almanac.com shows hit information for four well-known baseball players. Suppose that one hit from the table is randomly selected.

Name	Single	Double	Triple	Home Run	Total Hits
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
TOTAL	8,471	1,577	583	1,720	12,351

Table 3.22

109. Find P(Hit was made by Babe Ruth).

a. $\frac{1,518}{2,873}$
b. $\frac{2,873}{12,351}$
c. $\frac{583}{12,351}$
d. $\frac{4,189}{12,351}$

110. Find P(Hit was made by Ty Cobb | The hit was a Home Run).

a. $\frac{4,189}{12,351}$
b. $\frac{114}{1,720}$
c. $\frac{1,720}{4,189}$
d. $\frac{114}{12,351}$

111. Table 3.23 identifies a group of children by one of four hair colors, and by type of hair.

Hair Type	Brown	Blond	Black	Red	Totals
Wavy	20		15	3	43
Straight	80	15		12	
Totals		20			215

Table 3.23

a. Complete the table.
b. What is the probability that a randomly selected child will have wavy hair?
c. What is the probability that a randomly selected child will have either brown or blond hair?
d. What is the probability that a randomly selected child will have wavy brown hair?
e. What is the probability that a randomly selected child will have red hair, given that he or she has straight hair?
f. If B is the event of a child having brown hair, find the probability of the complement of B.
g. In words, what does the complement of B represent?

112. In a previous year, the weights of the members of a California football team and a Texas football team were published in a newspaper. The factual data were compiled into the following table. The weights in the column headings are in pounds.

Shirt #	≤ 210	211–250	251–290	> 290
1–33	21	5	0	0
34–66	6	18	7	4
66–99	6	12	22	5

Table 3.24

For the following, suppose that you randomly select one player from the California team or the Texas team.

a. Find the probability that his shirt number is from 1 to 33.
b. Find the probability that he weighs at most 210 pounds.
c. Find the probability that his shirt number is from 1 to 33 AND he weighs at most 210 pounds.
d. Find the probability that his shirt number is from 1 to 33 OR he weighs at most 210 pounds.
e. Find the probability that his shirt number is from 1 to 33 GIVEN that he weighs at most 210 pounds.

3.5 Tree and Venn Diagrams

Use the following information to answer the next two exercises. This tree diagram shows the tossing of an unfair coin followed by drawing one bead from a cup containing three red (*R*), four yellow (*Y*), and five blue (*B*) beads. For the coin, $P(H) = \frac{2}{3}$ and $P(T) = \frac{1}{3}$ where *H* is heads and *T* is tails.

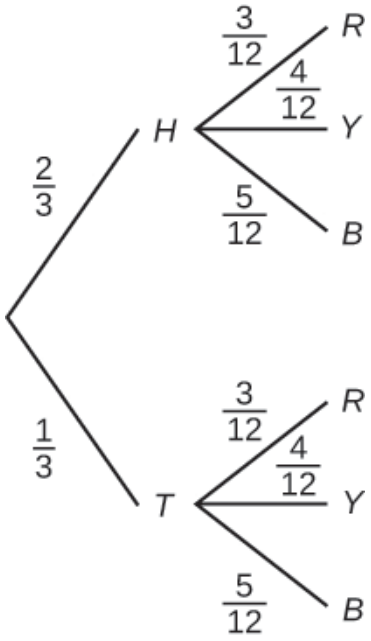


Figure 3.22

113. Find *P*(tossing a head on the coin AND a red bead).

a. $\frac{2}{3}$
b. $\frac{5}{15}$
c. $\frac{6}{36}$
d. $\frac{5}{36}$

114. Find *P*(blue bead).

a. $\frac{15}{36}$
b. $\frac{10}{36}$
c. $\frac{10}{12}$
d. $\frac{6}{36}$

115. A box of cookies contains three chocolate and seven butter cookies. Miguel randomly selects a cookie and eats it. Then he randomly selects another cookie and eats it. How many cookies did he take?

a. Draw the tree that represents the possibilities for the cookie selections. Write the probabilities along each branch of the tree.
b. Are the probabilities for the flavor of the second cookie that Miguel selects independent of his first selection? Explain.
c. For each complete path through the tree, write the event it represents and find the probabilities.
d. Let *S* be the event that both cookies selected were the same flavor. Find *P*(*S*).
e. Let *T* be the event that the cookies selected were different flavors. Find *P*(*T*) by two different methods by using the complement rule and by using the branches of the tree. Your answers should be the same with both methods.
f. Let *U* be the event that the second cookie selected is a butter cookie. Find *P*(*U*).

Bringing It Together: Homework

116. A previous year, the weights of the members of a California football team and a Texas football team were published in a newspaper The factual data are compiled into Table 3.25.

Shirt#	≤ 210	211–250	251–290	290≤
1–33	21	5	0	0
34–66	6	18	7	4
66–99	6	12	22	5

Table 3.25

For the following, suppose that you randomly select one player from the California team or the Texas team. If having a shirt number from one to 33 and weighing at most 210 pounds were independent events, then what should be true about $P(\text{Shirt\# 1–33} \mid \leq 210 \text{ pounds})$?

117. The probability that a male develops some form of cancer in his lifetime is .4567. The probability that a male has at least one false-positive test result, meaning the test comes back for cancer when the man does not have it, is .51. Some of the following questions do not have enough information for you to answer them. Write *not enough information* for those answers. Let C = a man develops cancer in his lifetime and P = a man has at least one false-positive.
   a. $P(C)$ = _____
   b. $P(P \mid C)$ = _____
   c. $P(P \mid C')$ = _____
   d. If a test comes up positive, based upon numerical values, can you assume that man has cancer? Justify numerically and explain why or why not.

118. Given events G and H: $P(G) = .43$; $P(H) = .26$; $P(H \text{ AND } G) = .14$
   a. Find $P(H \text{ OR } G)$.
   b. Find the probability of the complement of event (H AND G).
   c. Find the probability of the complement of event (H OR G).

119. Given events J and K: $P(J) = .18$; $P(K) = .37$; $P(J \text{ OR } K) = .45$
   a. Find $P(J \text{ AND } K)$.
   b. Find the probability of the complement of event (J AND K).
   c. Find the probability of the complement of event (J OR K).

Use the following information to answer the next two exercises. Suppose that you have eight cards. Five are green and three are yellow. The cards are well shuffled.

120. Suppose that you randomly draw two cards, one at a time, **with replacement**.
Let G_1 = first card is green
Let G_2 = second card is green
   a. Draw a tree diagram of the situation.
   b. Find $P(G_1 \text{ AND } G_2)$.
   c. Find P(at least one green).
   d. Find $P(G_2 \mid G_1)$.
   e. Are G_2 and G_1 independent events? Explain why or why not.

121. Suppose that you randomly draw two cards, one at a time, **without replacement**.
G_1 = first card is green
G_2 = second card is green
   a. Draw a tree diagram of the situation.
   b. Find $P(G_1 \text{ AND } G_2)$.
   c. Find P(at least one green).
   d. Find $P(G_2 \mid G_1)$.
   e. Are G_2 and G_1 independent events? Explain why or why not.

Use the following information to answer the next two exercises. The percent of licensed U.S. drivers (from a recent year) who are female is 48.60. Of the females, 5.03 percent are age 19 and under; 81.36 percent are age 20–64; 13.61 percent are age 65 or over. Of the licensed U.S. male drivers, 5.04 percent are age 19 and under; 81.43 percent are age 20–64; 13.53 percent are age 65 or over.

122. Complete the following:
a. Construct a table or a tree diagram of the situation.
b. Find *P*(driver is female).
c. Find *P*(driver is age 65 or over | driver is female).
d. Find *P*(driver is age 65 or over AND female).
e. In words, explain the difference between the probabilities in Part c and Part d.
f. Find *P*(driver is age 65 or over).
g. Are being age 65 or over and being female mutually exclusive events? How do you know?

123. Suppose that 10,000 U.S. licensed drivers are randomly selected.
a. How many would you expect to be male?
b. Using the table or tree diagram, construct a contingency table of gender versus age group.
c. Using the contingency table, find the probability that out of the age 20–64 group, a randomly selected driver is female.

124. Approximately 86.5 percent of Americans commute to work by car, truck, or van. Out of that group, 84.6 percent drive alone and 15.4 percent drive in a carpool. Approximately 3.9 percent walk to work and approximately 5.3 percent take public transportation.
a. Construct a table or a tree diagram of the situation. Include a branch for all other modes of transportation to work.
b. Assuming that the walkers walk alone, what percent of all commuters travel alone to work?
c. Suppose that 1,000 workers are randomly selected. How many would you expect to travel alone to work?
d. Suppose that 1,000 workers are randomly selected. How many would you expect to drive in a carpool?

125. When the euro coin was introduced in 2002, two math professors had their statistics students test whether the Belgian one euro-coin was a fair coin. They spun the coin rather than tossing it and found that out of 250 spins, 140 showed a head (event *H*) while 110 showed a tail (event *T*). On that basis, they claimed that it is not a fair coin.
a. Based on the given data, find *P*(*H*) and *P*(*T*).
b. Use a tree to find the probabilities of each possible outcome for the experiment of spinning the coin twice.
c. Use the tree to find the probability of obtaining exactly one head in two spins of the coin.
d. Use the tree to find the probability of obtaining at least one head.

126. *Use the following information to answer the next two exercises.* The following are real data from Santa Clara County, California. As of a certain time, there had been a total of 3,059 documented cases of a disease in the county. They were grouped into the following categories, with risk factors of becoming ill with the disease labeled as Methods A, B, and C and Other:

	Method A	Method B	Method C	Other	Totals
Female	0	70	136	49	___
Male	2,146	463	60	135	___
Totals	___	___	___	___	___

Table 3.26

Suppose a person with a disease in Santa Clara County is randomly selected.
a. Find *P*(Person is female).
b. Find *P*(Person has a risk factor of method C).
c. Find *P*(Person is female OR has a risk factor of method B).

d. Find *P*(Person is female AND has a risk factor of method A).
e. Find *P*(Person is male AND has a risk factor of method B).
f. Find *P*(Person is female GIVEN person got the disease from method C).
g. Construct a Venn diagram. Make one group females and the other group method C.

127. Answer these questions using probability rules. Do NOT use the contingency table. Three thousand fifty-nine cases of a disease had been reported in Santa Clara County, California, through a certain date. Those cases will be our population. Of those cases, 6.4 percent obtained the disease through method C and 7.4 percent are female. Out of the females with the disease, 53.3 percent got the disease from method C.
a. Find *P*(Person is female).
b. Find *P*(Person obtained the disease through method C).
c. Find *P*(Person is female GIVEN person got the disease from method C)
d. Construct a Venn diagram representing this situation. Make one group females and the other group method C. Fill in all values as probabilities.

References

3.1 Terminology

Worldatlas. (2013). *Countries list by continent*. Retrieved from http://www.worldatlas.com/cntycont.htm

3.2 Independent and Mutually Exclusive Events

Gallup. (n.d.). Retrieved from www.gallup.com/

Lopez, S., and Sidhu, P. (2013, March 28). U.S. teachers love their lives, but struggle in the workplace. *Gallup Wellbeing*. http://www.gallup.com/poll/161516/teachers-love-lives-struggle-workplace.aspx

3.3 Two Basic Rules of Probability

Baseball Almanac. (2013). Retrieved from www.baseball-almanac.com

DiCamillo, Mark, and Field, M. The file poll. *Field Research Corporation*. Retrieved from http://www.field.com/fieldpollonline/subscribers/Rls2443.pdf

Field Research Corporation. (n.d.). Retrieved from www.field.com/fieldpollonline

Forum Research. (n.d.). *Mayor's approval down*. Retrieved from http://www.forumresearch.com/forms/News Archives/News Releases/74209_TO_Issues_-_Mayoral_Approval_%28Forum_Research%29%2820130320%29.pdf

Rider, D. (2011, Sept. 14). Ford support plummeting, poll suggests. *The Star*. Retrieved from http://www.thestar.com/news/gta/2011/09/14/ford_support_plummeting_poll_suggests.html

Shin, H. B., and Kominski, R. A. (2010 April). Language use in the United States: 2007 (American Community Survey Reports, ACS-12). Washington, DC: United States Census Bureau. Retrieved from http://www.census.gov/hhes/socdemo/language/data/acs/ACS-12.pdf

The Roper Center for Public Opinion Research. (n.d.). *Archives*. Retrieved from http://www.ropercenter.uconn.edu/

The Wall Street Journal. (n.d.). Retrieved from https://www.wsj.com/

U.S. Census Bureau. (n.d.). Retrieved from https://www.census.gov/

Wikipedia. (n.d.). *Roulette*. Retrieved from http://en.wikipedia.org/wiki/Roulette

3.4 Contingency Tables

American Red Cross. (2013). Blood Types. Retrieved from http://www.redcrossblood.org/learn-about-blood/bloodtypes

Centers for Disease Control and Prevention/National Center for Health Statistics, United States Department of Health and Human Services. (n.d.). Retrieved from https://www.cdc.gov/nchs/

Haiman, C. A., et al. (2006, Jan. 26). Ethnic and racial differences in the smoking-related risk of lung cancer. *The New England Journal of Medicine*. Retrieved from http://www.nejm.org/doi/full/10.1056/NEJMoa033250

Samuels, T. M. (2013). Strange facts about RH negative blood. *eHow Health*. Retrieved from http://www.ehow.com/facts_5552003_strange-rh-negative-blood.html

The Disaster Center Crime Pages. (n.d.). *United States: Uniform crime report – state statistics from 1960–2011*. Retrieved

from http://www.disastercenter.com/crime/

United Blood Services. (2011). *Human blood types*. Retrieved from http://www.unitedbloodservices.org/learnMore.aspx

United States Senate. (n.d.). Retrieved from www.senate.gov

3.5 Tree and Venn Diagrams

American Cancer Society. (n.d.). Retrieved from https://www.cancer.org/

Clara County Public Health Department. (n.d.). Retrieved from https://www.sccgov.org/sites/sccphd/en-us/pages/phd.aspx

Federal Highway Administration, U.S. Department of Transportation. (n.d.). Retrieved from https://www.fhwa.dot.gov/

The Data and Story Library. (1996). Retrieved from http://lib.stat.cmu.edu/DASL/

The Roper Center for Public Opinion Research. (2013). *Search for datasets*. Retrieved from http://www.ropercenter.uconn.edu/data_access/data/search_for_datasets.html

USA Today. (n.d.). Retrieved from https://www.usatoday.com/

U.S. Census Bureau. (n.d.). Retrieved from https://www.census.gov/

World Bank Group. (2013). *Environment*. Available online at http://data.worldbank.org/topic/environment

Solutions

1. a. $P(L') = P(S)$
b. $P(M \text{ OR } S)$
c. $P(F \text{ AND } L)$
d. $P(M \mid L)$
e. $P(L \mid M)$
f. $P(S \mid F)$
g. $P(F \mid L)$
h. $P(F \text{ OR } L)$
i. $P(M \text{ AND } S)$
j. $P(F)$

3. $P(N) = \frac{15}{42} = \frac{5}{14} = .36$

5. $P(C) = \frac{5}{42} = .12$

7. $P(G) = \frac{20}{150} = \frac{2}{15} = .13$

9. $P(R) = \frac{22}{150} = \frac{11}{75} = .15$

11. $P(O) = \frac{150-22-38-20-28-26}{150} = \frac{16}{150} = \frac{8}{75} = .11$

13. $P(E) = \frac{47}{194} = .24$

15. $P(N) = \frac{23}{194} = .12$

17. $P(S) = \frac{12}{194} = \frac{6}{97} = .06$

19. $\frac{13}{52} = \frac{1}{4} = .25$

21. $\frac{3}{6} = \frac{1}{2} = .5$

23. $P(R) = \frac{4}{8} = .5$

25. $P(O \text{ OR } H)$

27. $P(H \mid I)$

29. $P(N \mid O)$

31. $P(I \text{ OR } N)$

33. $P(I)$

35. The likelihood that an event will occur given that another event has already occurred.

37. 1

39. the probability of landing on an even number or a multiple of three

41. $P(J) = .3$

43. $P(Q \text{ AND } R) = P(Q)P(R)$
$.1 = (.4)P(R)$
$P(R) = .25$

45. 0.376

47. $C \mid L$ means, given the person chosen is a Latino Californian, the person is a registered voter who prefers life in prison without parole for a person convicted of first degree murder.

49. L AND C is the event that the person chosen is a voter of the ethnicity in question who prefers life without parole over the death penalty for a person convicted of first degree murder.

51. .6492

53. No, because $P(L \text{ AND } C)$ does not equal 0.

55. $P(\text{musician is a male AND had private instruction}) = \frac{15}{130} = \frac{3}{26} = .12$

57. $P(\text{being a female musician AND learning music in school}) = \frac{38}{130} = \frac{19}{65} = .29$
$P(\text{being a female musician})P(\text{learning music in school}) = \left(\frac{72}{130}\right)\left(\frac{62}{130}\right) = \frac{4,464}{16,900} = \frac{1,116}{4,225} = .26$
No, they are not independent because $P(\text{being a female musician AND learning music in school})$ is not equal to $P(\text{being a female musician})P(\text{learning music in school})$.

58.

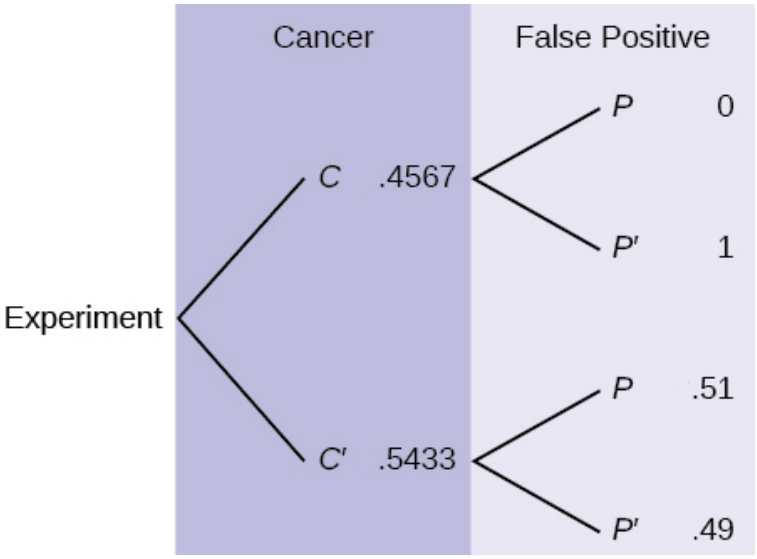


Figure 3.23

60. $\frac{35,065}{100,450}$

62. To pick one person from the study who is Japanese American AND uses the product 21 to 30 times a day means that the person has to meet both criteria: both Japanese American and uses the product 21 to 30 times a day. The sample space should include everyone in the study. The probability is $\frac{4,715}{100,450}$.

64. To pick one person from the study who is Japanese American given that person uses the product 21 to 30 times a day, means that the person must fulfill both criteria and the sample space is reduced to those who uses the product 21 to 30 times a day. The probability is $\frac{4715}{15,273}$.

67. a. You can't calculate the joint probability knowing the probability of both events occurring, which is not in the information given; the probabilities should be multiplied, not added; and probability is never greater than 100 percent
b. A home run by definition is a successful hit, so he has to have at least as many successful hits as home runs.

69. 0

71. .3571

73. .2142

75. Physician (83.7)

77. 83.7 − 79.6 = 4.1

79. *P*(Occupation < 81.3) = .5

81. a. The Forum Research surveyed 1,046 Torontonians.
b. 58 percent
c. 42 percent of 1,046 = 439 (rounding to the nearest integer)
d. .57
e. .60.

82. a. yes; *P*(getting a pork chop) = *P*(not getting a chicken breast)

b. getting a pork chop and getting a chicken breast
c. no

83. a. 20/40 = 1/2
b. 5/40 = 1/8
c. 39/40
d. 4/40 = 1/10
e. 33/40
f. 15/40 = 3/8
g. 0/40 = 0

84. Compute the probabilities.
a. 20/40 = 1/2
b. 8/40 = 1/5
c. 40/40 = 1
d. 16/40 = 2/5
e. 18/40 = 9/20
f. 40/40 = 1

85. a. {*G*1, *G*2, *G*3, *G*4, *G*5, *Y*1, *Y*2, *Y*3}
b. $\frac{5}{8}$
c. $\frac{2}{3}$
d. $\frac{2}{8}$
e. $\frac{6}{8}$
f. No, because *P*(*G* AND *E*) does not equal 0.

87.

NOTE

The coin toss is independent of the card picked first.

a. {(*G*,*H*) (*G*,*T*) (*B*,*H*) (*B*,*T*) (*R*,*H*) (*R*,*T*)}
b. *P*(*A*) = *P*(blue)*P*(head) = $\left(\frac{3}{10}\right)\left(\frac{1}{2}\right) = \frac{3}{20}$
c. Yes, *A* and *B* are mutually exclusive because they cannot happen at the same time; you cannot pick a card that is both blue and also (red or green). *P*(*A* AND *B*) = 0.
d. No, *A* and *C* are not mutually exclusive because they can occur at the same time. In fact, *C* includes all of the outcomes of *A*; if the card chosen is blue it is also (red or blue). *P*(*A* AND *C*) = *P*(*A*) = $\frac{3}{20.}$

89. a. *S* = {(*HHH*), (*HHT*), (*HTH*), (*HTT*), (*THH*), (*THT*), (*TTH*), (*TTT*)}
b. $\frac{4}{8}$
c. Yes, because if *A* has occurred, it is impossible to obtain two tails. In other words, *P*(*A* AND *B*) = 0.

91. a. If *Y* and *Z* are independent, then *P*(*Y* AND *Z*) = *P*(*Y*)*P*(*Z*), so *P*(*Y* OR *Z*) = *P*(*Y*) + *P*(*Z*) – *P*(*Y*)*P*(*Z*).
b. .5

93. iii i iv ii

95. a. *P*(*R*) = .44
b. *P*(*R*| *E*) = .56
c. *P*(*R*| *O*) = .31
d. No, whether the money is returned is not independent of which class the money was placed in. There are several ways to justify this mathematically, but one is that the money placed in economics classes is not

returned at the same overall rate; $P(R|E) \neq P(R)$.

e. No, this study definitely does not support that notion; in fact, it suggests the opposite. The money placed in the economics classrooms was returned at a higher rate than the money place in all classes collectively; $P(R|E) > P(R)$.

97. a. P(type O OR Rh–) = P(type O) + P(Rh–) – P(type O AND Rh–)

0.52 = 0.43 + 0.15 – P(type O AND Rh–); solve to find P(type O AND Rh–) = .06

6 percent of people have type O, Rh– blood

b. P(NOT(type O AND Rh–)) = 1 – P(type O AND Rh–) = 1 – .06 = .94

94 percent of people do not have type O, Rh– blood

99. a. Let C = be the event that the cookie contains chocolate. Let N = the event that the cookie contains nuts.
b. $P(C \text{ OR } N) = P(C) + P(N) - P(C \text{ AND } N)$ = .36 + .12 – .08 = .40
c. P(NEITHER chocolate NOR nuts) = 1 – $P(C \text{ OR } N)$ = 1 – .40 = .60

101. 0

103. $\frac{10}{67}$

105. $\frac{10}{34}$

107. d

110. b

112. a. $\frac{26}{106}$
b. $\frac{33}{106}$
c. $\frac{21}{106}$
d. $\left(\frac{26}{106}\right) + \left(\frac{33}{106}\right) - \left(\frac{21}{106}\right) = \left(\frac{38}{106}\right)$
e. $\frac{21}{33}$

114. a

117. a. $P(C)$ = .4567
b. not enough information
c. not enough information
d. no, because over half (0.51) of men have at least one false-positive text

119. a. $P(J \text{ OR } K) = P(J) + P(K) - P(J \text{ AND } K)$; .45 = .18 + .37 – $P(J \text{ AND } K)$; solve to find $P(J \text{ AND } K)$ = .10
b. P(NOT (J AND K)) = 1 – $P(J \text{ AND } K)$ = 1 – 010 = .90
c. P(NOT (J OR K)) = 1 – $P(J \text{ OR } K)$ = 1 – .45 = .55

120. a.

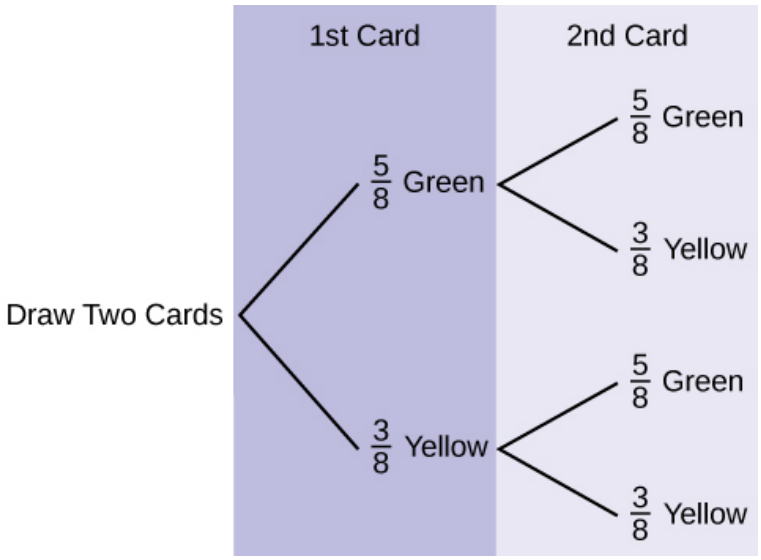


Figure 3.24

b. $P(GG) = \left(\frac{5}{8}\right)\left(\frac{5}{8}\right) = \frac{25}{64}$

c. $P(\text{at least one green}) = P(GG) + P(GY) + P(YG) = \frac{25}{64} + \frac{15}{64} + \frac{15}{64} = \frac{55}{64}$

d. $P(G \mid G) = \frac{5}{8}$

e. Yes, they are independent because the first card is placed back in the bag before the second card is drawn. The composition of cards in the bag remains the same from draw one to draw two.

122. a.

	<20	20–64	>64	Totals
Female	.0244	.3954	.0661	.486
Male	.0259	.4186	.0695	.514
Totals	.0503	.8140	.1356	1

Table 3.27

b. $P(F) = .486$

c. $P(>64 \mid F) = .1361$

d. $P(>64 \text{ and } F) = P(F)\,P(>64 \mid F) = (.486)(.1361) = .0661$

e. $P(>64 \mid F)$ is the percentage of female drivers who are 65 or older and $P(>64 \text{ and } F)$ is the percentage of drivers who are female and 65 or older.

f. $P(>64) = P(>64 \text{ and } F) + P(>64 \text{ and } M) = .1356$

g. No, being female and 65 or older are not mutually exclusive because they can occur at the same time P(>64 and F) = .0661.

124. a.

	Car, Truck or Van	Walk	Public Transportation	Other	Totals
Alone	.7318				
Not Alone	.1332				
Totals	.8650	.0390	.0530	.0430	1

Table 3.28

b. If we assume that all walkers are alone and that none from the other two groups travel alone (which is a big assumption) we have: *P*(Alone) = .7318 + .0390 = .7708.
c. Make the same assumptions as in (b) we have: (.7708)(1,000) = 771
d. (.1332)(1,000) = 133

126. The completed contingency table is as follows:

	Method A	Method B	Method C	Other	Totals
Female	0	70	136	49	**255**
Male	2,146	463	60	135	**2,804**
Totals	**2,146**	**533**	**196**	**184**	**3,059**

Table 3.29

a. $\frac{255}{3059}$
b. $\frac{196}{3059}$
c. $\frac{718}{3059}$
d. 0
e. $\frac{463}{3059}$
f. $\frac{136}{196}$
g.

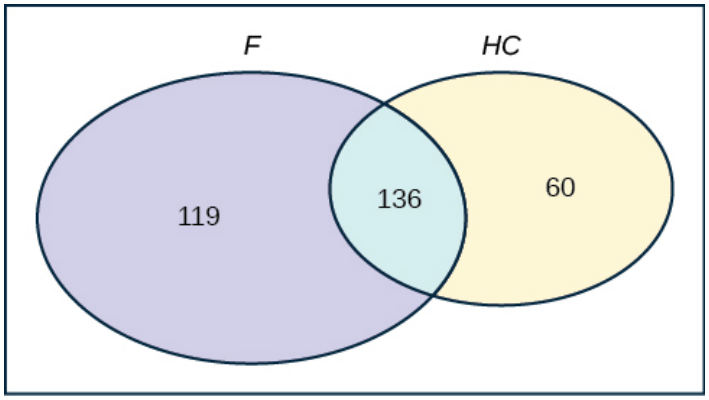


Figure 3.25

4 DISCRETE RANDOM VARIABLES

Figure 4.1 You can use probability and discrete random variables to calculate the likelihood of lightning striking the ground five times during a half-hour thunderstorm. (credit: Leszek Leszczynski)

Chapter Objectives

By the end of this chapter, the student should be able to do the following:

- Recognize and understand discrete probability distribution functions, in general.
- Calculate and interpret expected values.
- Recognize the binomial probability distribution and apply it appropriately.
- Recognize the poisson probability distribution and apply it appropriately.
- Recognize the geometric probability distribution and apply it appropriately.
- Recognize the hypergeometric probability distribution and apply it appropriately.
- Classify discrete word problems by their distributions.

Introduction

A student takes a 10-question, true-false quiz. Because the student had such a busy schedule, he or she could not study and guesses randomly at each answer. What is the probability of the student passing the test with at least a 70 percent?

Small companies might be interested in the number of long-distance phone calls their employees make during the peak time of the day. Suppose the average is 20 calls. What is the probability that the employees make more than 20 long-distance phone calls during the peak time?

These two examples illustrate two different types of probability problems involving discrete random variables. Recall that discrete data are data that you can count. A random variable is a variable whose values are numerical outcome of a probability experiment. We always describe a random variable in words and its values in numbers. The values of a random variable can vary with each repetition of an experiment.

Random Variable Notation

Uppercase letters such as X or Y denote a random variable. Lowercase letters like x or y denote the value of a random variable. If ***X* is a random variable, then *X* is written in words, and *x* is given as a number.**

The following are examples of random variables:

Example 1: Suppose a jar contains three marbles, one blue, one red, and one white. Randomly draw one marble from the jar. Let X = the possible number of red marbles to be drawn. The sample space for the drawing is red, white, and blue. Then, $x = 0,1$. If the marble we draw is red, then $x = 1$; otherwise, $x = 0$.

Example 2: Let X = the number of female children in a randomly selected family with only two kids. Here we are only interested in families with two kids, not families with one kid or more than two kids. The sample space for the genders of two-kid families is *MM, MF, FM, FF*. Here the first letter represents the gender of the older child and the second letter represents the gender of the younger child. F represents a female child and M represents a male child. For example, FM represents that the older child is a girl and the younger child is a boy, while MF represents that the older child is a boy and the younger child is a girl. Then, $x = 0,1,2$. A family has 0 female children if it has two boys (MM), a family has one female child if it has one boy and one girl (MF or FM), and a family has two female children if both kids are girls (FF).

Example 3: Let X = the number of heads you get when you toss three fair coins. The sample space for the toss of three fair coins is *TTT, THH, HTH, HHT, HTT, THT, TTH, HHH*. Here the first letter represents the result of the first toss, the second letter represents the result of the second toss, and the third letter represents the result of the third toss. T represents a tail and H represents a head. For example, THH means we get a tail in the first toss but a head in the second and third toss, while HHT means we get a head in the first and second toss but a tail in the third toss. Then, $x = 0, 1, 2, 3$. There are 0 heads if the result is TTT, one head if the result is THT, TTH, or HTT, two heads if the result is THH, HTH, or HHT, and three heads if the result is HHH.

COLLABORATIVE EXERCISE

Toss a coin 10 times and record the number of heads. After all members of the class have completed the experiment (tossed a coin 10 times and counted the number of heads), fill in Table 4.1. Let X = the number of heads in 10 tosses of the coin.

x	Frequency of x	Relative Frequency of x

Table 4.1

a. Which value(s) of x occurred most frequently?
b. If you tossed the coin 1,000 times, what values could x take on? Which value(s) of x do you think would occur most frequently?
c. What does the relative frequency column sum to?

4.1 Probability Distribution Function (PDF) for a Discrete Random Variable

There are two types of **random variables**, discrete random variables and continuous random variables. The values of a *discrete random variable* are countable, which means the values are obtained by counting. All random variables we discussed in previous examples are discrete random variables. We counted the number of red balls, the number of heads, or the number of female children to get the corresponding random variable values. The values of a *continuous random variable* are uncountable, which means the values are not obtained by counting. Instead, they are obtained by measuring. For example, let X = temperature of a randomly selected day in June in a city. The value of X can be 68°, 71.5°, 80.6°, or 90.32°. These values are obtained by measuring by a thermometer. Another example of a continuous random variable is the height of a randomly selected high school student. The value of this random variable can be 5'2", 6'1", or 5'8". Those values are obtained by measuring by a ruler.

A discrete **probability distribution function** has two characteristics:

1. Each probability is between zero and one, inclusive.

2. The sum of the probabilities is one.

EXAMPLE 4.1

A child psychologist is interested in the number of times a newborn baby's crying wakes its mother after midnight. For a random sample of 50 mothers, the following information was obtained. Let X = the number of times per week a newborn baby's crying wakes its mother after midnight. For this example, $x = 0, 1, 2, 3, 4, 5$.

$P(x)$ = probability that X takes on a value x.

x	$P(x)$
0	$P(x = 0) = \frac{2}{50}$
1	$P(x = 1) = \frac{11}{50}$
2	$P(x = 2) = \frac{23}{50}$
3	$P(x = 3) = \frac{9}{50}$
4	$P(x = 4) = \frac{4}{50}$
5	$P(x = 5) = \frac{1}{50}$

Table 4.2

X takes on the values 0, 1, 2, 3, 4, 5. This is a discrete PDF because we can count the number of values of x and also because of the following two reasons:

a. Each $P(x)$ is between zero and one, therefore inclusive
b. The sum of the probabilities is one, that is,

$$\frac{2}{50} + \frac{11}{50} + \frac{23}{50} + \frac{9}{50} + \frac{4}{50} + \frac{1}{50} = 1$$

TRY IT 4.1

A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a 12-hour shift. For a random sample of 50 patients, the following information was obtained. Let X = the number of times a patient rings the nurse during a 12-hour shift. For this exercise, $x = 0, 1, 2, 3, 4, 5$. $P(x)$ = the probability that X takes on value x. Why is this a discrete probability distribution function (two reasons)?

X	$P(x)$
0	$P(x = 0) = \frac{4}{50}$
1	$P(x = 1) = \frac{8}{50}$
2	$P(x = 2) = \frac{16}{50}$

Table 4.3

X	$P(x)$
3	$P(x = 3) = \frac{14}{50}$
4	$P(x = 4) = \frac{6}{50}$
5	$P(x = 5) = \frac{2}{50}$

Table 4.3

EXAMPLE 4.2

Suppose Nancy has classes **three days** a week. She attends classes three days a week **80 percent** of the time, **two days 15 percent** of the time, **one day 4 percent** of the time, and **no days 1 percent** of the time. Suppose one week is randomly selected.

Problem

Describe the random variable in words. Let X = the number of days Nancy _____.

Solution

a. Let X = the number of days Nancy attends class per week.

Problem

b. In this example, what are possible values of X?

Solution

b. 0, 1, 2, and 3

Problem

c. Suppose one week is randomly chosen. Construct a probability distribution table (called a PDF table) like the one in Example 4.1. The table should have two columns labeled x and $P(x)$.

Solution

c.

x	$P(x)$
0	.01
1	.04
2	.15
3	.80

Table 4.4

The sum of the $P(x)$ column is 0.01+0.04+0.15+0.80 = 1.00.

TRY IT 4.2

Jeremiah has basketball practice two days a week. 90 percent of the time, he attends both practices. Eight percent of the time, he attends one practice. Two percent of the time, he does not attend either practice. What is *X* and what values does it take on?

4.2 Mean or Expected Value and Standard Deviation

The **expected value** of a discrete random variable *X*, symbolized as *E(X)*, is often referred to as the *long-term average or* ***mean*** (symbolized as μ). This means that over the long term of doing an experiment over and over, you would expect this average. For example, let *X* = the number of heads you get when you toss three fair coins. If you repeat this experiment (toss three fair coins) a large number of times, the expected value of *X* is the number of heads you expect to get for each three tosses on average.

NOTE

To find the expected value, *E(X)*, or mean μ of a discrete random variable *X*, simply multiply each value of the random variable by its probability and add the products. The formula is given as $E(X) = \mu = \sum xP(x)$.

Here *x* represents values of the random variable *X*, *P*(*x*) represents the corresponding probability, and symbol $\sum$ represents the sum of all products *xP*(*x*). Here we use symbol μ for the mean because it is a parameter. It represents the mean of a population.

EXAMPLE 4.3

A men's soccer team plays soccer zero, one, or two days a week. The probability that they play zero days is .2, the probability that they play one day is .5, and the probability that they play two days is .3. Find the long-term average or expected value, μ, of the number of days per week the men's soccer team plays soccer.

To do the problem, first let the random variable *X* = the number of days the men's soccer team plays soccer per week. *X* takes on the values 0, 1, 2. Construct a PDF table adding a column *x*P(x)*, the product of the value *x* with the corresponding probability *P(x)*. In this column, you will multiply each *x* value by its probability.

x	*P*(*x*)	*x*P*(*x*)
0	.2	(0)(.2) = 0
1	.5	(1)(.5) = .5
2	.3	(2)(.3) = .6

Table 4.5 Expected Value Table This table is called an expected value table. The table helps you calculate the expected value or long-term average.

Add the last column $x * P(x)$ to get the expected value/mean of the random variable *X*.

$$E(X) = \mu = \sum xP(x) = 0 + .5 + .6 = 1.1$$

The expected value/mean is 1.1. The men's soccer team would, on the average, expect to play soccer 1.1 days per week.

The number 1.1 is the long-term average or expected value if the men's soccer team plays soccer week after week after week.

As you learned in Chapter 3, if you toss a fair coin, the probability that the result is heads is 0.5. This probability is a theoretical probability, which is what we expect to happen. This probability does not describe the short-term results of an experiment. If you flip a coin two times, the probability does not tell you that these flips will result in one head and one tail. Even if you flip a coin 10 times or 100 times, the probability does not tell you that you will get half tails and half heads. The probability gives information about what can be expected in the long term. To demonstrate this, Karl Pearson once tossed a fair coin 24,000 times! He recorded the results of each toss, obtaining heads 12,012 times. The relative frequency of heads is 12,012/24,000 = .5005, which is very close to the theoretical probability .5. In his experiment, Pearson illustrated the law of large numbers.

The law of large numbers states that, as the number of trials in a probability experiment increases, the difference between the theoretical probability of an event and the relative frequency approaches zero (the theoretical probability and the relative frequency get closer and closer together). The relative frequency is also called the experimental probability, a term that means what actually happens.

In the next example, we will demonstrate how to find the expected value and **standard deviation of a discrete probability distribution** by using relative frequency.

Like data, probability distributions have variances and standard deviations. The variance of a probability distribution is symbolized as σ^2 and the standard deviation of a probability distribution is symbolized as σ. Both are parameters since they summarize information about a population. To find the variance σ^2 of a discrete probability distribution, find each deviation from its expected value, square it, multiply it by its probability, and add the products. To find the standard deviation σ of a probability distribution, simply take the square root of variance σ^2. The formulas are given as below.

NOTE

The formula of the variance σ^2 of a discrete random variable X is

$$\sigma^2 = \sum (x-\mu)^2 P(x). \quad \textbf{4.1}$$

Here x represents values of the random variable X, μ is the mean of X, $P(x)$ represents the corresponding probability, and symbol $\sum$ represents the sum of all products $(x-\mu)^2 P(x)$.

To find the standard deviation, σ, of a discrete random variable X, simply take the square root of the variance σ^2.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum (x-\mu)^2 P(x)}$$

EXAMPLE 4.4

A researcher conducted a study to investigate how a newborn baby's crying after midnight affects the sleep of the baby's mother. The researcher randomly selected 50 new mothers and asked how many times they were awakened by their newborn baby's crying after midnight per week. Two mothers were awakened zero times, 11 mothers were awakened one time, 23 mothers were awakened two times, nine mothers were awakened three times, four mothers were awakened four times, and one mother was awakened five times. Find the expected value of the number of times a newborn baby's crying wakes its mother after midnight per week. Calculate the standard deviation of the variable as well.

To do the problem, first let the random variable X = the number of times a mother is awakened by her newborn's crying after midnight per week. X takes on the values 0, 1, 2, 3, 4, 5. Construct a PDF table as below. The column of $P(x)$ gives the experimental probability of each x value. We will use the relative frequency to get the probability. For example, the probability that a mother wakes up zero times is $\frac{2}{50}$ since there are two mothers out of 50 who were awakened zero times. The third column of the table is the product of a value and its probability, $xP(x)$.

x	$P(x)$	$xP(x)$
0	$P(x = 0) = \frac{2}{50}$	$(0)\left(\frac{2}{50}\right) = 0$
1	$P(x = 1) = \frac{11}{50}$	$(1)\left(\frac{11}{50}\right) = \frac{11}{50}$
2	$P(x = 2) = \frac{23}{50}$	$(2)\left(\frac{23}{50}\right) = \frac{46}{50}$
3	$P(x = 3) = \frac{9}{50}$	$(3)\left(\frac{9}{50}\right) = \frac{27}{50}$
4	$P(x = 4) = \frac{4}{50}$	$(4)\left(\frac{4}{50}\right) = \frac{16}{50}$
5	$P(x = 5) = \frac{1}{50}$	$(5)\left(\frac{1}{50}\right) = \frac{5}{50}$

Table 4.6

We then add all the products in the third column to get the mean/expected value of X.

$$E(X) = \mu = \sum xP(x) = 0 + \frac{11}{50} + \frac{46}{50} + \frac{27}{50} + \frac{16}{50} + \frac{5}{50} = \frac{105}{50} = 2.1 \quad \textbf{4.2}$$

Therefore, we expect a newborn to wake its mother after midnight 2.1 times per week, on the average.

To calculate the standard deviation σ, we add the fourth column $(x\text{-}\mu)^2$ and the fifth column $(x-\mu)^2 \bullet P(x)$ to get the following table:

x	$P(x)$	$xP(x)$	$(x\text{-}\mu)^2$	$(x\text{-}\mu)^2 \bullet P(x)$
0	$P(x = 0) = \frac{2}{50}$	$(0)\left(\frac{2}{50}\right) = 0$	$(0 - 2.1)^2 = 4.41$	$4.41 \bullet \frac{2}{50} = .1764$
1	$P(x = 1) = \frac{11}{50}$	$(1)\left(\frac{11}{50}\right) = \frac{11}{50}$	$(1 - 2.1)^2 = 1.21$	$1.21 \bullet \frac{11}{50} = .2662$
2	$P(x = 2) = \frac{23}{50}$	$(2)\left(\frac{23}{50}\right) = \frac{46}{50}$	$(2 - 2.1)^2 = .01$	$.01 \bullet \frac{23}{50} = .0046$
3	$P(x = 3) = \frac{9}{50}$	$(3)\left(\frac{9}{50}\right) = \frac{27}{50}$	$(3 - 2.1)^2 = .81$	$.81 \bullet \frac{9}{50} = .1458$
4	$P(x = 4) = \frac{4}{50}$	$(4)\left(\frac{4}{50}\right) = \frac{16}{50}$	$(4 - 2.1)^2 = 3.61$	$3.61 \bullet \frac{4}{50} = .2888$
5	$P(x = 5) = \frac{1}{50}$	$(5)\left(\frac{1}{50}\right) = \frac{5}{50}$	$(5 - 2.1)^2 = 8.41$	$8.41 \bullet \frac{1}{50} = .1682$

Table 4.7

We then add all the products in the 5th column to get the variance of X.

$$\sigma^2 = .1764 + .2662 + .0046 + .1458 + .2888 + .1682 = 1.05 \quad \textbf{4.3}$$

To get the standard deviation σ, we simply take the square root of variance σ^2.

$$\sigma = \sqrt{\sigma^2} = \sqrt{1.05} \approx 1.0247 \quad \textbf{4.4}$$

TRY IT 4.4

A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a 12-hour shift. For a random sample of 50 patients, the following information was obtained. What is the expected value?

x	$P(x)$
0	$P(x = 0) = \frac{4}{50}$
1	$P(x = 1) = \frac{8}{50}$
2	$P(x = 2) = \frac{16}{50}$
3	$P(x = 3) = \frac{14}{50}$
4	$P(x = 4) = \frac{6}{50}$
5	$P(x = 5) = \frac{2}{50}$

Table 4.8

EXAMPLE 4.5

Suppose you play a game of chance in which five numbers are chosen from 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. A computer randomly selects five numbers from zero to nine with replacement. You pay $2 to play and could profit $100,000 if you match all five numbers in order (you get your $2 back plus $100,000). Over the long term, what is your *expected* profit of playing the game?

To do this problem, set up a PDF table for the amount of money you can profit.

Let X = the amount of money you profit. If your five numbers match in order, you will win the game and will get your $2 back plus $100,000. That means your profit is $100,000. If your five numbers do not match in order, you will lose the game and lose your $2. That means your profit is -$2. Therefore, X takes on the values $100,000 and –$2. That is the second column x in the PDF table below.

To win, you must get all five numbers correct, in order. The probability of choosing the correct first number is $\frac{1}{10}$ because there are 10 numbers (from zero to nine) and only one of them is correct. The probability of choosing the correct second number is also $\frac{1}{10}$ because the selection is done with replacement and there are still 10 numbers (from zero to nine) for you to choose. Due to the same reason, the probability of choosing the correct third number, the correct fourth number, and the correct fifth number are also $\frac{1}{10}$. The selection of one number does not affect the selection of another number. That means the five selections are independent. The probability of choosing all five correct numbers and in order is equal to the product of the probabilities of choosing each number correctly.

$$\text{P(choosing all five numbers correctly)} \bullet \text{P(choosing 1st number correctly)} \bullet \text{P(choosing 2nd number correctly)} \bullet \text{P(choosing 5th number correctly)}$$
$$= (\tfrac{1}{10}) \bullet (\tfrac{1}{10}) \bullet (\tfrac{1}{10}) \bullet (\tfrac{1}{10}) \bullet (\tfrac{1}{10})$$
$$= .00001$$

Therefore, the probability of winning is .00001 and the probability of losing is 1 − .00001 = .99999. That is how we get the third column $P(x)$ in the PDF table below.

To get the fourth column $xP(x)$ in the table, we simply multiply the value x with the corresponding probability $P(x)$.

The PDF table is as follows:

	x	$P(x)$	$x*P(x)$
Loss	–2	.99999	(–2)(.99999) = –1.99998
Profit	100,000	.00001	(100000)(.00001) = 1

Table 4.9

We then add all the products in the last column to get the mean/expected value of X.

$$E(X) = \mu = \sum xP(x) = -1.99998 + 1 = -.9998.$$

Since –.99998 is about –1, you would, on average, expect to lose approximately $1 for each game you play. However, each time you play, you either lose $2 or profit $100,000. The $1 is the average or expected **loss** per game after playing this game over and over.

TRY IT 4.5

You are playing a game of chance in which four cards are drawn from a standard deck of 52 cards. You guess the suit of each card before it is drawn. The cards are replaced in the deck on each draw. You pay $1 to play. If you guess the right suit every time, you get your money back and $256. What is your expected profit of playing the game over the long term?

EXAMPLE 4.6

Suppose you play a game with a biased coin. You play each game by tossing the coin once. $P(\text{heads}) = \frac{2}{3}$ and $P(\text{tails}) = \frac{1}{3}$. If you toss a head, you pay $6. If you toss a tail, you win $10. If you play this game many times, will you come out ahead?

Problem

a. Define a random variable X.

Solution

a. X = amount of profit

Problem

b. Complete the following expected value table.

	x	___	___
WIN	10	$\frac{1}{3}$	___
LOSE	___	___	$\frac{-12}{3}$

Table 4.10

Solution

b.

	x	$P(x)$	$xP(x)$
WIN	10	$\frac{1}{3}$	$\frac{10}{3}$
LOSE	–6	$\frac{2}{3}$	$\frac{-12}{3}$

Table 4.11

Problem

c. What is the expected value, μ? Do you come out ahead?

Solution

c. Add the last column of the table. The expected value $E(X) = \mu = \frac{10}{3} + \left(-\frac{12}{3}\right) = -\frac{2}{3} \approx -.67$. You lose, on average, about 67 cents each time you play the game, so you do not come out ahead.

TRY IT 4.6

Suppose you play a game with a spinner. You play each game by spinning the spinner once. P(red) = $\frac{2}{5}$, P(blue) = $\frac{2}{5}$, and P(green) = $\frac{1}{5}$. If you land on red, you pay \$10. If you land on blue, you don't pay or win anything. If you land on green, you win \$10. Complete the following expected value table.

	x	$P(x)$	
Red			$-\frac{20}{5}$
Blue		$\frac{2}{5}$	
Green	10		

Table 4.12

Generally for probability distributions, we use a calculator or a computer to calculate μ and σ to reduce rounding errors. For some probability distributions, there are shortcut formulas for calculating μ and σ.

EXAMPLE 4.7

Problem

Toss a fair, six-sided die twice. Let X = the number of faces that show an even number. Construct a table like Table 4.12 and calculate the mean μ and standard deviation σ of X.

Solution

Tossing one fair six-sided die twice has the same sample space as tossing two fair six-sided dice. The sample space has 36 outcomes.

(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Table 4.13

Use the sample space to complete the following table.

x	$P(x)$	$xP(x)$	$(x - \mu)^2 \cdot P(x)$
0	$\frac{9}{36}$	0	$(0 - 1)^2 \cdot \frac{9}{36} = \frac{9}{36}$
1	$\frac{18}{36}$	$\frac{18}{36}$	$(1 - 1)^2 \cdot \frac{18}{36} = 0$
2	$\frac{9}{36}$	$\frac{18}{36}$	$(2 - 1)^2 \cdot \frac{9}{36} = \frac{9}{36}$

Table 4.14 Calculating μ and σ.

Add the values in the third column to find the expected value: $\mu = \frac{36}{36} = 1$. Use this value to complete the fourth column.

Add the values in the fourth column and take the square root of the sum: $\sigma = \sqrt{\frac{18}{36}} \approx .7071$.

Some of the more common discrete probability functions are binomial, geometric, hypergeometric, and Poisson. Most elementary courses do not cover the geometric, hypergeometric, and Poisson. Your instructor will let you know if he or she wishes to cover these distributions.

A probability distribution function is a pattern. You try to fit a probability problem into a **pattern** or distribution in order to perform the necessary calculations. These distributions are tools to make solving probability problems easier. Each distribution has its own special characteristics. Learning the characteristics enables you to distinguish among the different distributions.

4.3 Binomial Distribution (Optional)

There are three characteristics of a **binomial experiment**:

1. There are a fixed number of trials. Think of trials as repetitions of an experiment. The letter n denotes the number of trials.
2. There are only two possible outcomes, called *success* and *failure*, for each trial. The outcome that we are measuring is defined as a *success*, while the other outcome is defined as a *failure*. The letter p denotes the probability of a success on one trial, and q denotes the probability of a failure on one trial. $p + q = 1$.
3. The n trials are independent and are repeated using identical conditions. Because the n trials are independent, the outcome of one trial does not help in predicting the outcome of another trial. Another way of saying this is that for each individual trial, the probability, p, of a success and probability, q, of a failure remain the same. Let us look at several examples of a binomial experiment.

   Example 1: Toss a fair coin once and record the result.

   This is a binomial experiment since it meets all three characteristics. The number of trials $n = 1$. There are only two

outcomes, a head or a tail, of each trial. We can define a head as a success if we are measuring number of heads. For a fair coin, the probabilities of getting head or tail are both .5. So, $p = q - .5$. Both p and q remain the same from trial to trial. This experiment is also called a **Bernoulli trial**, named after Jacob Bernoulli who, in the late 1600s, studied such trials extensively. Any experiment that has characteristics two and three and where $n = 1$ is called a **Bernoulli trial**. A binomial experiment takes place when the number of successes is counted in one or more Bernoulli trials.

Example 2: Randomly guess a multiple choice question has A, B, C and D four options.

This is a binomial experiment since it meets all three characteristics. The number of trials $n = 1$. There are only two outcomes, guess correctly or guess wrong, of each trial. We can define guess correctly as a success. For a random guess (you have no clue at all), the probability of guessing correct should be $\frac{1}{4}$ because there are four options and only one option is correct. So, and $p = \frac{1}{4}$ and $q = 1 - p = 1 - \frac{1}{4} = \frac{3}{4}$. Both p and q remain the same from trial to trial. This experiment is also a Bernoulli trial. It meets the characteristics two and three and $n = 1$.

Example 3: Toss a fair coin five times and record the result.

This is a binomial experiment since it meets all three characteristics. The number of trials $n = 5$. There are only two outcomes, head or tail, of each trial. If we define head as a success, then $p = q = 0.5$. Both p and q remain the same for each trial. Since n = 5, this experiment is not a Bernoulli trial although it meets the characteristics two and three.

Example 4: Randomly guess 10 multiple choice questions in an exam. Each question has A, B, C and D four options.

This is a binomial experiment since it meets all three characteristics. The number of trials $n = 10$. There are only two outcomes, guess correctly or guess wrong, of each trial. We can define guess correctly as a success. As we explained in example 2, $p = \frac{1}{4}$ and $q = 1 - p = 1 - \frac{1}{4} = \frac{3}{4}$. Both p and q remain the same for each guess. Since $n = 10$, this experiment is not a Bernoulli trial.

The next two experiments are not binomial experiments.

Example 5: Randomly select two balls from a jar with five red balls and five blue balls without replacement. This means we select the first ball, and then without returning the selected ball into the jar, we will select the second ball.

This is not a binomial experiment since the third characteristic is not met. The number of trials $n = 2$. There are only two outcomes, a red ball or a blue ball, of each trial. If we define selecting a red ball as a success, then selecting a blue ball is a failure. The probability of getting the first ball red is $\frac{5}{10}$ since there are five red balls out of 10 balls. So, $p = \frac{5}{10}$ and $q = 1 - p = 1 - \frac{5}{10} = \frac{5}{10}$. However, p and q do not remain the same for the second trial. If the first ball selected is red, then the probability of getting the second ball red is $\frac{4}{9}$ since there are only four red balls out of nine balls. But if the first ball selected is blue, then the probability of getting the second ball red is $\frac{5}{9}$ since there are still five red balls out of nine balls.

Example 6: Toss a fair coin until a head appears.

This is not a binomial experiment since the first characteristic is not met. The number of trials n is not fixed. n could be 1 if a head appears from the first toss. n could be 2 if the first toss is a tail and the second toss is a head. So on and so forth.

More examples of binomial and non-binomial experiments will be discussed in this section later.

The outcomes of a binomial experiment fit a **binomial probability distribution**. The random variable X = the number of successes obtained in the n independent trials.

There are shortcut formulas for calculating mean μ, variance σ^2, and standard deviation σ of a binomial probability distribution. The formulas are given as below. The deriving of these formulas will not be discussed in this book.

$$\mu = np, \sigma^2 = npq, \sigma = \sqrt{npq}.$$

Here n is the number of trials, p is the probability of a success, and q is the probability of a failure.

EXAMPLE 4.8

At ABC High School, the withdrawal rate from an elementary physics course is 30 percent for any given term. This implies that, for any given term, 70 percent of the students stay in the class for the entire term. The random variable X = the number of students who withdraw from the randomly selected elementary physics class. Since we are measuring the number of students who withdrew, a *success* is defined as an individual who withdrew.

TRY IT 4.8

The state health board is concerned about the amount of fruit available in school lunches. Forty-eight percent of schools in the state offer fruit in their lunches every day. This implies that 52 percent do not. What would a *success* be in this case?

EXAMPLE 4.9

Suppose you play a game that you can only either win or lose. The probability that you win any game is 55 percent, and the probability that you lose is 45 percent. Each game you play is independent. If you play the game 20 times, write the function that describes the probability that you win 15 of the 20 times. Here, if you define X as the number of wins, then X takes on the values 0, 1, 2, 3, . . ., 20. The probability of a success is $p = 0.55$. The probability of a failure is $q = .45$. The number of trials is $n = 20$. The probability question can be stated mathematically as $P(x = 15)$. If you define X as the number of losses, then a *success* is defined as a loss and a *failure* is defined as a win. A *success* does not necessarily represent a good outcome. It is simply the outcome that you are measuring. X still takes on the values of 0, 1, 2, 3, . . ., 20. The probability of a success is $p = .45$. The probability of a failure is $q = .55$.

TRY IT 4.9

A trainer is teaching a dolphin to do tricks. The probability that the dolphin successfully performs the trick is 35 percent, and the probability that the dolphin does not successfully perform the trick is 65 percent. Out of 20 attempts, you want to find the probability that the dolphin succeeds 12 times. State the probability question mathematically.

EXAMPLE 4.10

Problem

A fair coin is flipped 15 times. Each flip is independent. What is the probability of getting more than 10 heads? Let X = the number of heads in 15 flips of the fair coin. X takes on the values 0, 1, 2, 3, . . ., 15. Since the coin is fair, $p = .5$ and $q = .5$. The number of trials $n = 15$. State the probability question mathematically.

Solution

$P(x > 10)$

TRY IT 4.10

A fair, six-sided die is rolled 10 times. Each roll is independent. You want to find the probability of rolling a one more than three times. State the probability question mathematically.

EXAMPLE 4.11

Approximately 70 percent of statistics students do their homework in time for it to be collected and graded. Each student does homework independently. In a statistics class of 50 students, what is the probability that at least 40 will do their homework on time? Students are selected randomly.

Problem

a. This is a binomial problem because there is only a success or a _____, there are a fixed number of trials, and the probability of a success is .70 for each trial.

Solution

a. failure

Problem

b. If we are interested in the number of students who do their homework on time, then how do we define X?

Solution

b. X = the number of statistics students who do their homework on time

Problem

c. What values does x take on?

Solution

c. 0, 1, 2, . . ., 50

Problem

d. What is a *failure*, in words?

Solution

d. Failure is defined as a student who does not complete his or her homework on time.

The probability of a success is $p = .70$. The number of trials is $n = 50$.

Problem

e. If $p + q = 1$, then what is q?

Solution

e. $q = .30$

Problem

f. The words *at least* translate as what kind of inequality for the probability question $P(x$ ___ $40)$?

Solution

f. greater than or equal to ($\geq$)
The probability question is $P(x \geq 40)$.

TRY IT 4.11

Sixty-five percent of people pass the state driver's exam on the first try. A group of 50 individuals who have taken the driver's exam is randomly selected. Give two reasons why this is a binomial problem.

Notation for the Binomial: B = Binomial Probability Distribution Function

$X \sim B(n, p)$

Read this as *X is a random variable with a binomial distribution*. The parameters are *n* and *p*: *n* = number of trials, *p* = probability of a success on each trial.

EXAMPLE 4.12

It has been stated that about 41 percent of adult workers have a high school diploma but do not pursue any further education. If 20 adult workers are randomly selected, find the probability that at most 12 of them have a high school diploma but do not pursue any further education. How many adult workers do you expect to have a high school diploma but do not pursue any further education?

Let *X* = the number of workers who have a high school diploma but do not pursue any further education.

X takes on the values 0, 1, 2, . . ., 20 where $n = 20$, $p = .41$, and $q = 1 - .41 = .59$. $X \sim B(20, .41)$

Find $P(x \leq 12)$. There is a formula to define the probability of a binomial distribution $P(x)$. We can use the formula to find $P(x \leq 12)$. But the calculation is tedious and time consuming, and people usually use a graphing calculator, software, or binomial table to get the answer. Use a graphing calculator, you can get $P(x \leq 12) = .9738$. The instruction of TI-83, 83+, 84, 84+ is given below.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Go into 2nd DISTR. The syntax for the instructions are as follows:

To calculate the probability of a value $P(x = value)$: **use binompdf(*n, p*, number)**. Here binompdf represents binomial probability density function. It is used to find the probability that a binomial random variable is equal to an exact value. n is the number of trials, p is the probability of a success, and number is the value. If *number* is left out, which means use **binompdf(n, p)**, then all the probabilities $P(x = 0), P(x = 1), \ldots, P(x = n)$ will be calculated.

To calculate the cumulative probability $P(x \leq value)$: **use binomcdf(*n, p*, number)**. Here binomcdf represents binomial cumulative distribution function. It is used to determine the probability of *at most* type of problem, the probability that a binomial random variable is less than or equal to a value. n is the number of trials, p is the probability of a success, and number is the value. If *number* is left out, all the cumulative probabilities $P(x \leq 0), P(x \leq 1), \ldots, P(x \leq n)$ will be calculated.

To calculate the cumulative probability $P(x \geq value)$: **use 1 - binomcdf(*n, p*, number)**. *n* is the number of trials, *p* is the probability of a success, and number is the value. TI calculators do not have a built-in function to find the probability that a binomial random variable is greater than a value. However, we can use the fact that $P(x > value) = 1 - P(x \leq value)$ to find the answer.

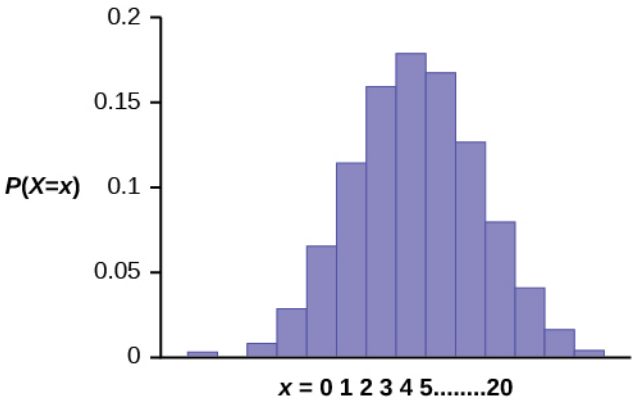
For this problem: After you are in 2nd DISTR, arrow down to binomcdf. Press ENTER. Enter 20,.41,12). The result is *P*(*x* ≤ 12) = .9738.

NOTE

If you want to find $P(x = 12)$, use the pdf (binompdf). If you want to find $P(x > 12)$, use 1 – binomcdf(20,.41,12).

The probability that at most 12 workers have a high school diploma but do not pursue any further education is .9738.

The graph of $X \sim B(20, .41)$ is as follows.



The previous graph is called a probability distribution histogram. It is made of a series of vertical bars. The *x*-axis of each bar is the value of X = the number of workers who have only a high school diploma, and the height of that bar is the probability of that value occurring.

The number of adult workers that you expect to have a high school diploma but not pursue any further education is the mean, $\mu = np = (20)(.41) = 8.2$.

The formula for the variance is $\sigma^2 = npq$. The standard deviation is $\sigma = \sqrt{npq}$.

$\sigma = \sqrt{(20)(.41)(.59)} = 2.20$.

The following is the interpretation of the mean $\mu = 8.2$ and standard deviation $\sigma = 2.20$:

If you randomly select 20 adult workers, and do that over and over, you expect around eight adult workers out of 20 to have a high school diploma but do not pursue any further education on average. And you expect that to vary by about two workers on average.

TRY IT 4.12

About 32 percent of students participate in a community volunteer program outside of school. If 30 students are selected at random, find the probability that at most 14 of them participate in a community volunteer program outside of school. Use the TI-83+ or TI-84 calculator to find the answer.

EXAMPLE 4.13

Problem

A store releases a 560-page art supply catalog. Eight of the pages feature signature artists. Suppose we randomly sample 100 pages. Let X = the number of pages that feature signature artists.

a. What values does x take on?
b. What is the probability distribution? Find the following probabilities:
   i. the probability that two pages feature signature artists
   ii. the probability that at most six pages feature signature artists
   iii. the probability that more than three pages feature signature artists
c. Using the formulas, calculate the (i) mean and (ii) standard deviation.

Solution

a. $x = 0, 1, 2, 3, 4, 5, 6, 7, 8$
b. This is a binomial experiment since all three characteristics are met. Each page is a trial. Since we sample 100 pages, the number of trials is $n = 100$. For each page, there are two possible outcomes, features signature artists or does not feature signature artists. Since we are measuring the number of pages that feature signature artists, a page that features signature artists is defined as a success and a page that does not feature signature artists is defined as

a failure. There are 8 out of 560 pages that feature signature artists. Therefore the probability of a success $p = \frac{8}{560}$ and the probability of a failure $q = 1 - p = 1 - \frac{8}{560} = \frac{552}{560}$.

Both p and q remain the same for each page. Therefore, X is a binomial random variable, and it can be written as $X \sim B\left(100, \frac{8}{560}\right)$.

We can use a graphing calculator to answer Parts i to iii.

i. $P(x = 2)$ = binompdf$\left(100, \frac{8}{560}, 2\right)$ = .2466
ii. $P(x \leq 6)$ = binomcdf$\left(100, \frac{8}{560}, 6\right)$ = .9994
iii. $P(x > 3) = 1 - P(x \leq 3)$ = 1 – binomcdf$\left(100, \frac{8}{560}, 3\right)$ = 1 – .9443 = .0557

c. i. mean = np = (100)$\left(\frac{8}{560}\right) = \frac{800}{560} \approx 1.4286$
ii. standard deviation = $\sqrt{npq} = \sqrt{(100)\left(\frac{8}{560}\right)\left(\frac{552}{560}\right)} \approx 1.1867$

TRY IT 4.13

According to a poll, 60 percent of American adults prefer saving over spending. Let X = the number of American adults out of a random sample of 50 who prefer saving to spending.

a. What is the probability distribution for X?
b. Use your calculator to find the following probabilities:
   i. The probability that 25 adults in the sample prefer saving over spending
   ii. The probability that at most 20 adults prefer saving
   iii. The probability that more than 30 adults prefer saving
c. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.

EXAMPLE 4.14

The lifetime risk of developing a specific disease is about 1 in 78 (1.28 percent). Suppose we randomly sample 200 people. Let X = the number of people who will develop the disease.

Problem

a. What is the probability distribution for X?
b. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.
c. Use your calculator to find the probability that at most eight people develop the disease.
d. Is it more likely that five or six people will develop the disease? Justify your answer numerically.

Solution

a. This is a binomial experiment since all three characteristics are met. Each person is a trial. Since we sample 200 people, the number of trials is n = 200. For each person, there are two possible outcomes: will develop the disease or not. Since we are measuring the number of people who will develop the disease, a person who will develop the disease is defined as a success and a person who will not develop the disease is defined as a failure. The risk of developing the disease is 1.28 percent. Therefore the probability of a success, $p = 1.28$ percent, $.0128$, and the probability of a failure, $q = 1 - p = 1 - .0128 = .9872$. Both p and q remain the same for each person. Therefore, X is a binomial random variable and it can be written as $X \sim B(200, .0128)$.

We can use a graphing calculator to answer Questions c and d.

b. i. Mean = np = 200(.0128) = 2.56
ii. Standard Deviation = $\sqrt{npq} = \sqrt{(200)(0.128)(.9872)} \approx 1.5897$

c. Using the TI-83, 83+, 84 calculator with instructions as provided in Example 4.12:
$P(x \leq 8)$ = binomcdf(200, .0128, 8) = .9988

d. $P(x = 5)$ = binompdf(200, .0128, 5) = .0707
$P(x = 6)$ = binompdf(200, .0128, 6) = .0298
So $P(x = 5) > P(x = 6)$; it is more likely that five people will develop the disease than six.

TRY IT 4.14

During the 2013 regular basketball season, a player had the highest field goal completion rate in the league. This player scored with 61.3 percent of his shots. Suppose you choose a random sample of 80 shots made by this player during the 2013 season. Let X = the number of shots that scored points.

a. What is the probability distribution for X?
b. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.
c. Use your calculator to find the probability that this player scored with 60 of these shots.
d. Find the probability that this player scored with more than 50 of these shots.

EXAMPLE 4.15

The following example illustrates a problem that is *not* binomial. It violates the condition of independence. ABC High School has a student advisory committee made up of 10 staff members and six students. The committee wishes to choose a chairperson and a recorder. What is the probability that the chairperson and recorder are both students? The names of all committee members are put into a box, and two names are drawn *without replacement*. The first name drawn determines the chairperson and the second name the recorder. There are two trials. However, the trials are not independent because the outcome of the first trial affects the outcome of the second trial. The probability of a student on the first draw is $\frac{6}{16}$ because there are six students out of 16 members (10 staff members + six students). If the first draw selects a student, then the probability of a student on the second draw is $\frac{5}{16}$ because there are only five students out of 15 members. If the first draw selects a staff member, then the probability of a student on the second draw is $\frac{6}{15}$ because there are still six students out of 15 members. The probability of drawing a student's name changes for each of the trials and, therefore, violates the condition of independence.

TRY IT 4.15

A lacrosse team is selecting a captain. The names of all the seniors are put into a hat, and the first three that are drawn will be the captains. The names are not replaced once they are drawn (one person cannot be two captains). You want to see if the captains all play the same position. State whether this problem is binomial or not and state why.

4.4 Geometric Distribution (Optional)

There are three main characteristics of a **geometric experiment**:

1. Repeating independent Bernoulli trials until a success is obtained. Recall that a Bernoulli trial is a binomial experiment with number of trials $n = 1$. In other words, you keep repeating what you are doing until the first success. Then you stop. For example, you throw a dart at a bull's-eye until you hit the bull's-eye. The first time you hit the bull's-eye is a *success* so you stop throwing the dart. It might take six tries until you hit the bull's-eye. You can think of the trials as failure, failure, failure, failure, failure, success, **stop**.
2. In theory, the number of trials could go on forever. There must be at least one trial.
3. The probability, p, of a success and the probability, q, of a failure do not change from trial to trial. $p + q = 1$ and $q = 1 - p$. For example, the probability of rolling a three when you throw one fair die is $\frac{1}{6}$. This is true no matter how many times you roll the die. Suppose you want to know the probability of getting the first three on the fifth roll. On rolls one through four, you do not get a face with a three. The probability for each of the rolls is $q = \frac{5}{6}$, the probability of a failure. The probability of getting a three on the fifth roll is $\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{1}{6}\right) = .0804$.

X = the number of independent trials until the first success.

p = the probability of a success, $q = 1 - p$ = the probability of a failure.

There are shortcut formulas for calculating mean μ, variance σ^2, and standard deviation σ of a geometric probability distribution. The formulas are given as below. The deriving of these formulas will not be discussed in this book.

$$\mu = \frac{1}{p}, \sigma^2 = (\frac{1}{p})(\frac{1}{p} - 1), \sigma = \sqrt{(\frac{1}{p})(\frac{1}{p} - 1)}$$

EXAMPLE 4.16

Suppose a game has two outcomes, win or lose. You repeatedly play that game **until** you lose. The probability of losing is $p = 0.57$.

If we let X = the number of games you play until you lose (includes the losing game), then X is a geometric random variable. All three characteristics are met. Each game you play is a Bernoulli trial, either win or lose. You would need to play at least one game before you stop. X takes on the values 1, 2, 3, . . . (could go on indefinitely). Since we are measuring the number of games you play until you lose, we define a success as losing a game and a failure as winning a game. The probability of a success $p = .57$ and the probability of a failure $q = 1 - p = 1 - 0.57 = 0.43$. Both p and q remain the same from game to game.

If we want to find the probability that it takes five games until you lose, then the probability could be written as $P(x = 5)$. We will explain how to find a geometric probability later in this section.

TRY IT 4.16

You throw darts at a board until you hit the center area. Your probability of hitting the center area is $p = 0.17$. You want to find the probability that it takes eight throws until you hit the center. What values does X take on?

EXAMPLE 4.17

A safety engineer feels that 35 percent of all industrial accidents in her plant are caused by failure of employees to follow instructions. She decides to look at the accident reports (selected randomly and replaced in the pile after reading) *until* she finds one that shows an accident caused by failure of employees to follow instructions.

If we let X = the number of accidents the safety engineer must examine until she finds a report showing an accident caused by employee failure to follow instructions, then X is a geometric random variable. All three characteristics are met. Each accident report she reads is a Bernoulli trial: the accident was either caused by failure of employees to follow instructions or not. She would need to read at least one accident report before she stops. X takes on the values 1, 2, 3, . . . (could go on indefinitely). Since we are measuring the number of reports she needs to read until one that shows an accident caused by failure of employees to follow instructions, we define a success as an accident caused by failure of employees to follow instructions. If an accident was caused by another reason, the report is defined as a failure. The probability of a success $p = .35$ and the probability of a failure $q = 1 - p = 1 - .35 = .65$. Both p and q remain the same from report to report.

If we want to find the probability that the safety engineer will have to examine at least three reports until she finds a report showing an accident caused by employee failure to follow instructions, then the probability could be written as $p = .35$. If we want to find how many reports, on average, the safety engineer would *expect* to look at until she finds a report showing an accident caused by employee failure to follow instructions, we need to find the expected value $E(x)$. We will explain how to solve these questions later in this section.

TRY IT 4.17

An instructor feels that 15 percent of students get below a C on their final exam. She decides to look at final exams

(selected randomly and replaced in the pile after reading) until she finds one that shows a grade below a C. We want to know the probability that the instructor will have to examine at least 10 exams until she finds one with a grade below a C. What is the probability question stated mathematically?

EXAMPLE 4.18

Suppose that you are looking for a student at your college who lives within five miles of you. You know that 55 percent of the 25,000 students do live within five miles of you. You randomly contact students from the college *until* one says he or she lives within five miles of you. What is the probability that you need to contact four people?

This is a geometric problem because you may have a number of failures before you have the one success you desire. Also, the probability of a success stays the same each time you ask a student if he or she lives within five miles of you. There is no definite number of trials (number of times you ask a student).

Problem

a. Let X = the number of _____ you must ask _____ one says yes.

Solution

a. Let X = the number of *students* you must ask *until* one says yes.

Problem

b. What values does X take on?

Solution

b. 1, 2, 3, . . ., (total number of students)

Problem

c. What are p and q?

Solution

c. $p = .55$; $q = .45$

Problem

d. The probability question is P(_____).

Solution

d. $P(x = 4)$

TRY IT 4.18

You need to find a store that carries a special printer ink. You know that of the stores that carry printer ink, 10 percent of them carry the special ink. You randomly call each store until one has the ink you need. What are p and q?

Notation for the Geometric: G = Geometric Probability Distribution Function

$X \sim G(p)$

Read this as *X is a random variable with a* ***geometric distribution***. The parameter is p; p = the probability of a success for each trial.

EXAMPLE 4.19

Assume that the probability of a defective computer component is 0.02. Components are randomly selected. Find the

probability that the first defect is caused by the seventh component tested. How many components do you expect to test until one is found to be defective?

Let X = the number of computer components tested until the first defect is found.

X takes on the values 1, 2, 3, . . . where $p = .02$. $X \sim G(.02)$

Find $P(x = 7)$. There is a formula to define the probability of a geometric distribution $P(x)$. We can use the formula to find $P(x = 7)$. But since the calculation is tedious and time consuming, people usually use a graphing calculator or software to get the answer. Using a graphing calculator, you can get $P(x = 7) = .0177$. The instruction of TI83, 83+, 84, 84+ is given below.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Go into 2nd DISTR. The syntax for the instructions are as follows:

To calculate the probability of a value $P(x = \text{value})$, use geometpdf(p, number). Here geometpdf represents geometric probability density function. It is used to find the probability that a geometric random variable is equal to an exact value. p is the probability of a success and number is the value.

To calculate the cumulative probability $P(x \leq \text{value})$, use geometcdf(p, number). Here geometcdf represents geometric cumulative distribution function. It is used to determine the probability of "at most" type of problem, the probability that a geometric random variable is less than or equal to a value. p is the probability of a success and number is the value.

To find $P(x = 7)$, enter 2nd DISTR, arrow down to geometpdf(. Press ENTER. Enter .02,7). The result is $P(x = 7) = .0177$.

If we need to find $P(x \leq 7)$ enter 2nd DISTR, arrow down to geometcdf(. Press ENTER. Enter .02,7). The result is $(x \leq= 7) = .1319$.

The graph of $X \sim G(.02)$ is

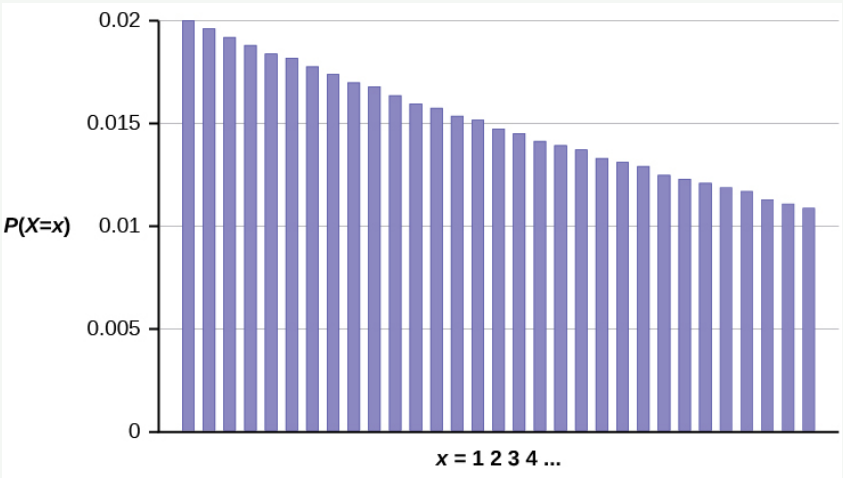


Figure 4.2

The previous probability distribution histogram gives all the probabilities of X. The x-axis of each bar is the value of X = the number of computer components tested until the first defect is found, and the height of that bar is the probability of that value occurring. For example, the x value of the first bar is 1 and the height of the first bar is 0.02. That means the probability that the first computer components tested is defective is .02.

The expected value or mean of X is $E(X) = \mu = \frac{1}{p} = \frac{1}{.02} = 50$.

The variance of X is $\sigma^2 = (\frac{1}{p})(\frac{1}{p} - 1) = (\frac{1}{.02})(\frac{1}{.02} - 1) = (50)(49) = 2{,}450$

The standard deviation of X is $\sigma = \sqrt{\sigma^2} = \sqrt{2{,}450} = 49.5$

Here is how we interpret the mean and standard deviation. The number of components that you would expect to test until you find the first defective one is 50 (which is the mean). And you expect that to vary by about 50 computer components (which is the standard deviation) on average.

TRY IT 4.19

The probability of a defective steel rod is .01. Steel rods are selected at random. Find the probability that the first defect occurs on the ninth steel rod. Use the TI-83+ or TI-84 calculator to find the answer.

EXAMPLE 4.20

Problem

The lifetime risk of developing pancreatic cancer is about one in 78 (1.28 percent). Let X = the number of people you ask until one says he or she has pancreatic cancer. Then X is a discrete random variable with a geometric distribution: $X \sim G\left(\frac{1}{78}\right)$ or $X \sim G(.0128)$.

a. What is the probability that you ask 10 people before one says he or she has pancreatic cancer?
b. What is the probability that you must ask 20 people?
c. Find the (i) mean and (ii) standard deviation of X.

Solution

a. $P(x = 10)$ = geometpdf(.0128, 10) = .0114
b. $P(x = 20)$ = geometpdf(.0128, 20) = .01
c. i. Mean = $\mu = \frac{1}{p} = \frac{1}{.0128} = 78$

ii. $\sigma = \sqrt{\sigma^2} = \sqrt{\left(\frac{1}{p}\right)\left(\frac{1}{p} - 1\right)} = \sqrt{\left(\frac{1}{.0128}\right)\left(\frac{1}{.0128} - 1\right)} = \sqrt{(78)(78-1)} = \sqrt{6{,}006} = 77.4984 \approx 77$

The number of people whom you would expect to ask until one says he or she has pancreatic cancer is 78. And you expect that to vary by about 77 people on average.

TRY IT 4.20

The literacy rate for a nation measures the proportion of people age 15 and over who can read and write. The literacy rate for women in Afghanistan is 12 percent. Let X = the number of Afghani women you ask until one says that she is literate.

a. What is the probability distribution of X?
b. What is the probability that you ask five women before one says she is literate?
c. What is the probability that you must ask 10 women?
d. Find the (i) mean and (ii) standard deviation of X.

4.5 Hypergeometric Distribution (Optional)

There are five characteristics of a **hypergeometric** experiment:

1. You take samples from *two* groups.
2. You are concerned with a group of interest, called the first group.
3. You sample *without replacement* from the combined groups. For example, you want to choose a softball team from a combined group of 11 men and 13 women. The team consists of 10 players.

4. Each pick is *not* independent, since sampling is without replacement. In the softball example, the probability of picking a woman first is $\frac{13}{24}$. The probability of picking a man second is $\frac{11}{23}$ if a woman was picked first. It is $\frac{10}{23}$ if a man was picked first. The probability of the second pick depends on what happened in the first pick.
5. You are **not** dealing with Bernoulli trials.

The outcomes of a hypergeometric experiment fit a **hypergeometric probability** distribution. The random variable X = the number of items from the group of interest.

EXAMPLE 4.21

Problem

A candy dish contains 100 jelly beans and 80 gumdrops. Fifty candies are picked at random. What is the probability that 35 of the 50 are gumdrops? The two groups are jelly beans and gumdrops. Since the probability question asks for the probability of picking gumdrops, the group of interest (first group) is gumdrops. The size of the group of interest (first group) is 80. The size of the second group is 100. The size of the sample is 50 (jelly beans or gumdrops). Let X = the number of gumdrops in the sample of 50. X takes on the values $x = 0, 1, 2, \ldots, 50$. What is the probability statement written mathematically?

Solution

$P(x = 35)$

TRY IT 4.21

A bag contains letter tiles. 44 of the tiles are vowels, and 56 are consonants. Seven tiles are picked at random. You want to know the probability that four of the seven tiles are vowels. What is the group of interest, the size of the group of interest, and the size of the sample?

EXAMPLE 4.22

Problem

Suppose a shipment of 100 DVD players is known to have 10 defective players. An inspector randomly chooses 12 for inspection. He is interested in determining the probability that, among the 12 players, at most two are defective. The two groups are the 90 non-defective DVD players and the 10 defective DVD players. The group of interest (first group) is the defective group because the probability question asks for the probability of at most two defective DVD players. The size of the sample is 12 DVD players. They may be non-defective or defective. Let X = the number of defective DVD players in the sample of 12. X takes on the values $0, 1, 2, \ldots, 10$. X may not take on the values 11 or 12. The sample size is 12, but there are only 10 defective DVD players. Write the probability statement mathematically.

Solution

$P(x \leq 2)$

TRY IT 4.22

A gross of eggs contains 144 eggs. A particular gross is known to have 12 cracked eggs. An inspector randomly chooses 15 for inspection. She wants to know the probability that, among the 15, at most three are cracked. What is X, and what values does it take on?

EXAMPLE 4.23

You are president of an on-campus special events organization. You need a committee of seven students to plan a special birthday party for the president of the college. Your organization consists of 18 women and 15 men. You are

interested in the number of men on your committee. If the members of the committee are randomly selected, what is the probability that your committee has more than four men?

This is a hypergeometric problem because you are choosing your committee from two groups (men and women).

Problem
a. Are you choosing with or without replacement?

Solution
a. without

Problem
b. What is the group of interest?

Solution
b. the men

Problem
c. How many are in the group of interest?

Solution
c. 15 men

Problem
d. How many are in the other group?

Solution
d. 18 women

Problem
e. Let X = _____ on the committee. What values does X take on?

Solution
e. Let X = *the number of men* on the committee. $x = 0, 1, 2, \ldots, 7$.

Problem
f. The probability question is P(_____).

Solution
f. $P(x > 4)$

TRY IT 4.23

A palette has 200 milk cartons. Of the 200 cartons, it is known that 10 of them have leaked and cannot be sold. A stock clerk randomly chooses 18 for inspection. He wants to know the probability that among the 18, no more than two are leaking. Give five reasons why this is a hypergeometric problem.

Notation for the Hypergeometric: H = Hypergeometric Probability Distribution Function

$X \sim H(r, b, n)$

Read this as *X is a random variable with a hypergeometric distribution*. The parameters are r, b, and n: r = the size of the group of interest (first group), b = the size of the second group, n = the size of the chosen sample.

EXAMPLE 4.24

A school site committee is to be chosen randomly from six men and five women. If the committee consists of four members chosen randomly, what is the probability that two of them are men? How many men do you expect to be on the committee?

Let X = the number of men on the committee of four. The men are the group of interest (first group).

X takes on the values 0, 1, 2, 3, 4, where $r = 6$, $b = 5$, and $n = 4$. $X \sim H(6, 5, 4)$

Find $P(x = 2)$. $P(x = 2) = .4545$ (calculator or computer)

NOTE
Currently, the TI-83+ and TI-84 do not have hypergeometric probability functions. There are a number of computer packages, including Microsoft Excel, that do.

The probability that there are two men on the committee is about .45.

The graph of $X \sim H(6, 5, 4)$ is

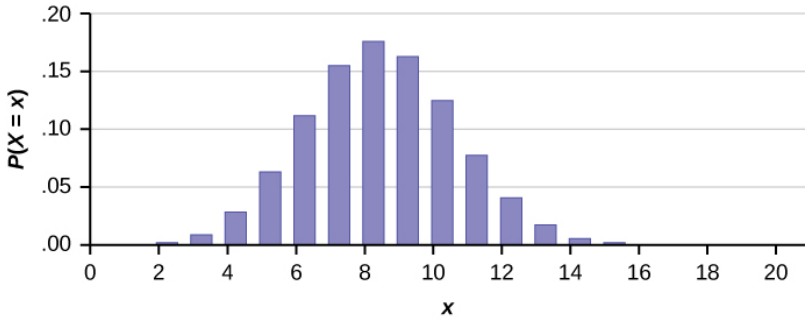


Figure 4.3

The y-axis contains the probability of X, where X = the number of men on the committee.

You would expect $m = 2.18$ (about two) men on the committee.

The formula for the mean is $\mu = \frac{nr}{r+b} = \frac{(4)(6)}{6+5} = 2.18$.

TRY IT 4.24

An intramural basketball team is to be chosen randomly from 15 boys and 12 girls. The team has 10 slots. You want to know the probability that eight of the players will be boys. What is the group of interest and the sample?

4.6 Poisson Distribution (Optional)

There are two main characteristics of a Poisson experiment.

1. The **Poisson probability distribution** gives the probability of a number of events occurring in a *fixed interval* of time or space if these events happen with a known average rate and independently of the time since the last event. For example, a book editor might be interested in the number of words spelled incorrectly in a particular book. It might be that, on the average, there are five words spelled incorrectly in 100 pages. The interval is the 100 pages.
2. The Poisson distribution may be used to approximate the binomial if the probability of success is *small* (such as .01) and the number of trials is *large* (such as 1,000). You will verify the relationship in the homework exercises. n is the number of trials, and p is the probability of a *success*.

The random variable X = the number of occurrences in the interval of interest.

EXAMPLE 4.25

The average number of loaves of bread put on a shelf in a bakery in a half-hour period is 12. Of interest is the number of loaves of bread put on the shelf in five minutes. The time interval of interest is five minutes. What is the probability that the number of loaves, selected randomly, put on the shelf in five minutes is three?

Let X = the number of loaves of bread put on the shelf in five minutes. If the average number of loaves put on the shelf in 30 minutes (half-hour) is 12, then the average number of loaves put on the shelf in five minutes is $\left(\frac{5}{30}\right)(12) = 2$ loaves of bread.

The probability question asks you to find $P(x = 3)$.

TRY IT 4.25

The average number of fish caught in an hour is eight. Of interest is the number of fish caught in 15 minutes. The time interval of interest is 15 minutes. What is the average number of fish caught in 15 minutes?

EXAMPLE 4.26

Problem

A bank expects to receive six bad checks per day, on average. What is the probability of the bank getting fewer than five bad checks on any given day? Of interest is the number of checks the bank receives in one day, so the time interval of interest is one day. Let X = the number of bad checks the bank receives in one day. If the bank expects to receive six bad checks per day then the average is six checks per day. Write a mathematical statement for the probability question.

Solution

$P(x < 5)$

TRY IT 4.26

An electronics store expects to have 10 returns per day on average. The manager wants to know the probability of the store getting fewer than eight returns on any given day. State the probability question mathematically.

EXAMPLE 4.27

You notice that a news reporter says "uh," on average, two times per broadcast. What is the probability that the news reporter says "uh" more than two times per broadcast?

This is a Poisson problem because you are interested in knowing the number of times the news reporter says "uh" during a broadcast.

Problem

a. What is the interval of interest?

Solution

a. one broadcast

Problem

b. What is the average number of times the news reporter says "uh" during one broadcast?

Solution
b. 2

Problem
c. Let X = _____. What values does X take on?

Solution
c. Let X = the number of times the news reporter says "uh" during one broadcast.
x = 0, 1, 2, 3, . . .

Problem
d. The probability question is P(_____).

Solution
d. $P(x > 2)$

TRY IT 4.27

An emergency room at a particular hospital gets an average of five patients per hour. A doctor wants to know the probability that the ER gets more than five patients per hour. Give the reason why this would be a Poisson distribution.

Notation for the Poisson: P = Poisson Probability Distribution Function

$X \sim P(\mu)$

Read this as *X is a random variable with a Poisson distribution*. The parameter is μ (or λ); μ (or λ) = the mean for the interval of interest.

EXAMPLE 4.28

Leah's answering machine receives about six telephone calls between 8 a.m. and 10 a.m. What is the probability that Leah receives more than one call in the next 15 minutes?

Let X = the number of calls Leah receives in 15 minutes. The *interval of interest* is 15 minutes or $\frac{1}{4}$ hour.

x = 0, 1, 2, 3, . . .

If Leah receives, on the average, six telephone calls in two hours, and there are eight 15-minute intervals in two hours, then Leah receives

$\left(\frac{1}{8}\right)(6) = .75$ calls in 15 minutes, on average. So, $\mu = .75$ for this problem.

$X \sim P(.75)$

Find $P(x > 1)$. $P(x > 1) = .1734$ (calculator or computer)

Note
The TI calculators use λ (lambda) for the mean.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

- Press 1 – and then press 2^{nd} DISTR.
- Arrow down to poissoncdf. Press ENTER.
- Enter (.75,1).

- The result is $P(x > 1) = .1734$.

The probability that Leah receives more than one telephone call in the next 15 minutes is about .1734 or $P(x > 1) = 1 - \text{poissoncdf}(.75, 1)$.

The graph of $X \sim P(.75)$ is

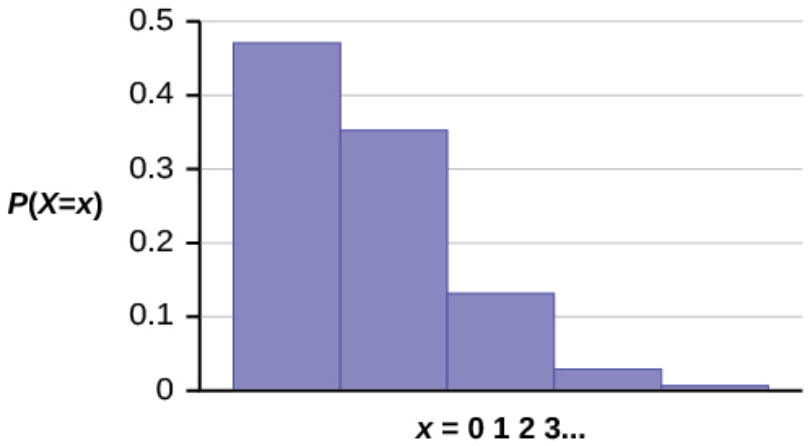


Figure 4.4

The *y*-axis contains the probability of x where X = the number of calls in 15 minutes.

TRY IT 4.28

A customer service center receives about 10 emails every half-hour. What is the probability that the customer service center receives more than four emails in the next six minutes? Use the TI-83+ or TI-84 calculator to find the answer.

EXAMPLE 4.29

According to Baydin, an email management company, an email user gets, on average, 147 emails per day. Let X = the number of emails an email user receives per day. The discrete random variable X takes on the values $x = 0, 1, 2 \ldots$. The random variable X has a Poisson distribution: $X \sim P(147)$. The mean is 147 emails.

Problem

a. What is the probability that an email user receives exactly 160 emails per day?
b. What is the probability that an email user receives at most 160 emails per day?
c. What is the standard deviation?

Solution

a. $P(x = 160) = \text{poissonpdf}(147, 160) \approx .0180$
b. $P(x \leq 160) = \text{poissoncdf}(147, 160) \approx .8666$
c. Standard Deviation = $\sigma = \sqrt{\mu} = \sqrt{147} \approx 12.1244$

TRY IT 4.29

According to a recent poll girls between the ages of 14 and 17 send an average of 187 text messages each day. Let X = the number of texts that a girl aged 14 to 17 sends per day. The discrete random variable X takes on the values $x = 0$,

1, 2 .... The random variable X has a Poisson distribution: $X \sim P(187)$. The mean is 187 text messages.

a. What is the probability that a teen girl sends exactly 175 texts per day?
b. What is the probability that a teen girl sends at most 150 texts per day?
c. What is the standard deviation?

EXAMPLE 4.30

Text message users receive or send an average of 41.5 text messages per day.

Problem

a. How many text messages does a text message user receive or send per hour?
b. What is the probability that a text message user receives or sends two messages per hour?
c. What is the probability that a text message user receives or sends more than two messages per hour?

Solution

a. Let X = the number of texts that a user sends or receives in one hour. The average number of texts received per hour is $\frac{41.5}{24} \approx 1.7292$.
b. $X \sim P(1.7292)$, so $P(x = 2)$ = poissonpdf(1.7292, 2) ≈ .2653
c. $P(x > 2) = 1 - P(x \leq 2) = 1 -$ poissoncdf(1.7292, 2) ≈ 1 – .7495 = .2505

TRY IT 4.30

Scientists recently researched the busiest airport in the world. On average, there are 2,500 arrivals and departures each day.

a. How many airplanes arrive and depart the airport per hour?
b. What is the probability that there are exactly 100 arrivals and departures in one hour?
c. What is the probability that there are at most 100 arrivals and departures in one hour?

EXAMPLE 4.31

Problem

On May 13, 2013, starting at 4:30 p.m., the probability of low seismic activity for the next 48 hours in Alaska was reported as about 1.02 percent. Use this information for the next 200 days to find the probability that there will be low seismic activity in 10 of the next 200 days. Use both the binomial and Poisson distributions to calculate the probabilities. Are they close?

Solution

Let X = the number of days with low seismic activity. We are asked to find the probability over the next 200 days. Probability is given for a 48-hour period.

n = 200 days = 100 48-hour periods

Using the binomial distribution

$P(x = 10)$ = binompdf(100, .0102, 10) ≈ .0000000076

Using the Poisson distribution

Calculate $\mu = np = 100(.0102) \approx 1.02$
$P(x = 10)$ = poissonpdf(1.02, 10) ≈ .000000121

We expect the approximation to be good because n is large (greater than 20) and p is small (less than .05). The results are close—both probabilities reported are almost 0.

TRY IT 4.31

On May 13, 2013, starting at 4:30 p.m., the probability of moderate seismic activity for the next 48 hours in the Kuril Islands off the coast of Japan was reported at about 1.43 percent. Use this information for the next 100 days to find the probability that there will be low seismic activity in 5 of the next 100 days. Use both the binomial and Poisson distributions to calculate the probabilities. Are they close?

4.7 Discrete Distribution (Playing Card Experiment)

Stats Lab

Discrete Distribution (Playing Card Experiment)

Student Learning Outcomes

- The student will compare empirical data and a theoretical distribution to determine if an everyday experiment fits a discrete distribution.
- The student will compare technology-generated simulation and a theoretical distribution.
- The student will demonstrate an understanding of long-term probabilities.

Supplies

- One full deck of playing cards
- Programmable calculator

Procedure for Empirical Data

The experimental procedure for empirical data is to pick one card from a deck of shuffled cards.

1. The theoretical probability of picking a diamond from a deck is _____.
2. Shuffle a deck of cards.
3. Pick one card from it.
4. Record whether it was a diamond or not a diamond.
5. Put the card back and reshuffle.
6. Do this a total of 10 times.
7. Record the number of diamonds picked.
8. Let X = number of diamonds. Theoretically, $X \sim B$(_____,_____)

Procedure for Simulation

Repeat the experimental procedure using a programmable calculator.

1. Use the randInt function to generate data. Consider 1 to be spades, 2 to be hearts, 3 to be diamonds, and 4 to be clubs. Generate 10 draws of cards with four suits with randInt(1,4,10).
2. Let $X =$ number of diamonds. Theoretically, $X \sim B$(_____,_____).

Organize the Empirical Data

1. Record the number of diamonds picked for your class with playing cards in Table 4.15. Then calculate the relative frequency.

x	Frequency	Relative Frequency
0	_____	_____
1	_____	_____
2	_____	_____

Table 4.15

x	Frequency	Relative Frequency
3	_____	_____
4	_____	_____
5	_____	_____
6	_____	_____
7	_____	_____
8	_____	_____
9	_____	_____
10	_____	_____

Table 4.15

2. Calculate the following:
   a. $\overline{x}$ = _____
   b. s = _____
3. Construct a histogram of the empirical data.

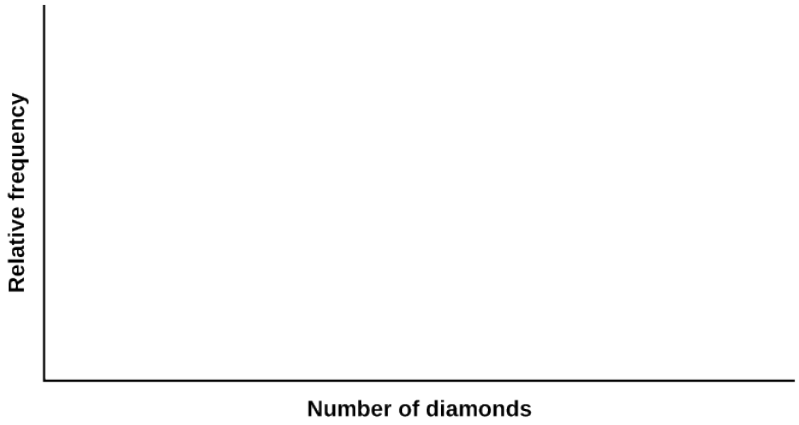


Figure 4.5

Organize the Simulation Data

1. Use Table 4.16 to record the number of diamonds picked for your class using the calculator simulation. Calculate the relative frequency.

X	Frequency	Relative Frequency
0	_____	_____
1	_____	_____
2	_____	_____
3	_____	_____

Table 4.16

X	Frequency	Relative Frequency
4	_____	_____
5	_____	_____
6	_____	_____
7	_____	_____
8	_____	_____
9	_____	_____
10	_____	_____

Table 4.16

2. Calculate the following:
   a. $\overline{x}$ = _____
   b. s = _____
3. Construct a histogram of the simulation data.

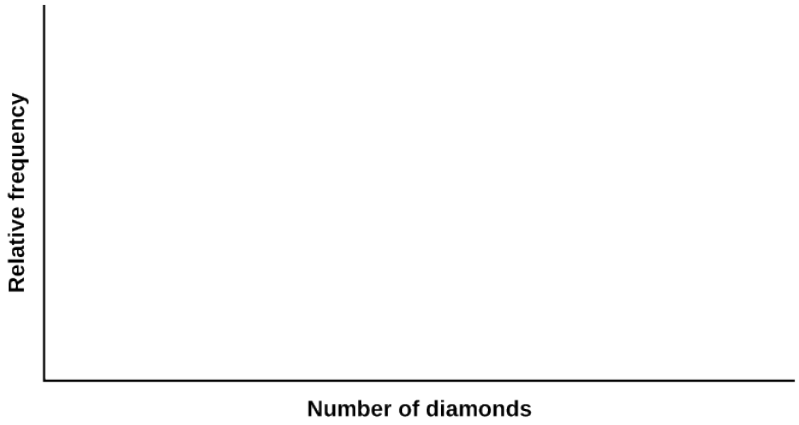


Figure 4.6

Theoretical Distribution

a. Build the theoretical PDF chart based on the distribution in the Procedure section.

x	P(x)
0	
1	
2	
3	
4	

Table 4.17

x	$P(x)$
5	
6	
7	
8	
9	
10	

Table 4.17

b. Calculate the following:
   a. μ = _____
   b. σ = _____

c. Construct a histogram of the theoretical distribution.

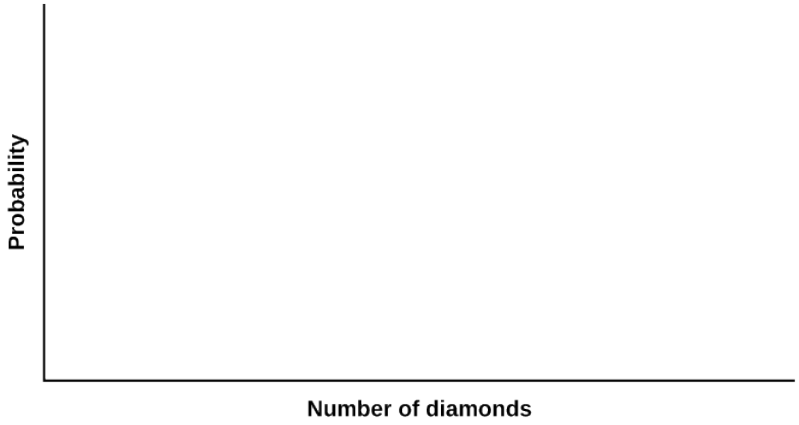


Figure 4.7

Using the Data

NOTE

RF = relative frequency

Use the table from the Theoretical Distribution section to calculate the following answers. Round your answers to four decimal places.

- $P(x = 3)$ = _____
- $P(1 < x < 4)$ = _____
- $P(x \geq 8)$ = _____

Use the data from the Organize the Empirical Data section to calculate the following answers. Round your answers to four decimal places.

- $RF(x = 3)$ = _____
- $RF(1 < x < 4)$ = _____
- $RF(x \geq 8)$ = _____

Use the data from the Organize the Simulation Data section to calculate the following answers. Round your answers to four decimal places.

- $RF(x = 3)$ = _____
- $RF(1 < x < 4)$ = _____
- $RF(x \geq 8)$ = _____

Discussion Questions

For Questions 1 and 2, think about the shapes of the two graphs, the probabilities, the relative frequencies, the means, and the standard deviations.

1. Knowing that data vary, describe three similarities between the graphs and distributions of the theoretical, empirical, and simulation distributions. Use complete sentences.
2. Describe the three most significant differences between the graphs or distributions of the theoretical, empirical, and simulation distributions.
3. Using your answers from Questions 1 and 2, does it appear that the two sets of data fit the theoretical distribution? In complete sentences, explain why or why not.
4. Suppose that the experiment had been repeated 500 times. Would you expect Table 4.15, Table 4.16, or Table 4.17 to change, and how would it change? Why? Why wouldn't the other table(s) change?

4.8 Discrete Distribution (Lucky Dice Experiment)

Stats Lab

Discrete Distribution (Lucky Dice Experiment)

Student Learning Outcomes

- The student will compare empirical data and a theoretical distribution to determine if a Tet gambling game fits a discrete distribution.
- The student will demonstrate an understanding of long-term probabilities.

Supplies

- One "Lucky Dice" game or three regular dice
- One programming calculator

Procedure

Round answers to relative frequency and probability problems to four decimal places.

1. The experimental procedure is to bet on one object. Then, roll three Lucky Dice and count the number of matches. The number of matches will decide your profit.
2. What is the theoretical probability of one die matching the object?
3. Choose one object to place a bet on. Roll the three Lucky Dice. Count the number of matches.
4. Let X = number of matches. Theoretically, $X \sim B$(_____,_____)
5. Let Y = profit per game.

Organize the Data

In Table 4.18, fill in the y-value that corresponds to each x-value. Next, record the number of matches picked for your class. Then, calculate the relative frequency.

1. Complete the table.

x	y	Frequency	Relative Frequency
0			
1			
2			
3			

Table 4.18

2. Calculate the following:
   a. $\overline{x}$ = _____
   b. s_x = _____
   c. $\overline{y}$ = _____
   d. s_y = _____
3. Explain what $\overline{x}$ represents.
4. Explain what $\overline{y}$ represents.
5. Based upon the experiment, answer the following questions:
   a. What was the average profit per game?
   b. Did this represent an average win or loss per game?
   c. How do you know? Answer in complete sentences.
6. Construct a histogram of the empirical data.

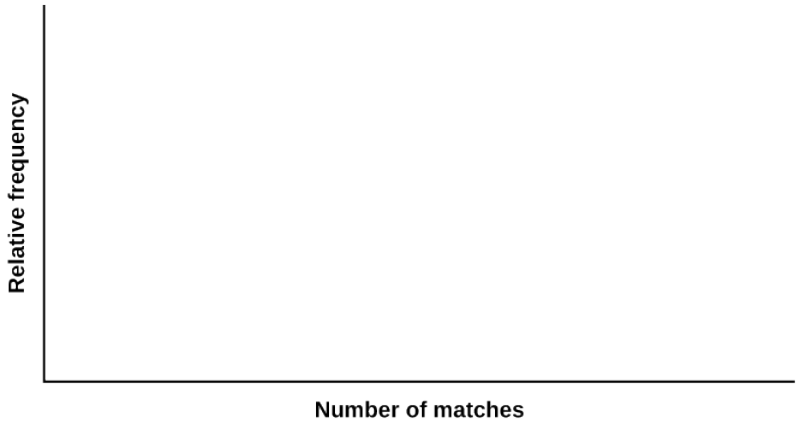


Figure 4.8

Theoretical Distribution

Build the theoretical PDF chart for *x* and *y* based on the distribution from the Procedure section.

1.

x	y	P(x) = P(y)
0		
1		
2		
3		

Table 4.19

2. Calculate the following:
   a. μ_X = _____
   b. σ_X = _____
   c. μ_X = _____
3. Explain what μ_X represents.
4. Explain what μ_Y represents.
5. Based upon theory, answer the following questions:
   a. What was the expected profit per game?
   b. Did the expected profit represent an average win or loss per game?
   c. How do you know? Answer in complete sentences.
6. Construct a histogram of the theoretical distribution.

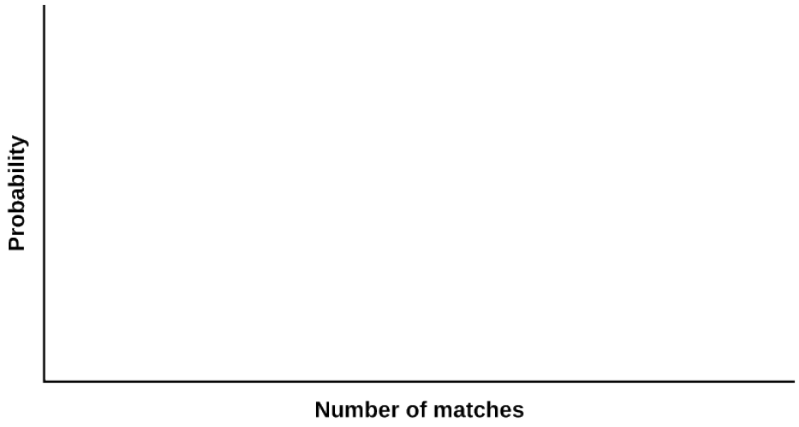


Figure 4.9

Use the Data

Note

RF = relative frequency

Use the data from the Theoretical Distribution section to calculate the following answers. Round your answers to four decimal places.

1. $P(x = 3)$ = _____
2. $P(0 < x < 3)$ = _____
3. $P(x \geq 2)$ = _____

Use the data from the Organize the Data section to calculate the following answers. Round your answers to four decimal places.

1. $RF(x = 3)$ = _____
2. $RF(0 < x < 3)$ = _____
3. $RF(x \geq 2)$ = _____

Discussion Question

For Questions 1 and 2, consider the graphs, the probabilities, the relative frequencies, the means, and the standard deviations.

1. Knowing that data vary, describe three similarities between the graphs and distributions of the theoretical and empirical distributions. Use complete sentences.
2. Describe the three most significant differences between the graphs or distributions of the theoretical and empirical distributions.
3. Thinking about your answers to Questions 1 and 2, does it appear that the data fit the theoretical distribution? In

complete sentences, explain why or why not.

4. Suppose that the experiment had been repeated 500 times. Would you expect Table 4.18 or Table 4.19 to change, and how would it change? Why? Why wouldn't the other table change?

Key Terms

Bernoulli trials an experiment with the following characteristics:

1. There are only two possible outcomes called *success* and *failure* for each trial
2. The probability p of a success is the same for any trial (so the probability $q = 1 - p$ of a failure is the same for any trial)

binomial experiment a statistical experiment that satisfies the following three conditions:

1. There are a fixed number of trials, n
2. There are only two possible outcomes, called *success* and, *failure*, for each trial; the letter p denotes the probability of a success on one trial, and q denotes the probability of a failure on one trial
3. The n trials are independent and are repeated using identical conditions

binomial probability distribution a discrete random variable (RV) that arises from Bernoulli trials; there are a fixed number, n, of independent trials

Independent means that the result of any trial (for example, trial one) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: $X \sim B(n, p)$. The mean is $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of the following exactly x successes in n trials is

$P(X = x) = \binom{n}{x} p^x q^{n-x}$

expected value expected arithmetic average when an experiment is repeated many times; also called the mean; notations μ; for a discrete random variable (RV) with probability distribution function $P(x)$,the definition can also be written in the form $\mu = \sum xP(x)$

geometric distribution a discrete random variable (RV) that arises from the Bernoulli trials; the trials are repeated until the first success.

The geometric variable X is defined as the number of trials until the first success. Notation $X \sim G(p)$. The mean is $\mu = \frac{1}{p}$ and the standard deviation is $\sigma = \sqrt{\frac{1}{p}\left(\frac{1}{p} - 1\right)}$. The probability of exactly x failures before the first success is given by the formula

$$P(X = x) = p(1-p)^{x-1}$$

.

geometric experiment a statistical experiment with the following properties:

1. There are one or more Bernoulli trials with all failures except the last one, which is a success
2. In theory, the number of trials could go on forever; there must be at least one trial
3. The probability, p, of a success and the probability, q, of a failure do not change from trial to trial

hypergeometric experiment a statistical experiment with the following properties:

1. You take samples from two groups
2. You are concerned with a group of interest, called the first group
3. You sample without replacement from the combined groups
4. Each pick is not independent, since sampling is without replacement
5. You are not dealing with Bernoulli trials

hypergeometric probability a discrete random variable (RV) that is characterized by the following:

1. The experiment uses a fixed number of trials.
2. The probability of success is not the same from trial to trial

We sample from two groups of items when we are interested in only one group. X is defined as the number of successes out of the total number of items chosen. Notation $X \sim H(r, b, n)$, where r = the number of items in the group of interest, b = the number of items in the group not of interest, and n = the number of items chosen.

mean a number that measures the central tendency; a common name for mean is *average*

The term *mean* is a shortened form of *arithmetic mean*. By definition, the mean for a sample (denoted by $\overline{x}$) is $\overline{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$ and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

mean of a probability distribution the long-term average of many trials of a statistical experiment

Poisson probability distribution a discrete random variable (RV) that counts the number of times a certain event will occur in a specific interval; characteristics of the variable:

- The probability that the event occurs in a given interval is the same for all intervals
- The events occur with a known mean and independently of the time since the last event

The distribution is defined by the mean μ of the event in the interval. Notation $X \sim P(\mu)$. The mean is $\mu = np$. The standard deviation is $\sigma = \sqrt{\mu}$. The probability of having exactly x successes in r trials is $P(X = x) = (e^{-\mu})\frac{\mu^x}{x!}$. The Poisson distribution is often used to approximate the binomial distribution, when n is *large* and p is *small* (a general rule is that n should be greater than or equal to 20 and p should be less than or equal to .05).

probability distribution function (PDF) a mathematical description of a discrete random variable (*RV*), given either in the form of an equation (formula) or in the form of a table listing all the possible outcomes of an experiment and the probability associated with each outcome

random variable (RV) a characteristic of interest in a population being studied; common notation for variables are uppercase Latin letters *X*, *Y*, *Z*, . . . ; common notation for a specific value from the domain (set of all possible values of a variable) are lowercase Latin letters *x*, *y*, and *z*

For example, if *X* is the number of children in a family, then *x* represents a specific integer 0, 1, 2, 3, . . . ; variables in statistics differ from variables in intermediate algebra in the two following ways:

- The domain of the random variable (RV) is not necessarily a numerical set; the domain may be expressed in words; for example, if *X* = hair color then the domain is {black, blond, gray, green, orange}
- We can tell what specific value *x* the random variable *X* takes only after performing the experiment

standard deviation of a probability distribution a number that measures how far the outcomes of a statistical experiment are from the mean of the distribution

the law of large numbers as the number of trials in a probability experiment increases, the difference between the theoretical probability of an event and the relative frequency probability approaches zero

Chapter Review

4.1 Probability Distribution Function (PDF) for a Discrete Random Variable

The characteristics of a probability distribution function (PDF) for a discrete random variable are as follows:

1. Each probability is between zero and one, inclusive (*inclusive* means to include zero and one)
2. The sum of the probabilities is one

4.2 Mean or Expected Value and Standard Deviation

The expected value, or mean, of a discrete random variable predicts the long-term results of a statistical experiment that has been repeated many times. The standard deviation of a probability distribution is used to measure the variability of possible outcomes.

4.3 Binomial Distribution (Optional)

A statistical experiment can be classified as a binomial experiment if the following conditions are met:

1. There are a fixed number of trials, n
2. There are only two possible outcomes, called *success* and *failure*, for each trial; the letter p denotes the probability of a success on one trial and q denotes the probability of a failure on one trial
3. The n trials are independent and are repeated using identical conditions

The outcomes of a binomial experiment fit a binomial probability distribution. The random variable X = the number of successes obtained in the n independent trials. The mean of X can be calculated using the formula $\mu = np$, and the standard deviation is given by the formula $\sigma = \sqrt{npq}$.

4.4 Geometric Distribution (Optional)

There are three characteristics of a geometric experiment:

1. There are one or more Bernoulli trials with all failures except the last one, which is a success
2. In theory, the number of trials could go on forever; there must be at least one trial
3. The probability, p, of a success and the probability, q, of a failure are the same for each trial

In a geometric experiment, define the discrete random variable X as the number of independent trials until the first success. We say that X has a geometric distribution and write $X \sim G(p)$ where p is the probability of success in a single trial.

The mean of the geometric distribution $X \sim G(p)$ is $\mu = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1}{p}\left(\frac{1}{p} - 1\right)}$.

4.5 Hypergeometric Distribution (Optional)

A **hypergeometric experiment** is a statistical experiment with the following properties:

1. You take samples from two groups
2. You are concerned with a group of interest, called the first group
3. You sample without replacement from the combined groups
4. Each pick is not independent, since sampling is without replacement
5. You are not dealing with Bernoulli trials

The outcomes of a hypergeometric experiment fit a hypergeometric probability distribution. The random variable X = the number of items from the group of interest. The distribution of X is denoted $X \sim H(r, b, n)$, where r = the size of the group of interest (first group), b = the size of the second group, and n = the size of the chosen sample. It follows that $n \leq r + b$.

The mean of X is $\mu = \frac{nr}{r+b}$ and the standard deviation is $\sigma = \sqrt{\frac{rbn(r+b-n)}{(r+b)^2 \quad (r+b-1)}}$.

4.6 Poisson Distribution (Optional)

A **Poisson probability distribution** of a discrete random variable gives the probability of a number of events occurring in a *fixed interval* of time or space, if these events happen at a known average rate and independently of the time since the last event. The Poisson distribution may be used to approximate the binomial, if the probability of success is *small* (less than or equal to .05) and the number of trials is *large* (greater than or equal to 20).

Formula Review

4.2 Mean or Expected Value and Standard Deviation

Mean or Expected Value: $\mu = \sum_{x \in X} xP(x)$

Standard Deviation: $\sigma = \sqrt{\sum_{x \in X} (x-\mu)^2 P(x)}$

4.3 Binomial Distribution (Optional)

$X \sim B(n, p)$ means that the discrete random variable X has a binomial probability distribution with n trials and probability of success p.

X = the number of successes in n independent trials

n = the number of independent trials

X takes on the values $x = 0, 1, 2, 3, \ldots, n$

p = the probability of a success for any trial

q = the probability of a failure for any trial

$p + q = 1$

$q = 1 - p$

The mean of X is $\mu = np$. The standard deviation of X is $\sigma = \sqrt{npq}$.

4.4 Geometric Distribution (Optional)

$X \sim G(p)$ means that the discrete random variable X has a geometric probability distribution with probability of success in a single trial p.

X = the number of independent trials until the first success

X takes on the values $x = 1, 2, 3, \ldots$

p = the probability of a success for any trial

q = the probability of a failure for any trial
$p + q = 1 : q = 1 - p$

The mean is $\mu = \frac{1}{p}$.

The standard deviation is $\sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1}{p}\left(\frac{1}{p} - 1\right)}$.

4.5 Hypergeometric Distribution (Optional)

$X \sim H(r, b, n)$ means that the discrete random variable X has a hypergeometric probability distribution with r = the size of the group of interest (first group), b = the size of the second group, and n = the size of the chosen sample.

X = the number of items from the group of interest that are in the chosen sample, and X may take on the values $x = 0, 1, \ldots,$ up to the size of the group of interest. The minimum value for X may be larger than zero in some instances.

$n \leq r + b$

The mean of X is given by the formula $\mu = \frac{nr}{r+b}$ and the standard deviation is $= \sqrt{\frac{rbn(r+b-n)}{(r+b)^2(r+b-1)}}$.

4.6 Poisson Distribution (Optional)

$X \sim P(\mu)$ means that X has a Poisson probability distribution where X = the number of occurrences in the interval of interest.

X takes on the values $x = 0, 1, 2, 3, \ldots$

The mean μ is typically given.

The variance is $\sigma^2 = \mu$, and the standard deviation is $\sigma = \sqrt{\mu}$.

When $P(\mu)$ is used to approximate a binomial distribution, $\mu = np$ where n represents the number of independent trials and p represents the probability of success in a single trial.

Practice

4.1 Probability Distribution Function (PDF) for a Discrete Random Variable

Use the following information to answer the next five exercises: A company wants to evaluate its attrition rate, or in other words, how long new hires stay with the company. Over the years, the company has established the following probability distribution:

Let X = the number of years a new hire will stay with the company.

Let $P(x)$ = the probability that a new hire will stay with the company x years.

1. Complete Table 4.20 using the data provided.

x	$P(x)$
0	.12
1	.18
2	.30
3	.15
4	
5	.10
6	.05

Table 4.20

2. $P(x = 4)$ = _____
3. $P(x \geq 5)$ = _____
4. On average, how long would you expect a new hire to stay with the company?
5. What does the column "$P(x)$" sum to?

Use the following information to answer the next four exercises: A baker is deciding how many batches of muffins to make to sell in his bakery. He wants to make enough to sell every one and no fewer. Through observation, the baker has established a probability distribution.

x	$P(x)$
1	.15
2	.35
3	.40
4	.10

Table 4.21

6. Define the random variable X.
7. What is the probability the baker will sell more than one batch? $P(x > 1) =$ _____
8. What is the probability the baker will sell exactly one batch? $P(x = 1) =$ _____
9. On average, how many batches should the baker make?

Use the following information to answer the next two exercises: Ellen has music practice three days a week. She practices for all of the three days 85 percent of the time, two days 8 percent of the time, one day 4 percent of the time, and no days 3 percent of the time. One week is selected at random.

10. Define the random variable X.
11. Construct a probability distribution table for the data.
12. We know that for a probability distribution function to be discrete, it must have two characteristics. One is that the sum of the probabilities is one. What is the other characteristic?

Use the following information to answer the next five exercises: Javier volunteers in community events each month. He does not do more than five events in a month. He attends exactly five events 35 percent of the time, four events 25 percent of the time, three events 20 percent of the time, two events 10 percent of the time, one event 5 percent of the time, and no events 5 percent of the time.

13. Define the random variable X.
14. What values does x take on?
15. Construct a PDF table.
16. Find the probability that Javier volunteers for fewer than three events each month. $P(x < 3) =$ _____
17. Find the probability that Javier volunteers for at least one event each month. $P(x > 0) =$ _____

4.2 Mean or Expected Value and Standard Deviation

18. Complete the expected value table.

x	$P(x)$	$x*P(x)$
0	.2	
1	.2	
2	.4	
3	.2	

Table 4.22

19. Find the expected value from the expected value table.

x	$P(x)$	$x*P(x)$
2	.1	2(.1) = .2
4	.3	4(.3) = 1.2

Table 4.23

x	$P(x)$	$x*P(x)$
6	.4	6(.4) = 2.4
8	.2	8(.2) = 1.6

Table 4.23

20. Find the standard deviation.

x	$P(x)$	$x*P(x)$	$(x-\mu)^2P(x)$
2	0.1	2(.1) = .2	$(2–5.4)^2(.1) = 1.156$
4	0.3	4(.3) = 1.2	$(4–5.4)^2(.3) = .588$
6	0.4	6(.4) = 2.4	$(6–5.4)^2(.4) = .144$
8	0.2	8(.2) = 1.6	$(8–5.4)^2(.2) = 1.352$

Table 4.24

21. Identify the mistake in the probability distribution table.

x	$P(x)$	$x*P(x)$
1	.15	.15
2	.25	.50
3	.30	.90
4	.20	.80
5	.15	.75

Table 4.25

22. Identify the mistake in the probability distribution table.

x	$P(x)$	$x*P(x)$
1	.15	.15
2	.25	.40
3	.25	.65

Table 4.26

x	$P(x)$	$x*P(x)$
4	.20	.85
5	.15	1

Table 4.26

Use the following information to answer the next five exercises: A physics professor wants to know what percent of physics majors will spend the next several years doing postgraduate research. He has the following probability distribution:

x	$P(x)$	$x*P(x)$
1	.35	
2	.20	
3	.15	
4		
5	.10	
6	.05	

Table 4.27

23. Define the random variable X.

24. Define $P(x)$, or the probability of x.

25. Find the probability that a physics major will do postgraduate research for four years. $P(x = 4)$ = _____

26. Find the probability that a physics major will do postgraduate research for at most three years. $P(x \leq 3)$ = _____

27. On average, how many years would you expect a physics major to spend doing postgraduate research?

Use the following information to answer the next seven exercises: A ballet instructor is interested in knowing what percent of each year's class will continue on to the next so that she can plan what classes to offer. Over the years, she has established the following probability distribution:

- Let X = the number of years a student will study ballet with the teacher.
- Let $P(x)$ = the probability that a student will study ballet x years.

28. Complete Table 4.28 using the data provided.

x	$P(x)$	$x*P(x)$
1	.10	
2	.05	
3	.10	

Table 4.28

x	$P(x)$	$x*P(x)$
4		
5	.30	
6	.20	
7	.10	

Table 4.28

29. In words, define the random variable X.

30. $P(x = 4)$ = _____

31. $P(x < 4)$ = _____

32. On average, how many years would you expect a child to study ballet with this teacher?

33. What does the column $P(x)$ sum to and why?

34. What does the column $x*P(x)$ sum to and why?

35. You are playing a game by drawing a card from a standard deck and replacing it. If the card is a face card, you win $30. If it is not a face card, you pay $2. There are 12 face cards in a deck of 52 cards. What is the expected value of playing the game?

36. You are playing a game by drawing a card from a standard deck and replacing it. If the card is a face card, you win $30. If it is not a face card, you pay $2. There are 12 face cards in a deck of 52 cards. Should you play the game?

4.3 Binomial Distribution (Optional)

Use the following information to answer the next eight exercises: Researchers collected data from 203,967 incoming first-time, full-time freshmen from 270 four-year colleges and universities in the United States. Of those students, 71.3 percent replied that, yes, they agreed with a recent federal law that was passed.

Suppose that you randomly pick eight first-time, full-time freshmen from the survey. You are interested in the number who agreed with that law.

37. In words, define the random variable X.

38. $X \sim$ _____(_____,_____)

39. What values does the random variable X take on?

40. Construct the probability distribution function (PDF).

x	$P(x)$

Table 4.29

41. On average (μ), how many would you expect to answer yes?

42. What is the standard deviation (σ)?

43. What is the probability that at most five of the freshmen reply yes?

44. What is the probability that at least two of the freshmen reply yes?

4.4 Geometric Distribution (Optional)

Use the following information to answer the next six exercises: Researchers collected data from 203,967 incoming first-time, full-time freshmen from 270 four-year colleges and universities in the United States. Of those students, 71.3 percent replied that, yes, they agree with a recent law that was passed. Suppose that you randomly select freshman from the study until you find one who replies yes. You are interested in the number of freshmen you must ask.

45. In words, define the random variable X.

46. $X \sim$ _____(_____,_____)

47. What values does the random variable X take on?

48. Construct the probability distribution function (PDF). Stop at $x = 6$.

x	$P(x)$
1	
2	
3	
4	
5	
6	

Table 4.30

49. On average (μ), how many freshmen would you expect to have to ask until you found one who replies yes?

50. What is the probability that you will need to ask fewer than three freshmen?

4.5 Hypergeometric Distribution (Optional)

Use the following information to answer the next five exercises: Suppose that a group of statistics students is divided into two groups: business majors and non-business majors. There are 16 business majors in the group and seven non-business majors in the group. A random sample of nine students is taken. We are interested in the number of business majors in the sample.

51. In words, define the random variable X.

52. $X \sim$ _____(_____,_____)

53. What values does X take on?

54. Find the standard deviation.

55. On average (μ), how many would you expect to be business majors?

4.6 Poisson Distribution (Optional)

Use the following information to answer the next six exercises: On average, a clothing store gets 120 customers per day.

56. Assume the event occurs independently in any given day. Define the random variable X.

57. What values does X take on?

58. What is the probability of getting 150 customers in one day?

59. What is the probability of getting 35 customers in the first four hours? Assume the store is open 12 hours each day.

60. What is the probability that the store will have more than 12 customers in the first hour?

61. What is the probability that the store will have fewer than 12 customers in the first two hours?

62. Which type of distribution can the Poisson model be used to approximate? When would you do this?

Use the following information to answer the next six exercises: On average, eight teens in the United States die from motor vehicle injuries per day. As a result, states across the country are debating raising the driving age.

63. Assume the event occurs independently in any given day. In words, define the random variable *X*.

64. *X* ~ _____(_____,_____)

65. What values does *X* take on?

66. For the given values of the random variable *X*, fill in the corresponding probabilities.

67. Is it likely that there will be no teens killed from motor vehicle injuries on any given day in the United States? Justify your answer numerically.

68. Is it likely that there will be more than 20 teens killed from motor vehicle injuries on any given day in the United States? Justify your answer numerically.

Homework

4.1 Probability Distribution Function (PDF) for a Discrete Random Variable

69. Suppose that the PDF for the number of years it takes to earn a bachelor of science (B.S.) degree is given in Table 4.31.

x	*P*(*x*)
3	.05
4	.40
5	.30
6	.15
7	.10

Table 4.31

a. In words, define the random variable *X*.
b. What does it mean that the values 0, 1, and 2 are not included for *x* in the PDF?

4.2 Mean or Expected Value and Standard Deviation

70. A theater group holds a fund-raiser. It sells 100 raffle tickets for $5 apiece. Suppose you purchase four tickets. The prize is two passes to a Broadway show, worth a total of $150.
a. What are you interested in here?
b. In words, define the random variable *X*.
c. List the values that *X* may take on.
d. Construct a PDF.
e. If this fund-raiser is repeated often and you always purchase four tickets, what would be your expected average winnings per raffle?

71. A game involves selecting a card from a regular 52-card deck and tossing a coin. The coin is a fair coin and is

equally likely to land on heads or tails.

- If the card is a face card, and the coin lands on heads, you win $6.
- If the card is a face card, and the coin lands on tails, you win $2.
- If the card is not a face card, you lose $2, no matter what the coin shows.

a. Find the expected value for this game (expected net gain or loss).
b. Explain what your calculations indicate about your long-term average profits and losses on this game.
c. Should you play this game to win money?

72. You buy a ticket to a raffle that costs $10 per ticket. There are only 100 tickets available to be sold in this raffle. In this raffle there are one $500 prize, two $100 prizes, and four $25 prizes. Find your expected gain or loss.

73. Complete the PDF and answer the questions.

x	$P(x)$	$xP(x)$
0	.3	
1	.2	
2		
3	.4	

Table 4.32

a. Find the probability that $x = 2$.
b. Find the expected value.

74. Suppose that you are offered the following deal: You roll a die. If you roll a six, you win $10. If you roll a four or five, you win $5. If you roll a one, two, or three, you pay $6.

a. What are you ultimately interested in here (the value of the roll or the money you win)?
b. In words, define the random variable X.
c. List the values that X may take on.
d. Construct a PDF.
e. Over the long run of playing this game, what are your expected average winnings per game?
f. Based on numerical values, should you take the deal? Explain your decision in complete sentences.

75. A venture capitalist, willing to invest $1,000,000, has three investments to choose from: The first investment, a software company, has a 10 percent chance of returning $5,000,000 profit, a 30 percent chance of returning $1,000,000 profit, and a 60 percent chance of losing the million dollars. The second company, a hardware company, has a 20 percent chance of returning $3,000,000 profit, a 40 percent chance of returning $1,000,000 profit, and a 40 percent chance of losing the million dollars. The third company, a biotech firm, has a 10 percent chance of returning $6,000,000 profit, a 70 percent of no profit or loss, and a 20 percent chance of losing the million dollars.

a. Construct a PDF for each investment.
b. Find the expected value for each investment.
c. Which is the safest investment? Why do you think so?
d. Which is the riskiest investment? Why do you think so?
e. Which investment has the highest expected return, on average?

76. Suppose that 20,000 married adults in the United States were randomly surveyed as to the number of children they have. The results are compiled and are used as theoretical probabilities. Let X = the number of children married people have.

x	$P(x)$	$xP(x)$
0	.10	
1	.20	
2	.30	
3		
4	.10	
5	.05	
6 (or more)	.05	

Table 4.33

a. Find the probability that a married adult has three children.
b. In words, what does the expected value in this example represent?
c. Find the expected value.
d. Is it more likely that a married adult will have two to three children or four to six children? How do you know?

77. Suppose that the PDF for the number of years it takes to earn a bachelor of science (B.S.) degree is given as in Table 4.34.

x	$P(x)$
3	.05
4	.40
5	.30
6	.15
7	.10

Table 4.34

On average, how many years do you expect it to take for an individual to earn a B.S.?

78. People visiting video rental stores often rent more than one DVD at a time. The probability distribution for DVD rentals per customer at Video to Go is given in the following table. There is a five-video limit per customer at this store, so nobody ever rents more than five DVDs.

x	$P(x)$
0	.03
1	.50

Table 4.35

x	$P(x)$
2	.24
3	
4	.07
5	.04

Table 4.35

a. Describe the random variable X in words.
b. Find the probability that a customer rents three DVDs.
c. Find the probability that a customer rents at least four DVDs.
d. Find the probability that a customer rents at most two DVDs.

Another shop, Entertainment Headquarters, rents DVDs and video games. The probability distribution for DVD rentals per customer at this shop is given as follows. They also have a five-DVD limit per customer.

x	$P(x)$
0	.35
1	.25
2	.20
3	.10
4	.05
5	.05

Table 4.36

e. At which store is the expected number of DVDs rented per customer higher?
f. If Video to Go estimates that they will have 300 customers next week, how many DVDs do they expect to rent next week? Answer in sentence form.
g. If Video to Go expects 300 customers next week, and Entertainment Headquarters projects that they will have 420 customers, for which store is the expected number of DVD rentals for next week higher? Explain.
h. Which of the two video stores experiences more variation in the number of DVD rentals per customer? How do you know that?

79. A "friend" offers you the following deal: For a $10 fee, you may pick an envelope from a box containing 100 seemingly identical envelopes. However, each envelope contains a coupon for a free gift.

- Ten of the coupons are for a free gift worth $6.
- Eighty of the coupons are for a free gift worth $8.
- Six of the coupons are for a free gift worth $12.
- Four of the coupons are for a free gift worth $40.

Based upon the financial gain or loss over the long run, should you play the game?

a. Yes, I expect to come out ahead in money.
b. No, I expect to come out behind in money.
c. It doesn't matter. I expect to break even.

80. A university has 14 statistics classes scheduled for its Summer 2013 term. One class has space available for 30 students, eight classes have space for 60 students, one class has space for 70 students, and four classes have

space for 100 students.

a. What is the average class size assuming each class is filled to capacity?
b. Space is available for 980 students. Suppose that each class is filled to capacity and select a statistics student at random. Let the random variable X equal the size of the student's class. Define the PDF for X.
c. Find the mean of X.
d. Find the standard deviation of X.

81. In a raffle, there are 250 prizes of $5, 50 prizes of $25, and 10 prizes of $100. Assuming that 10,000 tickets are to be issued and sold, what is a fair price to charge to break even?

4.3 Binomial Distribution (Optional)

82. According to a recent article the average number of babies born with significant hearing loss (deafness) is approximately two per 1,000 babies in a healthy baby nursery. The number climbs to an average of 30 per 1,000 babies in an intensive care nursery.
Suppose that 1,000 babies from healthy baby nurseries were randomly surveyed. Find the probability that exactly two babies were born deaf.

Use the following information to answer the next four exercises: Recently, a nurse commented that when a patient calls the medical advice line claiming to have the flu, the chance that he or she truly has the flu (and not just a nasty cold) is only about 4 percent. Of the next 25 patients calling in claiming to have the flu, we are interested in how many actually have the flu.

83. Define the random variable and list its possible values.

84. State the distribution of X.

85. Find the probability that at least four of the 25 patients actually have the flu.

86. On average, for every 25 patients calling in, how many do you expect to have the flu?

87. People visiting video rental stores often rent more than one DVD at a time. The probability distribution for DVD rentals per customer at Video to Go is given Table 4.37. There is a five-video limit per customer at this store, so nobody ever rents more than five DVDs.

x	$P(x)$
0	.03
1	.50
2	.24
3	
4	.07
5	.04

Table 4.37

a. Describe the random variable X in words.
b. Find the probability that a customer rents three DVDs.
c. Find the probability that a customer rents at least four DVDs.
d. Find the probability that a customer rents at most two DVDs.

88. A school newspaper reporter decides to randomly survey 12 students to see if they will attend Tet (Vietnamese New Year) festivities this year. Based on past years, she knows that 18 percent of students attend Tet festivities. We are interested in the number of students who will attend the festivities.

a. In words, define the random variable X.
b. List the values that X may take on.

c. Give the distribution of X. X ~ _____(_____,_____)
d. How many of the 12 students do we expect to attend the festivities?
e. Find the probability that at most four students will attend.
f. Find the probability that more than two students will attend.

Use the following information to answer the next three exercises: The probability that a local hockey team will win any given game is 0.3694 based on a 13-year win history of 382 wins out of 1,034 games played (as of a certain date). An upcoming monthly schedule contains 12 games.

89. What is the expected number of wins for that upcoming month?
a. 1.67
b. 12
c. $\frac{382}{1043}$
d. 4.43

Let X = the number of games won in that upcoming month.

90. What is the probability that the team wins six games in that upcoming month?
a. .1476
b. .2336
c. .7664
d. .8903

91. What is the probability that the team wins at least five games in that upcoming month
a. .3694
b. .5266
c. .4734
d. .2305

92. A student takes a 10-question true-false quiz, but did not study and randomly guesses each answer. Find the probability that the student passes the quiz with a grade of at least 70 percent of the questions correct.

93. A student takes a 32-question multiple choice exam, but did not study and randomly guesses each answer. Each question has three possible choices for the answer. Find the probability that the student guesses **more than** 75 percent of the questions correctly.

94. Six different colored dice are rolled. Of interest is the number of dice that show a one.
a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ _____(_____,_____)
d. On average, how many dice would you expect to show a one?
e. Find the probability that all six dice show a one.
f. Is it more likely that three or that four dice will show a one? Use numbers to justify your answer numerically.

95. More than 96 percent of the very largest colleges and universities (more than 15,000 total enrollments) have some online offerings. Suppose you randomly pick 13 such institutions. We are interested in the number that offer distance learning courses.
a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ _____(_____,_____)
d. On average, how many schools would you expect to offer such courses?
e. Find the probability that at most 10 offer such courses.
f. Is it more likely that 12 or that 13 will offer such courses? Use numbers to justify your answer numerically and answer in a complete sentence.

96. Suppose that about 85 percent of graduating students attend their graduation. A group of 22 graduating students is randomly chosen.
a. In words, define the random variable X.
b. List the values that X may take on.

c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. How many are expected to attend their graduation?
e. Find the probability that 17 or 18 attend.
f. Based on numerical values, would you be surprised if all 22 attended graduation? Justify your answer numerically.

97. At the Fencing Center, 60 percent of the fencers use the foil as their main weapon. We randomly survey 25 fencers at the Fencing Center. We are interested in the number of fencers who do **not** use the foil as their main weapon.
a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. How many are expected to **not** to use the foil as their main weapon?
e. Find the probability that six do **not** use the foil as their main weapon.
f. Based on numerical values, would you be surprised if all 25 did **not** use foil as their main weapon? Justify your answer numerically.

98. Approximately 8 percent of students at a local high school participate in after-school sports all four years of high school. A group of 60 seniors is randomly chosen. Of interest is the number who participated in after-school sports all four years of high school.
a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. How many seniors are expected to have participated in after-school sports all four years of high school?
e. Based on numerical values, would you be surprised if none of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.
f. Based upon numerical values, is it more likely that four or that five of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.

99. The chance of an IRS audit for a tax return reporting more than $25,000 in income is about 2 percent per year. We are interested in the expected number of audits a person with that income has in a 20-year period. Assume each year is independent.
a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. How many audits are expected in a 20-year period?
e. Find the probability that a person is not audited at all.
f. Find the probability that a person is audited more than twice.

100. It has been estimated that only about 30 percent of California residents have adequate earthquake supplies. Suppose you randomly survey 11 California residents. We are interested in the number who have adequate earthquake supplies.
a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. What is the probability that at least eight have adequate earthquake supplies?
e. Is it more likely that none or that all of the residents surveyed will have adequate earthquake supplies? Why?
f. How many residents do you expect will have adequate earthquake supplies?

101. There are two similar games played for Chinese New Year and Vietnamese New Year. In the Chinese version, fair dice with numbers 1, 2, 3, 4, 5, and 6 are used, along with a board with those numbers. In the Vietnamese version, fair dice with pictures of a gourd, fish, rooster, crab, crayfish, and deer are used. The board has those six objects on it, also. We will play with bets being $1. The player places a bet on a number or object. The *house* rolls three dice. If none of the dice show the number or object that was bet, the house keeps the $1 bet. If one of the dice shows the number or object bet (and the other two do not show it), the player gets back his or her $1 bet, plus $1 profit. If two of the dice show the number or object bet (and the third die does not show it), the player gets back his or her $1 bet, plus $2 profit. If all three dice show the number or object bet, the player gets back his or her $1 bet, plus $3 profit. Let X = number of matches and Y = profit per game.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. List the values that Y may take on. Then, construct one PDF table that includes both X and Y and their probabilities.
e. Calculate the average expected matches over the long run of playing this game for the player.
f. Calculate the average expected earnings over the long run of playing this game for the player.
g. Determine who has the advantage, the player or the house.

102. According to the World Bank, only 9 percent of the population of Uganda had access to electricity as of 2009. Suppose we randomly sample 150 people in Uganda. Let X = the number of people who have access to electricity.
a. What is the probability distribution for X?
b. Using the formulas, calculate the mean and standard deviation of X.
c. Use your calculator to find the probability that 15 people in the sample have access to electricity.
d. Find the probability that at most 10 people in the sample have access to electricity.
e. Find the probability that more than 25 people in the sample have access to electricity.

103. The literacy rate for a nation measures the proportion of people age 15 and over who can read and write. The literacy rate in Afghanistan is 28.1 percent. Suppose you choose 15 people in Afghanistan at random. Let X = the number of people who are literate.
a. Sketch a graph of the probability distribution of X.
b. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.
c. Find the probability that more than five people in the sample are literate. Is it more likely that three people or four people are literate?

4.4 Geometric Distribution (Optional)

104. A consumer looking to buy a used red sports car will call dealerships until she finds a dealership that carries the car. She estimates the probability that any independent dealership will have the car will be 28 percent. We are interested in the number of dealerships she must call.
a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. On average, how many dealerships would we expect her to have to call until she finds one that has the car?
e. Find the probability that she must call at most four dealerships.
f. Find the probability that she must call three or four dealerships.

105. Suppose that the probability that an adult in America will watch the Super Bowl is 40 percent. Each person is considered independent. We are interested in the number of adults in America we must survey until we find one who will watch the Super Bowl.
a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. How many adults in America do you expect to survey until you find one who will watch the Super Bowl?
e. Find the probability that you must ask seven people.
f. Find the probability that you must ask three or four people.

106. It has been estimated that only about 30 percent of California residents have adequate earthquake supplies. Suppose we are interested in the number of California residents we must survey until we find a resident who does **not** have adequate earthquake supplies.
a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. $X \sim$ _____(_____,_____)
d. What is the probability that we must survey just one or two residents until we find a California resident who does not have adequate earthquake supplies?
e. What is the probability that we must survey at least three California residents until we find a California resident who does not have adequate earthquake supplies?

f. How many California residents do you expect to need to survey until you find a California resident who **does not** have adequate earthquake supplies?
g. How many California residents do you expect to need to survey until you find a California resident who **does** have adequate earthquake supplies?

107. In one of its spring catalogs, a retailer advertised footwear on 29 of its 192 catalog pages. Suppose we randomly survey 20 pages. We are interested in the number of pages that advertise footwear. Each page may be picked more than once.
a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(_____,_____)
d. How many pages do you expect to advertise footwear on them?
e. Is it probable that all 20 will advertise footwear on them? Why or why not?
f. What is the probability that fewer than 10 will advertise footwear on them?
g. Reminder: A page may be picked more than once. We are interested in the number of pages that we must randomly survey until we find one that has footwear advertised on it. Define the random variable *X* and give its distribution.
h. What is the probability that you only need to survey at most three pages in order to find one that advertises footwear on it?
i. How many pages do you expect to need to survey in order to find one that advertises footwear?

108. Suppose that you are performing the probability experiment of rolling one fair six-sided die. Let *F* be the event of rolling a four or a five. You are interested in how many times you need to roll the die to obtain the first four or five as the outcome.
- *p* = probability of success (event *F* occurs)
- *q* = probability of failure (event *F* does not occur)

a. Write the description of the random variable *X*.
b. What are the values that *X* can take on?
c. Find the values of *p* and *q*.
d. Find the probability that the first occurrence of event *F* (rolling a four or five) is on the second trial.

109. Ellen has music practice three days a week. She practices for all of the three days 85 percent of the time, two days 8 percent of the time, one day 4 percent of the time, and no days 3 percent of the time. One week is selected at random. What values does *X* take on?

110. Researchers investigate the prevalence of a particular infectious disease in countries around the world. According to their data, "Prevalence of this disease refers to the percentage of people ages 15 to 49 who are infected with it." In South Africa, the prevalence of this disease is 17.3 percent. Let *X* = the number of people you test until you find a person infected with this disease.
a. Sketch a graph of the distribution of the discrete random variable *X*.
b. What is the probability that you must test 30 people to find one with this disease?
c. What is the probability that you must ask 10 people?
d. Find the (i) mean and (ii) standard deviation of the distribution of *X*.

111. According to a recent poll, 75 percent of millennials (people born between 1981 and 1995) have a profile on a social networking site. Let *X* = the number of millennials you ask until you find a person without a profile on a social networking site.
a. Describe the distribution of *X*.
b. Find the (i) mean and (ii) standard deviation of *X*.
c. What is the probability that you must ask 10 people to find one person without a social networking site?
d. What is the probability that you must ask 20 people to find one person without a social networking site?
e. What is the probability that you must ask *at most* five people?

4.5 Hypergeometric Distribution (Optional)

112. A group of martial arts students is planning on participating in an upcoming demonstration. Six are students of tae kwon do, and seven are students of shotokan karate. Suppose that eight students are randomly picked to be in the first demonstration. We are interested in the number of shotokan karate students in that first demonstration.

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(_____,_____)
d. How many shotokan karate students do we expect to be in that first demonstration?

113. In one of its spring catalogs, a retailer advertised footwear on 29 of its 192 catalog pages. Suppose we randomly survey 20 pages. We are interested in the number of pages that advertise footwear. Each page may be picked at most once.
a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(_____,_____)
d. How many pages do you expect to advertise footwear on them?
e. Calculate the standard deviation.

114. Suppose that a technology task force is being formed to study technology awareness among instructors. Assume that 10 people will be randomly chosen to be on the committee from a group of 28 volunteers, 20 who are technically proficient and eight who are not. We are interested in the number on the committee who are **not** technically proficient.
a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(_____,_____)
d. How many instructors do you expect on the committee who are **not** technically proficient?
e. Find the probability that at least five on the committee are not technically proficient.
f. Find the probability that at most three on the committee are not technically proficient.

115. Suppose that nine Massachusetts athletes are scheduled to appear at a charity benefit. The nine are randomly chosen from eight volunteers from the local basketball team and four volunteers from the local football team. We are interested in the number of football players picked.
a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(_____,_____)
d. Are you choosing the nine athletes with or without replacement?

116. A bridge hand is defined as 13 cards selected at random and without replacement from a deck of 52 cards. In a standard deck of cards, there are 13 cards from each suit: hearts, spades, clubs, and diamonds. What is the probability of being dealt a hand that does not contain a heart?
a. What is the group of interest?
b. How many are in the group of interest?
c. How many are in the other group?
d. Let *X* = _____. What values does *X* take on?
e. The probability question is *P*(_____).
f. Find the probability in question.
g. Find the (i) mean and (ii) standard deviation of *X*.

4.6 Poisson Distribution (Optional)

117. The switchboard in a Minneapolis law office gets an average of 5.5 incoming phone calls during the noon hour on Mondays. Experience shows that the existing staff can handle up to six calls in an hour. Let *X* = the number of calls received at noon.
a. Find the mean and standard deviation of *X*.
b. What is the probability that the office receives at most six calls at noon on Monday?
c. Find the probability that the law office receives six calls at noon. What does this mean to the law office staff who get, on average, 5.5 incoming phone calls at noon?
d. What is the probability that the office receives more than eight calls at noon?

118. The maternity ward at a hospital in the Philippines is one of the busiest in the world with an average of 60 births per day. Let *X* = the number of births in an hour.
a. Find the mean and standard deviation of *X*.

b. Sketch a graph of the probability distribution of *X*.
c. What is the probability that the maternity ward will deliver three babies in one hour?
d. What is the probability that the maternity ward will deliver at most three babies in one hour?
e. What is the probability that the maternity ward will deliver more than five babies in one hour?

119. A manufacturer of decorative string lights knows that 3 percent of its bulbs are defective. Using both the binomial and Poisson distributions, find the probability that a string of 100 lights contains at most four defective bulbs.

120. The average number of children a Japanese woman has in her lifetime is 1.37. Suppose that one Japanese woman is randomly chosen.
a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(_____,_____)
d. Find the probability that she has no children.
e. Find the probability that she has fewer children than the Japanese average.
f. Find the probability that she has more children than the Japanese average.

121. The average number of children a Spanish woman has in her lifetime is 1.47. Suppose that one Spanish woman is randomly chosen.
a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(_____,_____)
d. Find the probability that she has no children.
e. Find the probability that she has fewer children than the Spanish average.
f. Find the probability that she has more children than the Spanish average.

122. Fertile, female cats produce an average of three litters per year. Suppose that one fertile, female cat is randomly chosen. Answer the questions about the cat's probability of litters in one year.
a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____
d. Find the probability that she has no litters in one year.
e. Find the probability that she has at least two litters in one year.
f. Find the probability that she has exactly three litters in one year.

123. The chance of having an extra fortune in a fortune cookie is about 3 percent. Given a bag of 144 fortune cookies, we are interested in the number of cookies with an extra fortune. Two distributions may be used to solve this problem, but only use one distribution to solve the problem.
a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(_____,_____)
d. How many cookies do we expect to have an extra fortune?
e. Find the probability that none of the cookies have an extra fortune.
f. Find the probability that more than three have an extra fortune.
g. As *n* increases, what happens involving the probabilities using the two distributions? Explain in complete sentences.

124. According to the South Carolina Department of Mental Health website, for every 200 U.S. women, the average number who suffer from a particular disease is one. Out of a randomly chosen group of 600 U.S. women. Determine the following:
a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(_____,_____)
d. How many are expected to suffer from this disease?
e. Find the probability that no one suffers from this disease.
f. Find the probability that more than four suffer from this disease.

125. The chance of an IRS audit for a tax return reporting more than $25,000 in income is about 2 percent per year.

Suppose that 100 people with tax returns over $25,000 are randomly picked. We are interested in the number of people audited in one year. Use a Poisson distribution to anwer the following questions.

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(_____,_____)
d. How many are expected to be audited?
e. Find the probability that no one was audited.
f. Find the probability that at least three were audited.

126. Approximately 8 percent of students at a local high school participate in after-school sports all four years of high school. A group of 60 seniors is randomly chosen. Of interest is the number who participated in after-school sports all four years of high school.

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(_____,_____)
d. How many seniors are expected to have participated in after-school sports all four years of high school?
e. Based on numerical values, would you be surprised if none of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.
f. Based on numerical values, is it more likely that four or that five of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.

127. On average, Pierre, an amateur chef, drops three pieces of eggshell into every two cake batters he makes. Suppose that you buy one of his cakes.

a. In words, define the random variable *X*.
b. List the values that *X* may take on.
c. Give the distribution of *X*. *X* ~ _____(_____,_____)
d. On average, how many pieces of eggshell do you expect to be in the cake?
e. What is the probability that there will not be any pieces of eggshell in the cake?
f. Let's say that you buy one of Pierre's cakes each week for six weeks. What is the probability that there will not be any eggshell in any of the cakes?
g. Based upon the average given for Pierre, is it possible for there to be seven pieces of shell in the cake? Why?

Use the following information to answer the next two exercises: The average number of times per week that Mrs. Plum's cats wake her up at night because they want to play is 10. We are interested in the number of times her cats wake her up each week.

128. In words, what is the random variable *X*?

a. the number of times Mrs. Plum's cats wake her up each week
b. the number of times Mrs. Plum's cats wake her up each hour
c. the number of times Mrs. Plum's cats wake her up each night
d. the number of times Mrs. Plum's cats wake her up

129. Find the probability that her cats will wake her up no more than five times next week.

a. .5000
b. .9329
c. .0378
d. .0671

4.7 Discrete Distribution (Playing Card Experiment)

130. Use a programmable calculator to simulate a binomial distribution.

a. How would you use the randInt function to simulate the number of successes in five trials of an experiment with two outcomes, each of which has a .5 probability of occurring?
b. Use the randInt function to simulate 10 observations of the random variable in Part A.
c. Find the sample mean and sample standard deviation.
d. Compare the sample mean and sample standard deviation to the theoretical mean and the theoretical standard deviation.

References

4.2 Mean or Expected Value and Standard Deviation

Florida State University. (n.d.). *Class catalogue at the Florida State University*. Retrieved from https://apps.oti.fsu.edu/RegistrarCourseLookup/SearchFormLegacy

World Earthquakes. (2012). *World earthquakes: Live earthquake news and highlights*. Retrieved from http://www.worldearthquakes.com/index.php?option=ethq_prediction

4.3 Binomial Distribution (Optional)

American Cancer Society. (2013). *What are the key statistics about pancreatic cancer?* Retrieved from http://www.cancer.org/cancer/pancreaticcancer/detailedguide/pancreatic-cancer-key-statistics

Central Intelligence Agency. (n.d.). *The world factbook*. Retrieved from https://www.cia.gov/library/publications/theworld-factbook/geos/af.html

ESPN NBA. (2013). NBA statistics – 2013. Retrieved from http://espn.go.com/nba/statistics/_/seasontype/2

Newport, F. (2013). Americans still enjoy saving rather than spending: Few demographic differences seen in these views other than by income. *GALLUP Economy*. Retrieved from http://www.gallup.com/poll/162368/americansenjoy-saving-rather-spending.aspx

Pryor, J. H., et al. (2011). *The American freshman: National norms fall 2011*. Los Angeles, CA: Cooperative Institutional Research Program, Higher Education Research Institute. Retrieved from http://heri.ucla.edu/PDFs/pubs/TFS/Norms/Monographs/TheAmericanFreshman2011.pdf

Wikipedia. (n.d.). *Distance education*. Retrieved from http://en.wikipedia.org/wiki/Distance_education

World Bank Group. (2013). *Access to electricity (% of population)*. Retrieved from http://data.worldbank.org/indicator/EG.ELC.ACCS.ZS?order=wbapi_data_value_2009%20wbapi_data_value%20wbapi_data_value-first&sort=asc

4.4 Geometric Distribution (Optional)

Central Intelligence Agency. (n.d.). The world factbook. Retrieved from https://www.cia.gov/library/publications/theworld-factbook/geos/af.html

Pew Research Center. (n.d.). Millennials: A portrait of generation next. Retrieved from http://www.pewsocialtrends.org/files/2010/10/millennials-confident-connected-open-to-change.pdf

Pew Research. (2013). Millennials: confident. Executive Summary: Pew Research Social & Demographic Trends. Retrieved from http://www.pewsocialtrends.org/2010/02/24/millennials-confident-connected-open-tochange/

Pryor, J. H., et al. (2011). *The American freshman: National norms fall 2011*. Los Angeles: Cooperative Institutional Research Program, Higher Education Research Institute. Retrieved from http://heri.ucla.edu/PDFs/pubs/TFS/Norms/Monographs/ TheAmericanFreshman2011.pdf

The European Union and ICON-Institute. (2007/8). *Summary of the national risk and vulnerability assessment 2007/8: A profile of Afghanistan*. Retrieved from http://ec.europa.eu/europeaid/where/asia/documents/afgh_brochure_summary_en.pdf

The World Bank. (2013). Prevalence of HIV, total (% of populations ages 15-49). Retrieved from http://data.worldbank.org/indicator/SH.DYN.AIDS.ZS?order=wbapi_data_value_2011+wbapi_data_value+wbapi_data_value-last&sort=desc

UNICEF Television. (n.d.). *UNICEF reports on female literacy centers in Afghanistan established to teach women and girls basic reading and writing skills*. (Video). Retrieved from http://www.unicefusa.org/assets/video/afghan-femaleliteracy-centers.html

4.6 Poisson Distribution (Optional)

Centers for Disease Control and Prevention. (2012, Oct. 2). *Teen drivers: Get the facts*. Retrieved from http://www.cdc.gov/Motorvehiclesafety/Teen_Drivers/teendrivers_factsheet.html

Daily Mail. (2011, June 9). One born every minute: the maternity unit where mothers are THREE to a bed. Retrieved from http://www.dailymail.co.uk/news/article-2001422/Busiest-maternity-ward-planet-averages-60-babies-day-mothersbed.html

Department of Aviation at the Hartsfield-Jackson Atlanta International Airport. (2013). *ATL fact sheet*. Retrieved from

http://www.atlanta-airport.com/Airport/ATL/ATL_FactSheet.aspx

Lenhart, A. (2012). Teens, smartphones & testing: Texting volume is up while the frequency of voice calling is down. About one in four teens say they own smartphones. Pew Internet. Retrieved from http://www.pewinternet.org/~/media/Files/Reports/2012/PIP_Teens_Smartphones_and_Texting.pdf

Ministry of Health, Labour, and Welfare. (n.d.). *Children and childrearing*. Retrieved from http://www.mhlw.go.jp/english/policy/children/children-childrearing/index.html

Pew Internet. (2013). *How Americans use text messaging*. Retrieved from http://pewinternet.org/Reports/2011/Cell-Phone-Texting-2011/Main-Report.aspx

South Carolina Department of Mental Health. (2006). *Eating disorder statistics*. Retrieved from http://www.state.sc.us/dmh/anorexia/statistics.htm

The Guardian. (2011, June 8). Giving birth in Manila: The maternity ward at the Dr Jose Fabella Memorial Hospital in Manila, the busiest in the Philippines, where there is an average of 60 births a day. Retrieved from http://www.theguardian.com/world/gallery/2011/jun/08/philippines-health#/?picture=375471900&index=2

Vanderkam, L. (2012, Oct. 8). Stop checking your email, now. *CNNMoney*. Retrieved from http://management.fortune.cnn.com/2012/10/08/stop-checking-your-email-now/

World Earthquakes. (2012). *World earthquakes: Live earthquake news and highlights*. Retrieved from http://www.worldearthquakes.com/index.php?option=ethq_prediction

Solutions

1.

x	$P(x)$
0	.12
1	.18
2	.30
3	.15
4	.10
5	.10
6	.05

Table 4.38

3. .10 + .05 = .15

5. 1

7. .35 + .40 + .10 = .85

9. 1(.15) + 2(.35) + 3(.40) + 4(.10) = .15 + .70 + 1.20 + .40 = 2.45

11.

x	*P*(*x*)
0	.03
1	.04
2	.08
3	.85

Table 4.39

13. Let X = the number of events Javier volunteers for each month.

15.

x	*P*(*x*)
0	.05
1	.05
2	.10
3	.20
4	.25
5	.35

Table 4.40

17. 1 – .05 = .95

19. .2 + 1.2 + 2.4 + 1.6 = 5.4

21. The values of $P(x)$ do not sum to one.

23. Let X = the number of years a physics major will spend doing postgraduate research.

25. 1 – .35 – .20 – .15 – .10 – .05 = .15

27. 1(.35) + 2(.20) + 3(.15) + 4(.15) + 5(.10) + 6(.05) = .35 + .40 + .45 + .60 + .50 + .30 = 2.6 years

29. X is the number of years a student studies ballet with the teacher.

31. .10 + .05 + .10 = .25

33. The sum of the probabilities sum to one because it is a probability distribution.

35. $-2\left(\frac{40}{52}\right) + 30\left(\frac{12}{52}\right) = -1.54 + 6.92 = 5.38$

37. X = the number that reply *yes*

39. 0, 1, 2, 3, 4, 5, 6, 7, 8

41. 5.7

43. .4151

45. X = the number of freshmen selected from the study until one replied *yes* to the law that was passed.

47. 1,2,...

49. 1.4

51. X = the number of business majors in the sample.

53. 2, 3, 4, 5, 6, 7, 8, 9

55. 6.26

57. 0, 1, 2, 3, 4, ...

59. .0485

61. .0214

63. X = the number of United States teens who die from motor vehicle injuries per day.

65. 0, 1, 2, 3, 4, ...

67. no

71. The variable of interest is X, or the gain or loss, in dollars.
The face cards jack, queen, and king. There are (3)(4) = 12 face cards and 52 – 12 = 40 cards that are not face cards. We first need to construct the probability distribution for X. We use the card and coin events to determine the probability for each outcome, but we use the monetary value of X to determine the expected value.

Card Event	X net gain/loss	$P(X)$
Face Card and Heads	6	$\left(\frac{12}{52}\right)\left(\frac{1}{2}\right)=\left(\frac{6}{52}\right)$
Face Card and Tails	2	$\left(\frac{12}{52}\right)\left(\frac{1}{2}\right)=\left(\frac{6}{52}\right)$
(Not Face Card) and (H or T)	–2	$\left(\frac{40}{52}\right)(1)=\left(\frac{40}{52}\right)$

Table 4.41

- $\text{Expected value} = (6)\left(\frac{6}{52}\right)+(2)\left(\frac{6}{52}\right)+(-2)\left(\frac{40}{52}\right)=-\frac{32}{52}$
- Expected value = –$0.62, rounded to the nearest cent
- If you play this game repeatedly, over a long string of games, you would expect to lose 62 cents per game, on average.
- You should not play this game to win money because the expected value indicates an expected average loss.

73. a. .1
b. 1.6

75. a.

Software Company	
x	***P*(*x*)**
5,000,000	.10
1,000,000	.30
–1,000,000	.60

Table 4.42

Hardware Company	
x	***P*(*x*)**
3,000,000	.20
1,000,000	.40
–1,000,00	.40

Table 4.43

Biotech Firm	
x	***P*(*x*)**
6,000,000	.10
0	.70
–1,000,000	.20

Table 4.44

b. $200,000; $600,000; $400,000
c. third investment because it has the lowest probability of loss
d. first investment because it has the highest probability of loss
e. second investment

77. 4.85 years

79. b

81. Let X = the amount of money to be won on a ticket. The following table shows the PDF for X:

x	$P(x)$
0	.969
5	$\frac{250}{10{,}000}$ = .025
25	$\frac{50}{10{,}000}$ = .005
100	$\frac{10}{10{,}000}$ = .001

Table 4.45

Calculate the expected value of X.
0(.969) + 5(.025) + 25(.005) + 100(.001) = .35
A fair price for a ticket is $0.35. Any price over $0.35 will enable the lottery to raise money.

83. X = the number of patients calling in claiming to have the flu, who actually have the flu.
X = 0, 1, 2, ...25

85. .0165

87. a. X = the number of DVDs a Video to Go customer rents
b. .12
c. .11
d. .77

89. d. 4.43

91. c

93.
- X = number of questions answered correctly
- $X \sim B\left(32, \frac{1}{3}\right)$
- We are interested in MORE THAN 75 percent of 32 questions correct. 75 percent of 32 is 24. We want to find $P(x > 24)$. The event *more than 24* is the complement of *less than or equal to 24*.
- Using your calculator's distribution menu: 1 – binomcdf$\left(32, \frac{1}{3}, 24\right)$
- $P(x > 24) = 0$
- The probability of getting more than 75 percent of the 32 questions correct when randomly guessing is very small and practically zero.

95. a. X = the number of college and universities that offer online offerings.
b. 0, 1, 2, ..., 13
c. $X \sim B(13, 0.96)$
d. 12.48
e. .0135
f. $P(x = 12)$ = .3186 $P(x = 13)$ = 0.5882 More likely to get 13.

97. a. X = the number of fencers who do *not* use the foil as their main weapon
b. 0, 1, 2, 3,... 25
c. $X \sim B(25,.40)$
d. 10
e. .0442
f. The probability that all 25 not use the foil is almost zero. Therefore, it would be very surprising.

99. a. X = the number of audits in a 20-year period
b. 0, 1, 2, ..., 20
c. $X \sim B(20, .02)$
d. .4
e. .6676
f. .0071

101. 1. X = the number of matches
2. 0, 1, 2, 3
3. $X \sim B\left(3, \frac{1}{6}\right)$
4. In dollars: −1, 1, 2, 3
5. $\frac{1}{2}$
6. Multiply each Y value by the corresponding X probability from the PDF table. The answer is −.0787. You lose about eight cents, on average, per game.
7. The house has the advantage.

103. a. $X \sim B(15, .281)$

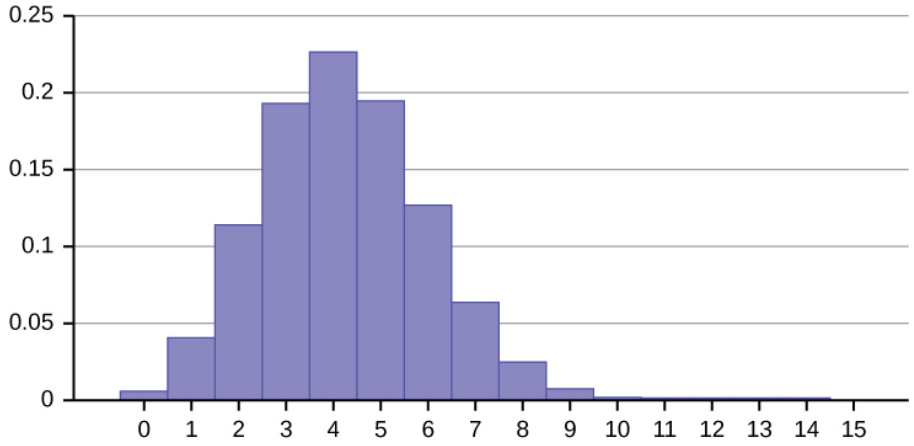


Figure 4.10

b. i. Mean = $\mu = np = 15(.281) = 4.215$
ii. Standard Deviation = $\sigma = \sqrt{npq} = \sqrt{15(.281)(.719)} = 1.7409$

c. $P(x > 5) = 1 - P(x \leq 5) = 1 - \text{binomcdf}(15, .281, 5) = 1 - 0.7754 = .2246$
$P(x = 3) = \text{binompdf}(15, .281, 3) = .1927$
$P(x = 4) = \text{binompdf}(15, .281, 4) = .2259$
It is more likely that four people are literate than three people are.

105. a. X = the number of adults in America who are surveyed until one says he or she will watch the Super Bowl.
b. $X \sim G(.40)$
c. 2.5
d. .0187
e. .2304

107.
a. X = the number of pages that advertise footwear
b. X takes on the values 0, 1, 2, ..., 20
c. $X \sim B(20, \frac{29}{192})$
d. 3.02
e. no
f. .9997
g. X = the number of pages we must survey until we find one that advertises footwear. $X \sim G(\frac{29}{192})$

h. .3881
i. 6.6207 pages

109. 0, 1, 2, and 3

111. a. $X \sim G(.25)$
b. i. mean = $\mu = \frac{1}{p} = \frac{1}{0.25} = 4$
ii. standard deviation = $\sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-.25}{.25^2}} \approx 3.4641$
c. $P(x = 10)$ = geometpdf(.25, 10) = .0188
d. $P(x = 20)$ = geometpdf(.25, 20) = .0011
e. $P(x \leq 5)$ = geometcdf(.25, 5) = .7627

113. a. X = the number of pages that advertise footwear
b. 0, 1, 2, 3, ..., 20
c. $X \sim H(29, 163, 20)$, $r = 29$, $b = 163$, $n = 20$
d. 3.03
e. 1.5197

115. a. X = the number of Patriots picked
b. 0, 1, 2, 3, 4
c. $X \sim H(4, 8, 9)$
d. without replacement

117. a. $X \sim P(5.5)$; $\mu = 5.5$; $\sigma = \sqrt{5.5} \approx 2.3452$
b. $P(x \leq 6)$ = poissoncdf(5.5, 6) ≈ .6860
c. There is a 15.7 percent probability that the law staff will receive more calls than they can handle.
d. $P(x > 8) = 1 - P(x \leq 8) = 1 -$ poissoncdf(5.5, 8) ≈ 1 – .8944 = .1056

119. Let X = the number of defective bulbs in a string.
Using the Poisson distribution:
- $\mu = np = 100(.03) = 3$
- $X \sim P(3)$
- $P(x \leq 4)$ = poissoncdf(3, 4) ≈ .8153

Using the binomial distribution
- $X \sim B(100, .03)$
- $P(x \leq 4)$ = binomcdf(100, .03, 4) ≈ .8179

The Poisson approximation is very good—the difference between the probabilities is only .0026.

121. a. X = the number of children for a Spanish woman
b. 0, 1, 2, 3,...
c. $X \sim P(1.47)$
d. .2299
e. .5679
f. .4321

123. a. X = the number of fortune cookies that have an extra fortune
b. 0, 1, 2, 3,... 144
c. $X \sim B(144, .03)$ or $P(4.32)$
d. 4.32
e. .0124 or .0133
f. .6300 or .6264

g. As *n* gets larger, the probabilities get closer together.

125. a. *X* = the number of people audited in one year
b. 0, 1, 2, ..., 100
c. *X* ~ *P*(2)
d. 2
e. .1353
f. .3233

127. a. *X* = the number of shell pieces in one cake
b. 0, 1, 2, 3,...
c. *X* ~ *P*(1.5)
d. 1.5
e. .2231
f. .0001
g. yes

129. d

130. a. You can use randInt (0,1,5) to generate five trials of the experiment. Count the number of 1's generated to determine the number of successes.
b. Student answers may vary.
c. Student answers may vary.
d. The theoretical mean is $(5)(.5) = 2.5$. The theoretical standard deviation is $\sqrt{(5)(.5)(0.5)} = \sqrt{1.25}$.

5 CONTINUOUS RANDOM VARIABLES

Figure 5.1 The heights of these radish plants are continuous random variables. (credit: Rev Stan)

Chapter Objectives

By the end of this chapter, the student should be able to do the following:

- Recognize and understand continuous probability density functions in general
- Recognize the uniform probability distribution and apply it appropriately
- Recognize the exponential probability distribution and apply it appropriately

Introduction

Continuous random variables have many applications. Baseball batting averages, IQ scores, the length of time a long-distance telephone call lasts, the amount of money a person carries, the length of time a computer chip lasts, and SAT scores are just a few. The field of reliability depends on a variety of continuous random variables.

Note

The values of discrete and continuous random variables can be ambiguous. For example, if X is equal to the number of miles (to the nearest mile) you drive to work, then X is a discrete random variable. You count the miles. If X is the distance you drive to work, then you measure values of X and X is a continuous random variable. For a second example, if X is equal to the number of books in a backpack, then X is a discrete random variable. If X is the weight of a book, then X is a continuous random variable because weights are measured. How the random variable is defined is very important.

Properties of Continuous Probability Distributions

The graph of a continuous probability distribution is a curve. Probability is represented by the area under the curve.

The curve is called the **probability density function** (abbreviated as **pdf**). We use the symbol $f(x)$ to represent the curve. $f(x)$ is the function that corresponds to the graph; we use the density function $f(x)$ to draw the graph of the probability distribution.

Area under the curve is given by a different function called the **cumulative distribution function** (abbreviated as **cdf**). The cumulative distribution function is used to evaluate probability as area.

- The outcomes are measured, not counted.
- The entire area under the curve and above the x-axis is equal to one.
- Probability is found for intervals of x values rather than for individual x values.
- $P(c < x < d)$ is the probability that the random variable X is in the interval between the values c and d. $P(c < x < d)$ is the area under the curve, above the x-axis, to the right of c and the left of d.
- $P(x = c) = 0$ The probability that x takes on any single individual value is zero. The area below the curve, above the x-axis, and between $x = c$ and $x = c$ has no width, and therefore no area (area = 0). Since the probability is equal to the area, the probability is also zero.
- $P(c < x < d)$ is the same as $P(c \leq x \leq d)$ because probability is equal to area.

We will find the area that represents probability by using geometry, formulas, technology, or probability tables. In general, calculus is needed to find the area under the curve for many probability density functions. When we use formulas to find the area in this textbook, we are using formulas that were found by using the techniques of integral calculus. However, because most students taking this course have not studied calculus, we will not be using calculus in this textbook.

There are many continuous probability distributions. When probability is modeled by use of a continuous probability distribution, the distribution used is selected to model and fit the particular situation in the best way.

In this chapter and the next, we will study the uniform distribution, the exponential distribution, and the normal distribution. The following graphs illustrate these distributions:

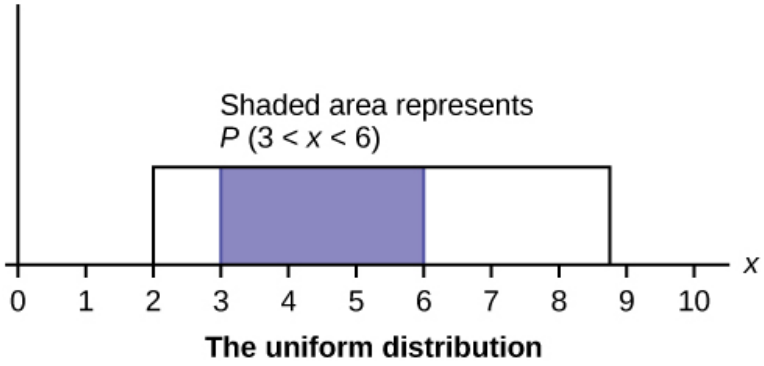


Figure 5.2 The graph shows a uniform distribution with the area between $x = 3$ and $x = 6$ shaded to represent the probability that the value of the random variable X is in the interval between three and six.

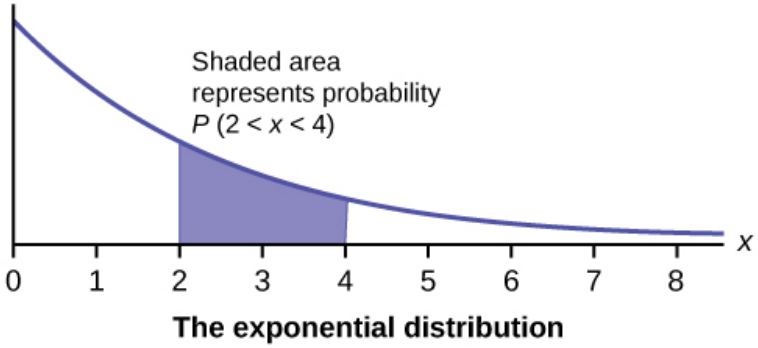


Figure 5.3 The graph shows an exponential distribution with the area between $x = 2$ and $x = 4$ shaded to represent the probability that the value of the random variable X is in the interval between two and four.

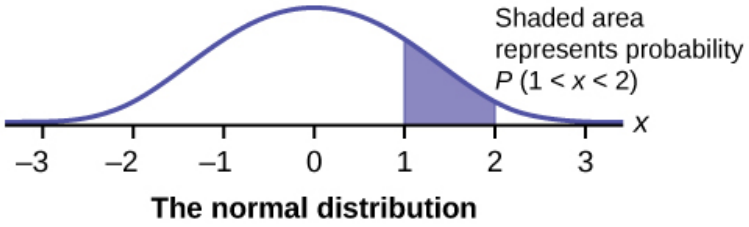


Figure 5.4 The graph shows the standard normal distribution with the area between $x = 1$ and $x = 2$ shaded to represent the probability that the value of the random variable X is in the interval between one and two.

5.1 Continuous Probability Functions

We begin by defining a continuous probability density function. We use the function notation $f(x)$. Intermediate algebra may have been your first formal introduction to functions. In the study of probability, the functions we study are special. We define the function $f(x)$ so that the area between it and the x-axis is equal to a probability. Since the maximum probability is one, the maximum area is also one. **For continuous probability distributions, PROBABILITY = AREA.**

EXAMPLE 5.1

Consider the function $f(x) = \frac{1}{20}$ for $0 \leq x \leq 20$. x = a real number. The graph of $f(x) = \frac{1}{20}$ is a horizontal line. However, since $0 \leq x \leq 20$, $f(x)$ is restricted to the portion between $x = 0$ and $x = 20$, inclusive.

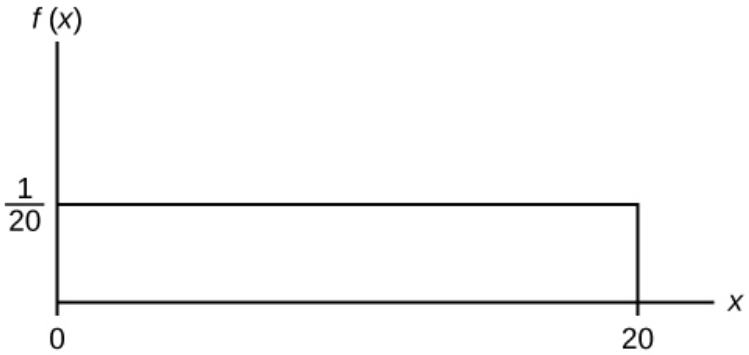


Figure 5.5

$f(x) = \frac{1}{20}$ **for** $0 \leq x \leq 20$.

The graph of $f(x) = \frac{1}{20}$ is a horizontal line segment when $0 \leq x \leq 20$.

The area between $f(x) = \frac{1}{20}$ where $0 \leq x \leq 20$ and the x-axis is the area of a rectangle with base = 20 and height = $\frac{1}{20}$.

$$\text{AREA} = 20\left(\frac{1}{20}\right) = 1$$

Suppose we want to find the area between $f(x) = \frac{1}{20}$ and the x-axis where $0 < x < 2$.

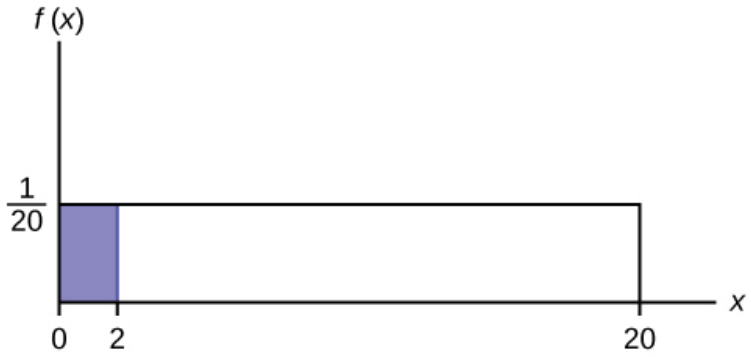


Figure 5.6

$$\text{AREA} = (2-0)\left(\frac{1}{20}\right) = 0.1$$

$(2-0) = 2 =$ base of a rectangle

Reminder

area of a rectangle = (base)(height)

The area corresponds to a probability. The probability that x is between zero and two is 0.1, which can be written mathematically as $P(0 < x < 2) = P(x < 2) = 0.1$.

Suppose we want to find the area between $f(x) = \frac{1}{20}$ and the x-axis where $4 < x < 15$.

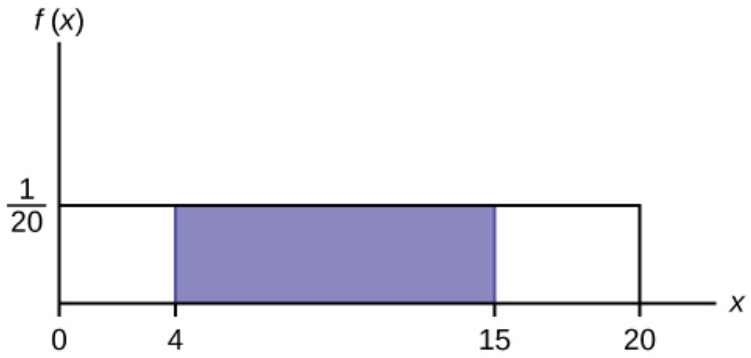


Figure 5.7

$$\text{AREA} = (15-4)\left(\tfrac{1}{20}\right) = 0.55$$

$(15-4) = 11 =$ the base of a rectangle

The area corresponds to the probability $P(4 < x < 15) = 0.55$.

Suppose we want to find $P(x = 15)$. On an x-y graph, $x = 15$ is a vertical line. A vertical line has no width (or zero width). Therefore, $P(x = 15) =$ (base)(height) $= (0)\left(\frac{1}{20}\right) = 0$

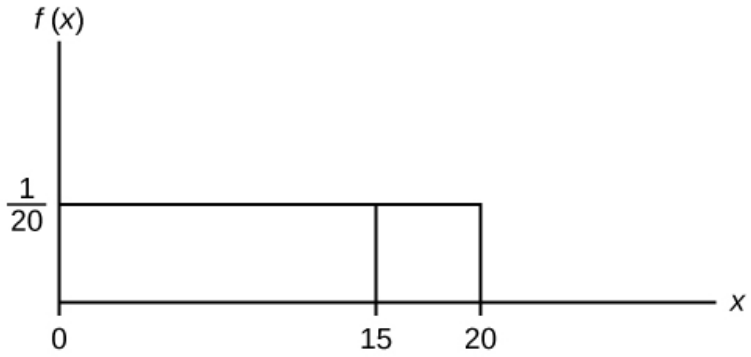


Figure 5.8

$P(X <= x)$, which can also be written as $P(X < x)$ for continuous distributions, is called the **cumulative distribution function** or CDF. Notice the *less than or equal to* symbol. We can also use the CDF to calculate $P(X > x)$. The CDF gives *area to the left* and $P(X > x)$ gives *area to the right*. We calculate $P(X > x)$ for continuous distributions as follows: $P(X > x) = 1 - P(X < x)$.

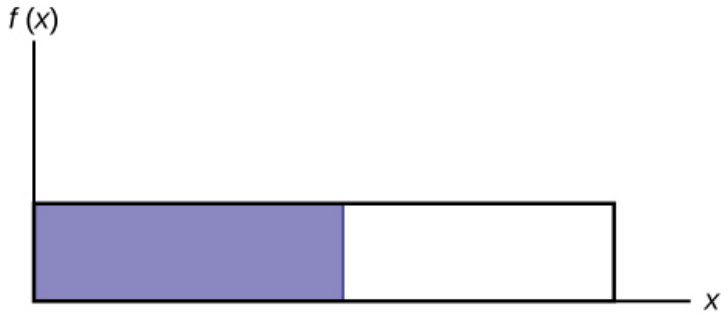


Figure 5.9

Label the graph with $f(x)$ and x. Scale the x and y axes with the maximum x and y values. $f(x) = \frac{1}{20}$, $0 \leq x \leq 20$.

To calculate the probability that x is between two values, look at the following graph. Shade the region between $x = 2.3$ and $x = 12.7$. Then calculate the shaded area of a rectangle.

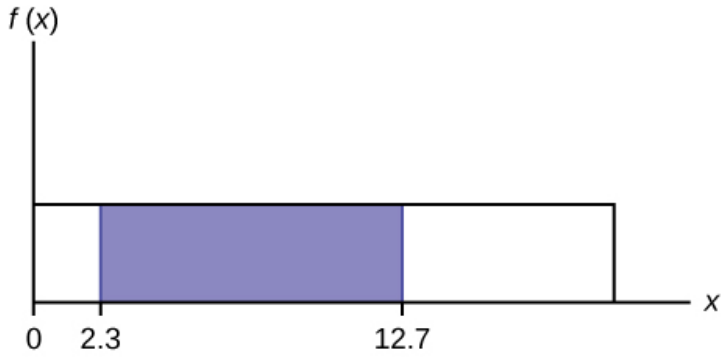


Figure 5.10

$$P(2.3 < x < 12.7) = (\text{base})(\text{height}) = (12.7 - 2.3)\left(\frac{1}{20}\right) = 0.52$$

TRY IT 5.1

Consider the function $f(x) = \frac{1}{8}$ for $0 \leq x \leq 8$. Draw the graph of $f(x)$ and find $P(2.5 < x < 7.5)$.

5.2 The Uniform Distribution

The uniform distribution is a continuous probability distribution and is concerned with events that are equally likely to occur. When working out problems that have a uniform distribution, be careful to note if the data are inclusive or exclusive of endpoints.

EXAMPLE 5.2

The data in Table 5.1 are 55 smiling times, in seconds, of an eight-week-old baby.

10.4	19.6	18.8	13.9	17.8	16.8	21.6	17.9	12.5	11.1	4.9
12.8	14.8	22.8	20.0	15.9	16.3	13.4	17.1	14.5	19.0	22.8
1.3	0.7	8.9	11.9	10.9	7.3	5.9	3.7	17.9	19.2	9.8
5.8	6.9	2.6	5.8	21.7	11.8	3.4	2.1	4.5	6.3	10.7
8.9	9.4	9.4	7.6	10.0	3.3	6.7	7.8	11.6	13.8	18.6

Table 5.1

The sample mean = 11.49 and the sample standard deviation = 6.23.

We will assume that the smiling times, in seconds, follow a uniform distribution between zero and 23 seconds, inclusive. This means that any smiling time from zero to and including 23 seconds is *equally likely*. The histogram that could be constructed from the sample is an empirical distribution that closely matches the theoretical uniform distribution.

Let X = length, in seconds, of an eight-week-old baby's smile.

The notation for the uniform distribution is

$X \sim U(a, b)$ where a = the lowest value of x and b = the highest value of x.

The probability density function is $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$.

For this example, $X \sim U(0, 23)$ and $f(x) = \frac{1}{23-0}$ for $0 \leq X \leq 23$.

Formulas for the theoretical mean and standard deviation are

$$\mu = \frac{a+b}{2} \text{ and } \sigma = \sqrt{\frac{(b-a)^2}{12}}$$

For this problem, the theoretical mean and standard deviation are

$$\mu = \frac{0 + 23}{2} = 11.50 \text{ seconds and } \sigma = \sqrt{\frac{(23 - 0)^2}{12}} = 6.64 \text{ seconds.}$$

Notice that the theoretical mean and standard deviation are close to the sample mean and standard deviation in this example.

TRY IT 5.2

The data that follow are the number of passengers on 35 different charter fishing boats. The sample mean = 7.9 and the sample standard deviation = 4.33. The data follow a uniform distribution where all values between and including zero and 14 are equally likely. State the values of a and b. Write the distribution in proper notation, and calculate the theoretical mean and standard deviation.

1	12	4	10	4	14	11
7	11	4	13	2	4	6
3	10	0	12	6	9	10
5	13	4	10	14	12	11
6	10	11	0	11	13	2

Table 5.2

EXAMPLE 5.3

Problem

a. Refer to Example 5.2. What is the probability that a randomly chosen eight-week-old baby smiles between two and 18 seconds?

Solution

$P(2 < x < 18)$ = (base)(height) = $(18 - 2)\left(\frac{1}{23}\right) = \frac{16}{23}$

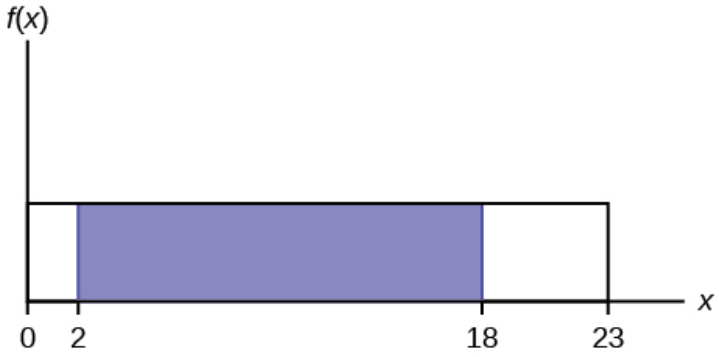


Figure 5.11

Problem

b. Find the 90th percentile for an eight-week-old baby's smiling time.

Solution

b. Ninety percent of the smiling times fall below the 90th percentile, k, so $P(x < k) = 0.90$.

$$P(x < k) = 0.90$$

$$\text{(base)(height)} = 0.90$$

$$(k - 0)\left(\frac{1}{23}\right) = 0.90$$

$$k = (23)(0.90) = 20.7$$

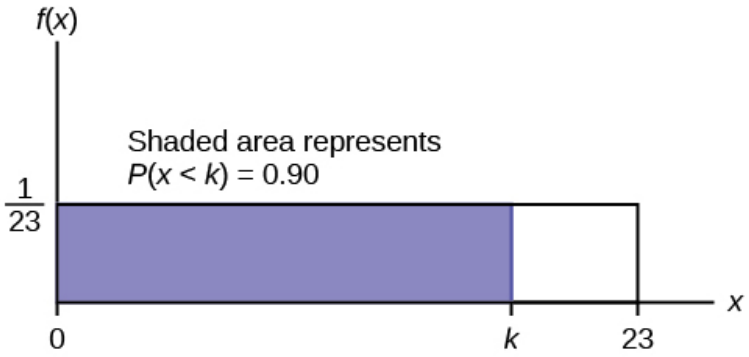


Figure 5.12

Problem

c. Find the probability that a random eight-week-old baby smiles more than 12 seconds *knowing* that the baby smiles *more than eight seconds*.

Solution

c. This probability question is a **conditional**. You are asked to find the probability that an eight-week-old baby smiles more than 12 seconds when you *already know* the baby has smiled for more than eight seconds.

Find $P(x > 12 \mid x > 8)$ There are two ways to do the problem. *For the first way*, use the fact that this is a conditional and changes the sample space. The graph illustrates the new sample space. You already know the baby smiled more than eight seconds.

for $8 < x < 23$

Write a new f(x): $f(x) = \frac{1}{23 - 8} = \frac{1}{15}$ for $8 < x < 23$.

$P(x > 12 \mid x > 8) = (23 - 12)\left(\frac{1}{15}\right) = \frac{11}{15}$

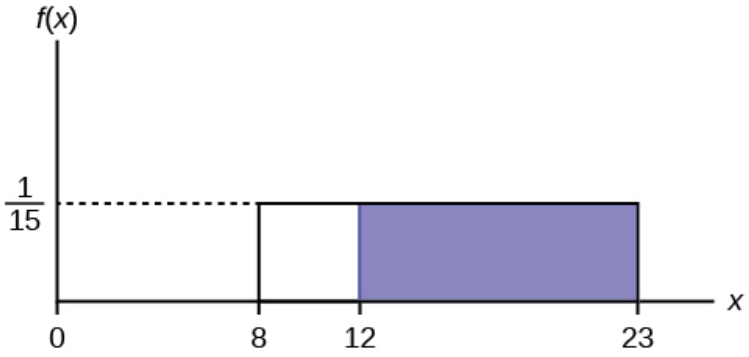


Figure 5.13

For the second way, use the conditional formula from Probability Topics with the original distribution.

$P(A \mid B) = \frac{P(A \text{ AND } B)}{P(B)}$

For this problem, A is $(x > 12)$ and B is $(x > 8)$.

So, $P(x > 12 \mid x > 8) = \frac{(x>12 \text{ AND } x>8)}{P(x>8)} = \frac{P(x>12)}{P(x>8)} = \frac{\frac{11}{23}}{\frac{15}{23}} = \frac{11}{15}$

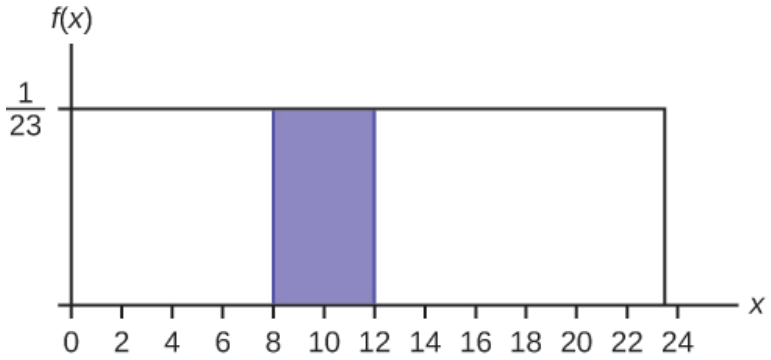


Figure 5.14

TRY IT 5.3

A distribution is given as $X \sim U(0, 20)$. What is $P(2 < x < 18)$? Find the 90th percentile.

EXAMPLE 5.4

The amount of time, in minutes, that a person must wait for a bus is uniformly distributed between zero and 15 minutes, inclusive.

Problem

a. What is the probability that a person waits fewer than 12.5 minutes?

Solution

a. Let X = the number of minutes a person must wait for a bus. $a = 0$ and $b = 15$. $X \sim U(0, 15)$. Write the probability density function. $f(x) = \frac{1}{15 - 0} = \frac{1}{15}$ for $0 \leq x \leq 15$.

Find $P(x < 12.5)$. Draw a graph.

$$P(x < k) = (\text{base})(\text{height}) = (12.5 - 0)\left(\frac{1}{15}\right) = 0.8333$$

The probability a person waits fewer than 12.5 minutes is 0.8333.

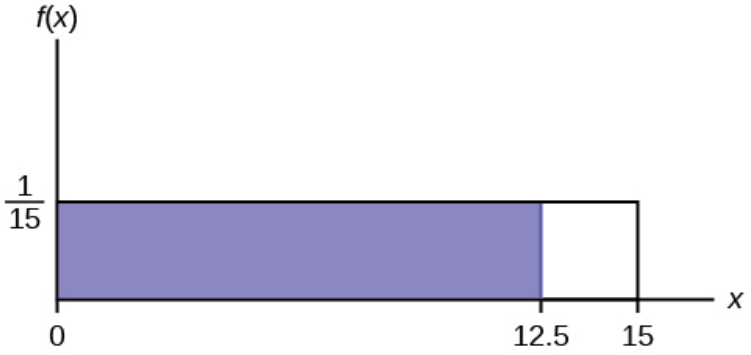


Figure 5.15

Problem

b. On the average, how long must a person wait? Find the mean, μ, and the standard deviation, σ.

Solution

b. $\mu = \frac{a + b}{2} = \frac{15 + 0}{2} = 7.5$. On the average, a person must wait 7.5 minutes.

$\sigma = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(15-0)^2}{12}} = 4.3$. The standard deviation is 4.3 minutes.

Problem

c. Ninety percent of the time, the minutes a person must wait falls below what value?

This question asks for the 90th percentile.

Solution

c. Find the 90th percentile. Draw a graph. Let k = the 90th percentile.

$P(x < k) = (\text{base})(\text{height}) = (k - 0)(\frac{1}{15})$

$0.90 = (k)\left(\frac{1}{15}\right)$

$k = (0.90)(15) = 13.5$

k is sometimes called a critical value.

The 90th percentile is 13.5 minutes. Ninety percent of the time, a person must wait at most 13.5 minutes.

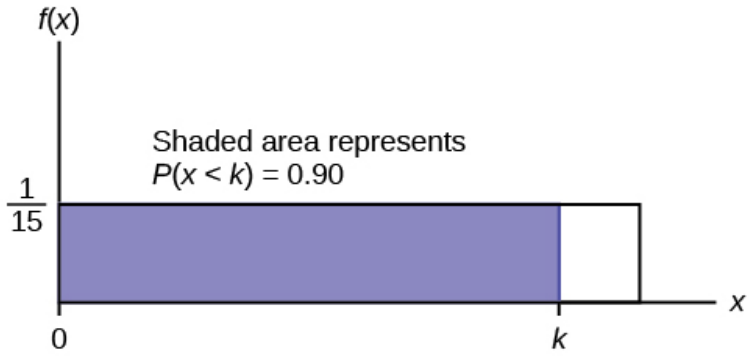


Figure 5.16

TRY IT 5.4

The total duration of baseball games in the major league in the 2011 season is uniformly distributed between 447 hours and 521 hours inclusive.

a. Find a and b and describe what they represent.
b. Write the distribution.
c. Find the mean and the standard deviation.
d. What is the probability that the duration of games for a team for the 2011 season is between 480 and 500 hours?
e. What is the 65th percentile for the duration of games for a team for the 2011 season?

EXAMPLE 5.5

Suppose the time it takes a nine-year old to eat a donut is between 0.5 and 4 minutes, inclusive. Let X = the time, in minutes, it takes a nine-year-old child to eat a doughnut. Then $X \sim U(0.5, 4)$.

Problem

a. The probability that a randomly selected nine-year-old child eats a doughnut in at least two minutes is _____.

Solution

a. 0.5714

Problem

b. Find the probability that a different nine-year-old child eats a doughnut in more than two minutes given that the child has already been eating the doughnut for more than 1.5 minutes.

The second question has a **conditional probability**. You are asked to find the probability that a nine-year-old child eats a doughnut in more than two minutes given that the child has already been eating the donut for more than 1.5 minutes. Solve the problem two different ways (see Example 5.3). You must reduce the sample space. *First way*: Since you know the child has already been eating the doughnut for more than 1.5 minutes, you are no longer starting at $a = 0.5$ minutes. Your starting point is 1.5 minutes.

Write a new f(x):

$$f(x) = \frac{1}{4 - 1.5} = \frac{2}{5} \text{ for } 1.5 \leq x \leq 4.$$

Find $P(x > 2 \mid x > 1.5)$. Draw a graph.

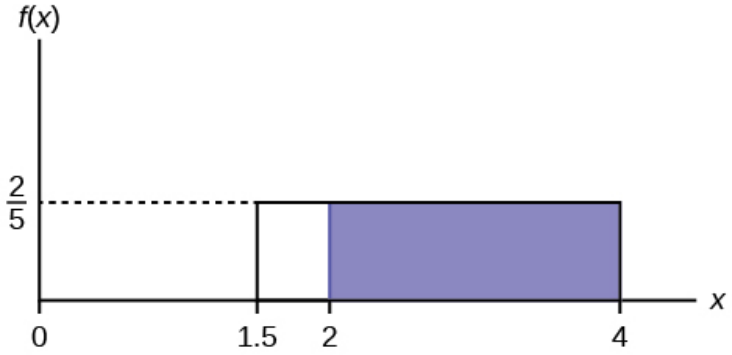


Figure 5.17

$$P(x > 2 \mid x > 1.5) = (\text{base})(\text{new height}) = (4 - 2)\left(\frac{2}{5}\right) = \frac{4}{5}$$

Solution

b. $\frac{4}{5}$

The probability that a nine-year-old child eats a donut in more than two minutes given that the child has already been eating the doughnut for more than 1.5 minutes is $\frac{4}{5}$.

Second way: Draw the original graph for $X \sim U(0.5, 4)$. Use the conditional formula

$$P(x > 2 \mid x > 1.5) = \frac{P(x > 2 \text{ AND } x > 1.5)}{P(x > 1.5)} = \frac{P(x > 2)}{P(x > 1.5)} = \frac{\frac{2}{3.5}}{\frac{2.5}{3.5}} = 0.8 = \frac{4}{5}$$

TRY IT 5.5

Suppose the time it takes a student to finish a quiz is uniformly distributed between six and 15 minutes, inclusive. Let X = the time, in minutes, it takes a student to finish a quiz. Then $X \sim U(6, 15)$.

Find the probability that a randomly selected student needs at least eight minutes to complete the quiz. Then find the probability that a different student needs at least eight minutes to finish the quiz given that she has already taken more than seven minutes.

EXAMPLE 5.6

Ace Heating and Air Conditioning Service finds that the amount of time a repairman needs to fix a furnace is uniformly distributed between 1.5 and four hours. Let x = the time needed to fix a furnace. Then $x \sim U(1.5, 4)$.

Problem

a. Find the probability that a randomly selected furnace repair requires more than two hours.
b. Find the probability that a randomly selected furnace repair requires less than three hours.
c. Find the 30^{th} percentile of furnace repair times.
d. The longest 25 percent of furnace repair times take at least how long? (In other words: find the minimum time for the longest 25 percent of repair times.) What percentile does this represent?
e. Find the mean and standard deviation

Solution

a. To find $f(x)$: $f(x) = \frac{1}{4 - 1.5} = \frac{1}{2.5}$ so $f(x) = 0.4$

$P(x > 2)$ = (base)(height) = (4 – 2)(0.4) = 0.8

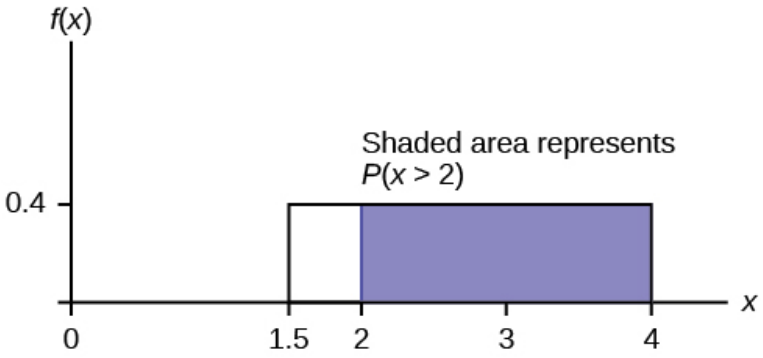


Figure 5.18 Uniform distribution between 1.5 and four with shaded area between two and four representing the probability that the repair time x is greater than two

Solution

b. $P(x < 3)$ = (base)(height) = (3 – 1.5)(0.4) = 0.6

The graph of the rectangle showing the entire distribution would remain the same. However the graph should be shaded between $x = 1.5$ and $x = 3$. Note that the shaded area starts at $x = 1.5$ rather than at $x = 0$. Because $X \sim U(1.5, 4)$, x cannot be less than 1.5.

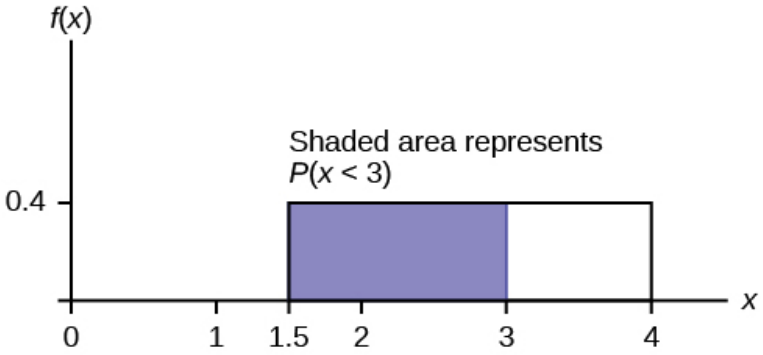


Figure 5.19 Uniform distribution between 1.5 and four with shaded area between 1.5 and three representing the probability that the repair time x is less than three

Solution

c.

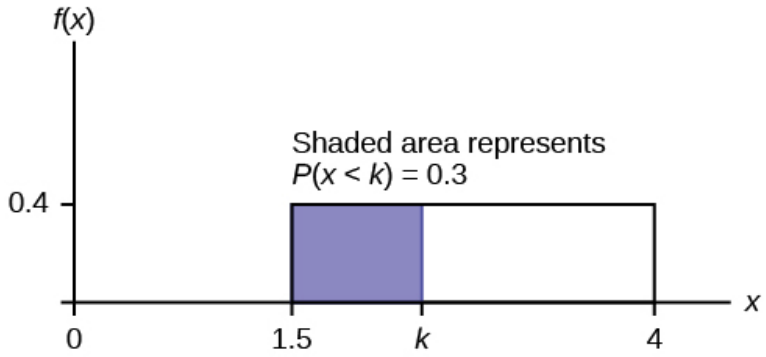


Figure 5.20 Uniform distribution between 1.5 and 4 with an area of 0.30 shaded to the left, representing the shortest 30 percent of repair times.

$P(x < k) = 0.30$
$P(x < k)$ = (base)(height) = $(k - 1.5)(0.4)$
$0.3 = (k - 1.5)(0.4)$; Solve to find k:
$0.75 = k - 1.5$, obtained by dividing both sides by 0.4
$k = 2.25$, obtained by adding 1.5 to both sides
The 30th percentile of repair times is 2.25 hours. 30 percent of repair times are 2.5 hours or less.

Solution
d.

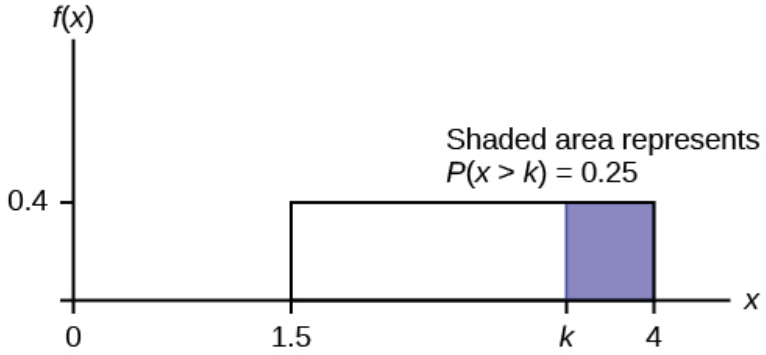


Figure 5.21 Uniform distribution between 1.5 and 4 with an area of 0.25 shaded to the right representing the longest 25 percent of repair times.

$P(x > k) = 0.25$
$P(x > k)$ = (base)(height) = $(4 - k)(0.4)$
$0.25 = (4 - k)(0.4)$; Solve for k:
$0.625 = 4 - k$,
obtained by dividing both sides by 0.4
$-3.375 = -k$,
obtained by subtracting four from both sides: **$k = 3.375$**
The longest 25 percent of furnace repairs take at least 3.375 hours (3.375 hours or longer).
Note: Since 25 percent of repair times are 3.375 hours or longer, that means that 75 percent of repair times are 3.375 hours or less. 3.375 hours is the *75th percentile* of furnace repair times.

Solution
e. $\mu = \frac{a+b}{2}$ and $\sigma = \sqrt{\frac{(b-a)^2}{12}}$
$\mu \frac{1.5+4}{2} 2.75$ hours and $\sigma = \sqrt{\frac{(4-1.5)^2}{12}} = 0.7217$ hours

TRY IT 5.6

The amount of time a service technician needs to change the oil in a car is uniformly distributed between 11 and 21 minutes. Let X = the time needed to change the oil on a car.

a. Write the random variable X in words. X = _____.
b. Write the distribution.
c. Graph the distribution.
d. Find $P(x > 19)$.
e. Find the 50^{th} percentile.

5.3 The Exponential Distribution (Optional)

The **exponential distribution** is often concerned with the amount of time until some specific event occurs. For example, the amount of time (beginning now) until an earthquake occurs has an exponential distribution. Other examples include the length, in minutes, of long-distance business telephone calls, and the amount of time, in months, a car battery lasts. It can be shown, too, that the value of the change that you have in your pocket or purse approximately follows an exponential distribution.

Values for an exponential random variable occur in the following way. There are fewer large values and more small values. For example, the amount of money customers spend in one trip to the supermarket follows an exponential distribution. There are more people who spend small amounts of money and fewer people who spend large amounts of money.

Exponential distributions are commonly used in calculations of product reliability, or the length of time a product lasts.

EXAMPLE 5.7

Let X = amount of time (in minutes) a postal clerk spends with his or her customer. The time is known to have an exponential distribution with the average amount of time equal to four minutes.

X is a **continuous random variable** since time is measured. It is given that $\mu = 4$ minutes. To do any calculations, you must know m, the decay parameter.

$m = \frac{1}{\mu}$. Therefore, $m = \frac{1}{4} = 0.25$.

The standard deviation, σ, is the same as the mean. $\mu = \sigma$

The distribution notation is $X \sim Exp(m)$. Therefore, $X \sim Exp(0.25)$.

The probability density function is $f(x) = me^{-mx}$. The number $e = 2.71828182846\ldots$ It is a number that is used often in mathematics. Scientific calculators have the key "e^x." If you enter one for x, the calculator will display the value e.

The curve is

$f(x) = 0.25e^{-0.25x}$ where x is at least zero and $m = 0.25$.

For example, $f(5) = 0.25e^{(-0.25)(5)} = 0.072$. The probability that the postal clerk spends five minutes with the customers is 0.072.

The graph is as follows:

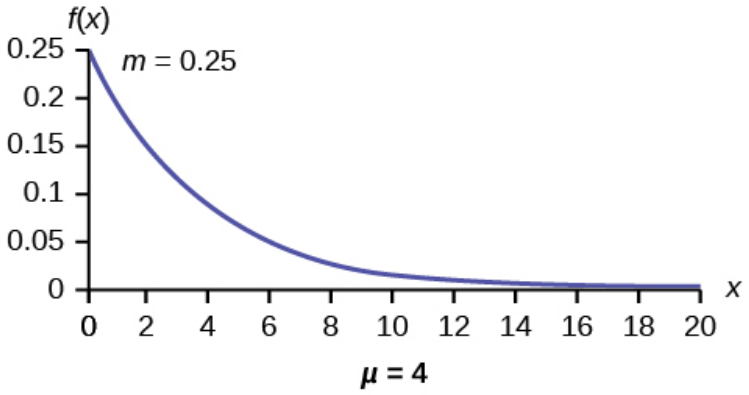


Figure 5.22

Notice the graph is a declining curve. When $x = 0$,

$f(x) = 0.25e^{(-0.25)(0)} = (0.25)(1) = 0.25 = m$. The maximum value on the y-axis is m.

TRY IT 5.7

The amount of time spouses shop for anniversary cards can be modeled by an exponential distribution with the average amount of time equal to eight minutes. Write the distribution, state the probability density function, and graph the distribution.

EXAMPLE 5.8

Problem

a. Using the information in Example 5.7, find the probability that a clerk spends four to five minutes with a randomly selected customer.

Solution

a. Find $P(4 < x < 5)$. The **cumulative distribution function (CDF)** gives the area to the left.

$$P(x < x) = 1 - e^{-mx}$$

$$P(x < 5) = 1 - e^{(-0.25)(5)} = 0.7135 \text{ and } P(x < 4) = 1 - e^{(-0.25)(4)} = 0.6321$$

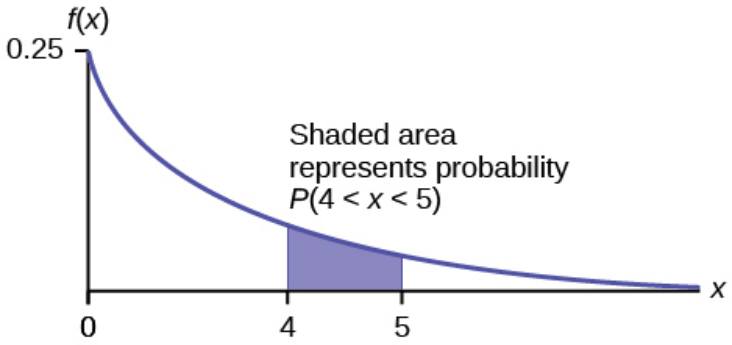


Figure 5.23

NOTE

You can do these calculations easily on a calculator.

The probability that a postal clerk spends four to five minutes with a randomly selected customer is $P(4 < x < 5) = P(x < 5) - P(x < 4) = 0.7135 - 0.6321 = 0.0814$.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

On the home screen, enter (1 – e^(–0.25*5))–(1–e^(–0.25*4)) or enter e^(–0.25*4) – e^(–0.25*5).

Problem

b. Half of all customers are finished within how long? (Find the 50th percentile).

Solution

b. Find the 50th percentile.

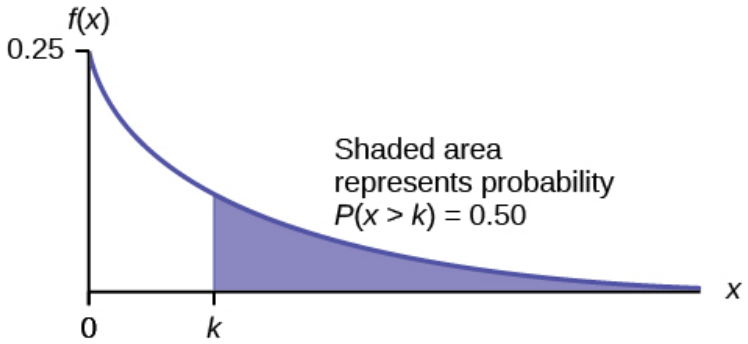


Figure 5.24

$P(x < k) = 0.50$, $k = 2.8$ minutes (calculator or computer)

Half of all customers are finished within 2.8 minutes.

You can also do the calculation as follows:

$$P(x < k) = 0.50 \text{ and } P(x < k) = 1 - e^{-0.25k}$$

Therefore, $0.50 = 1 - e^{-0.25k}$ and $e^{-0.25k} = 1 - 0.50 = 0.5$.

Take natural logs: $ln(e^{-0.25k}) = ln(0.50)$. So, $-0.25k = ln(0.50)$.

Solve for k: $k = \frac{ln(0.50)}{-0.25} = 2.8$ minutes. The calculator simplifies the calculation for percentile k. See the following two notes.

Note

A formula for the percentile k is $k = \frac{ln(1 - AreaToTheLeft)}{-m}$ where ln is the natural log.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

On the home screen, enter ln(1 – 0.50)/–0.25. Press the (–) for the negative.

Problem

c. Which is larger, the mean or the median?

Solution

c. From Part b, the median or 50th percentile is 2.8 minutes. The theoretical mean is four minutes. The mean is larger.

TRY IT 5.8

The number of days ahead travelers purchase their airline tickets can be modeled by an exponential distribution with the average amount of time equal to 15 days. Find the probability that a traveler will purchase a ticket fewer than 10 days in advance. How many days do half of all travelers wait?

COLLABORATIVE EXERCISE

Have each class member count the change he or she has in his or her pocket or purse. Your instructor will record the amounts in dollars and cents. Construct a histogram of the data taken by the class. Use five intervals. Draw a smooth curve through the bars. The graph should look approximately exponential. Then calculate the mean.

Let X = the amount of money a student in your class has in his or her pocket or purse.

The distribution for X is approximately exponential with mean, μ = _____ and m = _____. The standard deviation, σ = _____.

Draw the appropriate exponential graph. You should label the x– and y–axes, the decay rate, and the mean. Shade the area that represents the probability that one student has less than $0.40 in his or her pocket or purse. (Shade $P(x < 0.40)$).

EXAMPLE 5.9

On the average, a certain computer part lasts 10 years. The length of time the computer part lasts is exponentially distributed.

Problem

a. What is the probability that a computer part lasts more than seven years?

Solution

a. Let x = the amount of time (in years) a computer part lasts.

$$\mu = 10, \text{ so } m = \frac{1}{\mu} = \frac{1}{10} = 0.1$$

Find $P(x > 7)$. Draw the graph.

$$P(x > 7) = 1 - P(x < 7).$$

Since $P(X < x) = 1 - e^{-mx}$ then $P(X > x) = 1 - (1 - e^{-mx}) = e^{-mx}$

$P(x > 7) = e^{(-0.1)(7)} = 0.4966$. The probability that a computer part lasts more than seven years is 0.4966.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

On the home screen, enter e^(-.1*7).

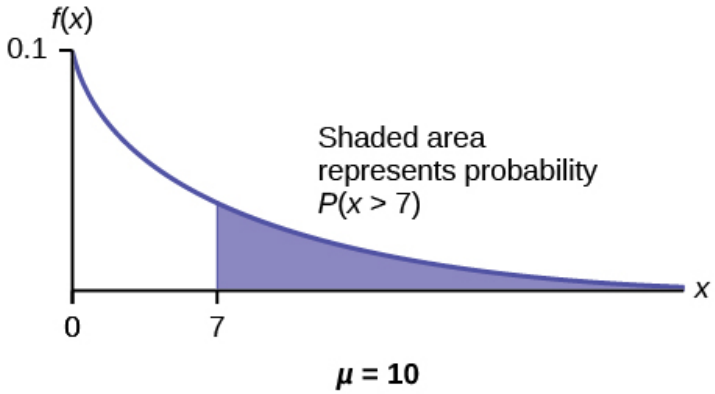


Figure 5.25

Problem

b. On the average, how long would five computer parts last if they are used one after another?

Solution

b. On the average, one computer part lasts 10 years. Therefore, five computer parts, if they are used one right after the other would last, on the average, (5)(10) = 50 years.

Problem

c. Eighty percent of computer parts last at most how long?

Solution

c. Find the 80th percentile. Draw the graph. Let k = the 80th percentile.

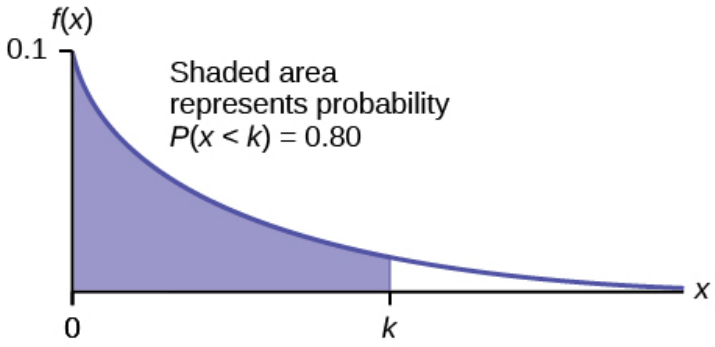


Figure 5.26

Solve for k:

$$k = \frac{ln(1-0.80)}{-0.1} = 16.1 \text{years.}$$

Eighty percent of the computer parts last at most 16.1 years.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

On the home screen, enter $\frac{\ln(1-0.80)}{-0.1}$.

Problem

d. What is the probability that a computer part lasts between nine and 11 years?

Solution

d. Find $P(9 < x < 11)$. Draw the graph.

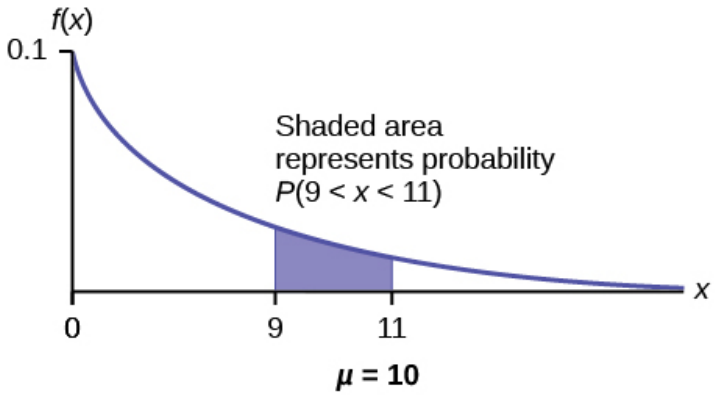


Figure 5.27

$P(9 < x < 11) = P(x < 11) - P(x < 9) = (1 - e^{(-0.1)(11)}) - (1 - e^{(-0.1)(9)}) = 0.6671 - 0.5934 = 0.0737$. The probability that a computer part lasts between nine and 11 years is 0.0737.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

On the home screen, enter e^(–0.1*9) – e^(–0.1*11).

TRY IT 5.9

On average, a pair of running shoes can last 18 months if used every day. The length of time running shoes last is exponentially distributed. What is the probability that a pair of running shoes last more than 15 months? On average, how long would six pairs of running shoes last if they are used one after the other? Eighty percent of running shoes last at most how long if used every day?

EXAMPLE 5.10

Suppose that the length of a phone call, in minutes, is an exponential random variable with decay parameter $\frac{1}{12}$. If another person arrives at a public telephone just before you, find the probability that you will have to wait more than five minutes. Let X = the length of a phone call, in minutes.

Problem

What is m, μ, and σ? The probability that you must wait more than five minutes is _____ .

Solution

- $m = \frac{1}{12}$

- $\mu = 12$
- $\sigma = 12$

$P(x > 5) = 0.6592$

TRY IT 5.10

Suppose that the distance, in miles, that people are willing to commute to work is an exponential random variable with a decay parameter $\frac{1}{20}$. Let X = the distance people are willing to commute in miles. What is m, μ, and σ? What is the probability that a person is willing to commute more than 25 miles?

EXAMPLE 5.11

The time spent waiting between events is often modeled using the exponential distribution. For example, suppose that an average of 30 customers per hour arrive at a store and the time between arrivals is exponentially distributed.

Problem

a. On average, how many minutes elapse between two successive arrivals?
b. When the store first opens, how long on average does it take for three customers to arrive?
c. After a customer arrives, find the probability that it takes less than one minute for the next customer to arrive.
d. After a customer arrives, find the probability that it takes more than five minutes for the next customer to arrive.
e. Seventy percent of the customers arrive within how many minutes of the previous customer?
f. Is an exponential distribution reasonable for this situation?

Solution

a. Since we expect 30 customers to arrive per hour (60 minutes), we expect on average one customer to arrive every two minutes on average.
b. Since one customer arrives every two minutes on average, it will take six minutes on average for three customers to arrive.
c. Let X = the time between arrivals, in minutes. By Part a, $\mu = 2$, so $m = \frac{1}{2} = 0.5$.
Therefore, $X \sim Exp(0.5)$.
The cumulative distribution function is $P(X < x) = 1 - e^{(-0.5)(x)}$.
Therefore $P(X < 1) = 1 - e^{(-0.5)(1)} \approx 0.3935$.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

1 - e^(–0.5) ≈ 0.3935

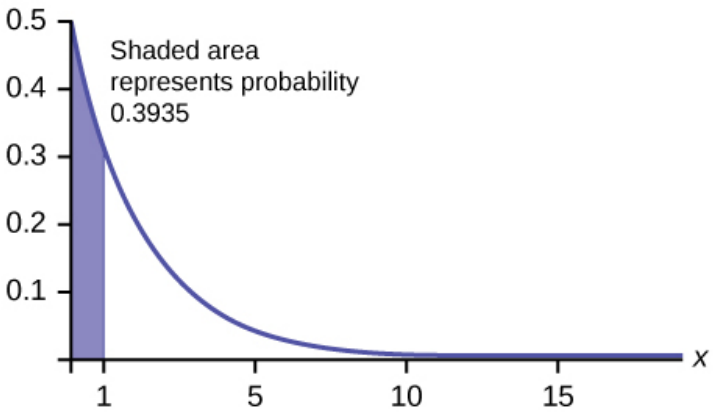


Figure 5.28

d. $P(X > 5) = 1 - P(X < 5) = 1 - (1 - e^{(-0.5)(5)}) = e^{-2.5} \approx 0.0821$.

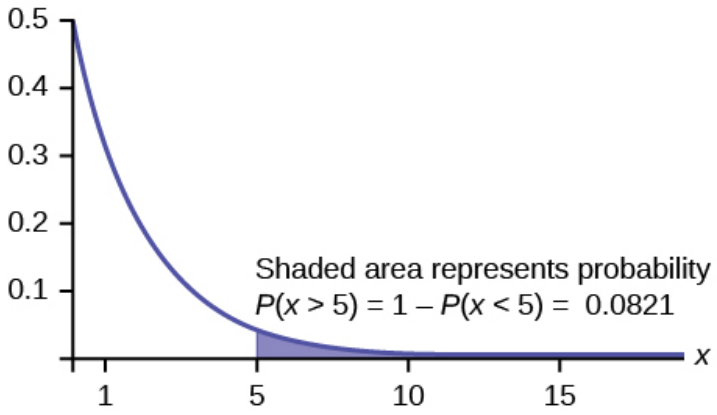


Figure 5.29

USING THE TI-83, 83+, 84, 84+ CALCULATOR

$1 - (1 - e \text{ ^ } (-0.50)(5))$ *or* $e \text{ ^ } (-0.50)(5)$

e. We want to solve $0.70 = P(X < x)$ for x.
Substituting in the cumulative distribution function gives $0.70 = 1 - e^{-0.5x}$, so that $e^{-0.5x} = 0.30$. Converting this to logarithmic form gives $-0.5x = ln(0.30)$, or $x = \frac{ln(0.30)}{-0.5} \approx 2.41$ minutes.
Thus, 70 percent of customers arrive within 2.41 minutes of the previous customer.
You are finding the 70th percentile k so you can use the formula

$$k = \frac{ln(1 - Area_To_The_Left_Of_k)}{(-m)}$$

$$k = \frac{ln(1-0.70)}{(-0.5)} \approx 2.41 \text{ minutes} \quad \textbf{5.1}$$

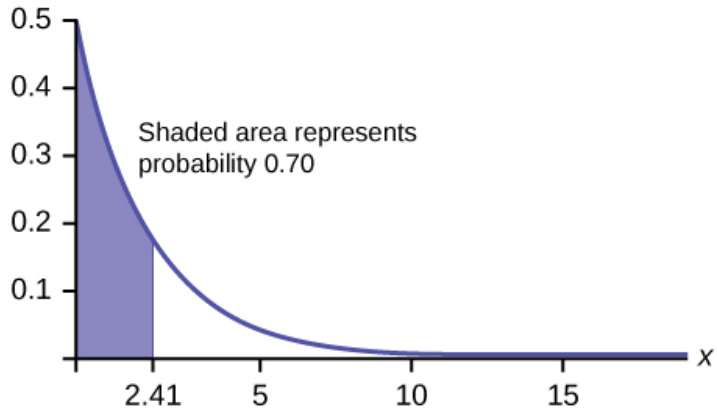


Figure 5.30

f. This model assumes that a single customer arrives at a time, which may not be reasonable since people might shop in groups, leading to several customers arriving at the same time. It also assumes that the flow of customers does not change throughout the day, which is not valid if some times of the day are busier than others.

TRY IT 5.11

Suppose that on a certain stretch of highway, cars pass at an average rate of five cars per minute. Assume that the duration of time between successive cars follows the exponential distribution.

a. On average, how many seconds elapse between two successive cars?
b. After a car passes by, how long on average will it take for another seven cars to pass by?
c. Find the probability that after a car passes by, the next car will pass within the next 20 seconds.
d. Find the probability that after a car passes by, the next car will not pass for at least another 15 seconds.

Memorylessness of the Exponential Distribution

In Example 5.7 recall that the amount of time between customers is exponentially distributed with a mean of two minutes ($X \sim Exp(0.5)$). Suppose that five minutes have elapsed since the last customer arrived. Since an unusually long amount of time has now elapsed, it would seem to be more likely for a customer to arrive within the next minute. With the exponential distribution, this is not the case—the additional time spent waiting for the next customer does not depend on how much time has already elapsed since the last customer. This is referred to as the **memoryless property**. Specifically, the **memoryless property** says the following

$$P(X > r + t \mid X > r) = P(X > t) \text{ for all } r \geq 0 \text{ and } t \geq 0$$

For example, if five minutes have elapsed since the last customer arrived, then the probability that more than one minute will elapse before the next customer arrives is computed by using $r = 5$ and $t = 1$ in the foregoing equation.

$$P(X > 5 + 1 \mid X > 5) = P(X > 1) = e^{(-0.5)(1)} \approx 0.6065.$$

This is the same probability as that of waiting more than one minute for a customer to arrive after the previous arrival.

The exponential distribution is often used to model the longevity of an electrical or a mechanical device. In Example 5.9, the lifetime of a certain computer part has the exponential distribution with a mean of ten years ($X \sim Exp(0.1)$). The **memoryless property** says that knowledge of what has occurred in the past has no effect on future probabilities. In this case it means that an old part is not any more likely to break down at any particular time than a brand new part. In other words, the part stays as good as new until it suddenly breaks. For example, if the part has already lasted ten years, then the probability that it lasts another seven years is $P(X > 17 \mid X > 10) = P(X > 7) = 0.4966$.

EXAMPLE 5.12

Refer to Example 5.7 where the time a postal clerk spends with his or her customer has an exponential distribution with a mean of four minutes. Suppose a customer has spent four minutes with a postal clerk. What is the probability that he or she will spend at least an additional three minutes with the postal clerk?

The decay parameter of X is $m = \frac{1}{4} = 0.25$, so $X \sim Exp(0.25)$.

The cumulative distribution function is $P(X < x) = 1 - e^{-0.25x}$.

We want to find $P(X > 7 \mid X > 4)$. The **memoryless property** says that $P(X > 7 \mid X > 4) = P(X > 3)$, so we just need to find the probability that a customer spends more than three minutes with a postal clerk.

This is $P(X > 3) = 1 - P(X < 3) = 1 - (1 - e^{-0.25 \cdot 3}) = e^{-0.75} \approx 0.4724$.

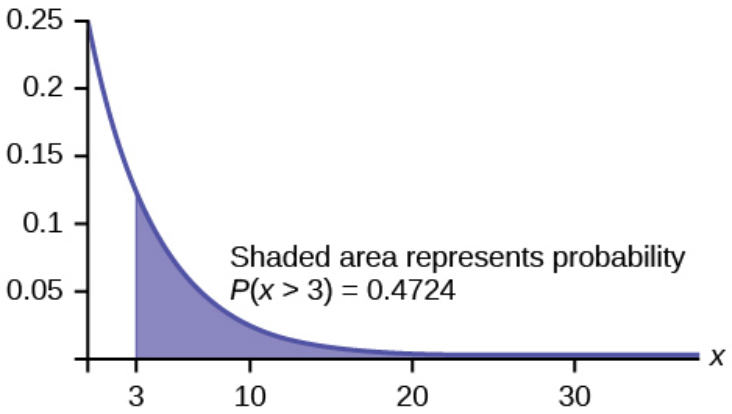


Figure 5.31

USING THE TI-83, 83+, 84, 84+ CALCULATOR

1–(1–e^(–0.25*3)) = e^(–0.25*3).

TRY IT 5.12

Suppose that the longevity of a light bulb is exponential with a mean lifetime of eight years. If a bulb has already lasted 12 years, find the probability that it will last a total of more than 19 years.

Relationship Between the Poisson and the Exponential Distribution

There is an interesting relationship between the exponential distribution and the Poisson distribution. Suppose that the time that elapses between two successive events follows the exponential distribution with a mean of μ units of time. Also assume that these times are independent, meaning that the time between events is not affected by the times between previous events. If these assumptions hold, then the number of events per unit time follows a Poisson distribution with mean $\lambda = 1/\mu$. Recall from the chapter on Discrete Random Variables that if X has the Poisson distribution with mean λ, then $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$. Conversely, if the number of events per unit time follows a Poisson distribution, then the amount of time between events follows the exponential distribution. ($k!$ = k*(k–1*)(k–2)*(k–3)...3*2*1)

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Suppose X has the Poisson distribution with mean λ. Compute $P(X = k)$ by entering 2nd, VARS(DISTR), C: poissonpdf(λ, k). To compute $P(X \leq k)$, enter 2nd, VARS (DISTR), D:poissoncdf(λ, k).

EXAMPLE 5.13

At a police station in a large city, calls come in at an average rate of four calls per minute. Assume that the time that elapses from one call to the next has the exponential distribution. Take note that we are concerned only with the rate at which calls come in, and we are ignoring the time spent on the phone. We must also assume that the times spent between calls are independent. This means that a particularly long delay between two calls does not mean that there will be a shorter waiting period for the next call. We may then deduce that the total number of calls received during a time period has the Poisson distribution.

Problem

a. Find the average time between two successive calls.
b. Find the probability that after a call is received, the next call occurs in less than 10 seconds.
c. Find the probability that exactly five calls occur within a minute.
d. Find the probability that fewer than five calls occur within a minute.
e. Find the probability that more than 40 calls occur in an eight-minute period.

Solution

a. On average four calls occur per minute, so 15 seconds, or $\frac{15}{60}$ = 0.25 minutes occur between successive calls on average.
b. Let T = time elapsed between calls. From Part a, $\mu = 0.25$, so $m = \frac{1}{0.25} = 4$. Thus, $T \sim Exp(4)$.
The cumulative distribution function is $P(T < t) = 1 - e^{-4t}$.
The probability that the next call occurs in less than 10 seconds (10 seconds = 1/6 minute) is
$P\left(T < \frac{1}{6}\right) = 1 - e^{(-4)\left(\frac{1}{6}\right)} \approx 0.4866.$

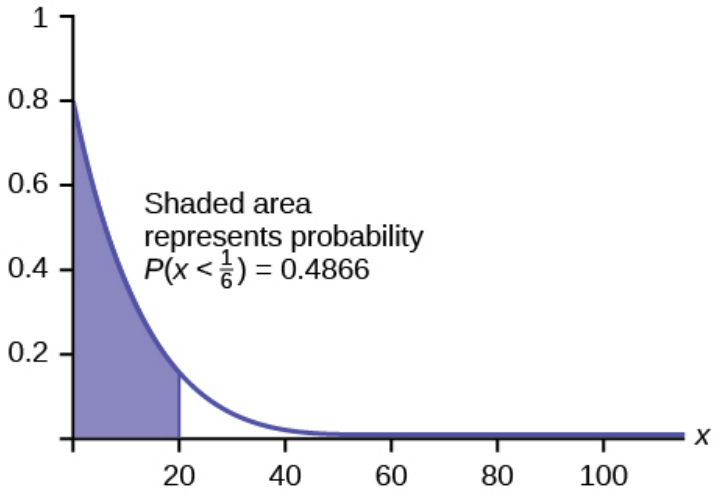


Figure 5.32

c. Let X = the number of calls per minute. As previously stated, the number of calls per minute has a Poisson distribution, with a mean of four calls per minute.
Therefore, $X \sim Poisson(4)$, and so $P(X = 5) = \frac{4^5 e^{-4}}{5!} \approx 0.1563$. (5! = (5)(4)(3)(2)(1))

USING THE TI-83, 83+, 84, 84+ CALCULATOR

poissonpdf(4, 5) = 0.1563

d. Keep in mind that X must be a whole number, so $P(X < 5) = P(X \leq 4)$.
To compute this, we could take $P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$.
Using technology, we see that $P(X \leq 4) = 0.6288$.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

poisssoncdf(4, 4) = 0.6288

e. Let Y = the number of calls that occur during an eight-minute period.
Since there is an average of four calls per minute, there is an average of (8)(4) = 32 calls during each eight minute period.
Hence, $Y \sim Poisson(32)$. Therefore, $P(Y > 40) = 1 - P(Y \leq 40) = 1 - 0.9294 = 0.0706$.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

1 – poissoncdf(32, 40). = 0.0706

TRY IT 5.13

In a small city, the number of automobile accidents occur with a Poisson distribution at an average of three per week.

a. Calculate the probability that at most two accidents occur in any given week.
b. What is the probability that there are at least two weeks between any two accidents?

5.4 Continuous Distribution

Stats Lab

Continuous Distribution

Student Learning Outcomes

- The student will compare and contrast empirical data from a random number generator with the uniform distribution.

Collect the Data

Use a random number generator to generate 50 values between zero and one (inclusive). List them in Table 5.3. Round the numbers to four decimal places or set the calculator MODE to four places.

1. Complete the table.

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

Table 5.3

2. Calculate the following:

a. $\overline{x}$ = _____
b. s = _____
c. first quartile = _____
d. third quartile = _____
e. median = _____

Organize the Data

1. Construct a histogram of the empirical data. Make eight bars.

Figure 5.33

2. Construct a histogram of the empirical data. Make five bars.

Figure 5.34

Describe the Data

1. In two to three complete sentences, describe the shape of each graph. (Keep it simple. Does the graph go straight across, does it have a V shape, does it have a hump in the middle or at either end (and so on). One way to help you determine a shape is to draw a smooth curve roughly through the top of the bars.)
2. Describe how changing the number of bars might change the shape.

Theoretical Distribution

1. In words, X = _____.
2. The theoretical distribution of X is $X \sim U(0,1)$.
3. In theory, based upon the distribution $X \sim U(0,1)$, complete the following.
   a. μ = _____
   b. σ = _____
   c. first quartile = _____
   d. third quartile = _____
   e. median = _____
4. Are the empirical values (the data) in the section titled Collect the Data close to the corresponding theoretical values? Why or why not?

Plot the Data

1. Construct a box plot of the data. Be sure to use a ruler to scale accurately and draw straight edges.
2. Do you notice any potential outliers? If so, which values are they? Either way, justify your answer numerically. (Recall that any data that are less than $Q_1 - 1.5(IQR)$ or more than $Q_3 + 1.5(IQR)$ are potential outliers. *IQR* means interquartile range.)

Compare the Data

1. For each of the following parts, use a complete sentence to comment on how the value obtained from the data compares to the theoretical value you expected from the distribution in the section titled Theoretical Distribution:
   a. minimum value: _____
   b. first quartile: _____
   c. median: _____
   d. third quartile: _____
   e. maximum value: _____
   f. width of *IQR*: _____
   g. overall shape: _____
2. Based on your comments in the section titled Collect the Data, how does the box plot fit or not fit what you would expect of the distribution in the section titled Theoretical Distribution?

Discussion Question

1. Suppose that the number of values generated was 500, not 50. How would that affect what you would expect the empirical data to be and the shape of its graph to look like?

Key Terms

conditional probability the likelihood that an event will occur given that another event has already occurred

decay parameter The decay parameter describes the rate at which probabilities decay to zero for increasing values of x.
It is the value m in the probability density function $f(x) = me^{(-mx)}$ of an exponential random variable.
It is also equal to $m = \frac{1}{\mu}$, where μ is the mean of the random variable.

exponential distribution a continuous random variable (RV) that appears when we are interested in the intervals of time between some random events, for example, the length of time between emergency arrivals at a hospital; the notation is $X \sim Exp(m)$.
The mean is $\mu = \frac{1}{m}$ and the standard deviation is $\sigma = \frac{1}{m}$. The probability density function is $f(x) = me^{-mx}$, $x \geq 0$ and the cumulative distribution function is $P(X \leq x) = 1 - e^{-mx}$.

memoryless property for an exponential random variable X, the statement that knowledge of what has occurred in the past has no effect on future probabilities
This means that the probability that X exceeds $x + k$, given that it has exceeded x, is the same as the probability that X would exceed k if we had no knowledge about it. In symbols we say that $P(X > x + k | X > x) = P(X > k)$.

Poisson distribution a distribution function that gives the probability of a number of events occurring in a fixed interval of time or space if these events happen with a known average rate and independently of the time since the last event; if there is a known average of λ events occurring per unit time, and these events are independent of each other, then the number of events X occurring in one unit of time has the Poisson distribution.
The probability of k events occurring in one unit time is equal to $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$.

uniform distribution a continuous random variable (RV) that has equally likely outcomes over the domain, $a < x < b$.
Notation—$X \sim U(a,b)$.
The mean is $\mu = \frac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$. The probability density function is $f(x) = \frac{1}{b-a}$ for $a < x < b$ or $a \leq x \leq b$. The cumulative distribution is $P(X \leq x) = \frac{x-a}{b-a}$.

Chapter Review

5.1 Continuous Probability Functions

The probability density function (pdf) is used to describe probabilities for continuous random variables. The area under the density curve between two points corresponds to the probability that the variable falls between those two values. In other words, the area under the density curve between points a and b is equal to $P(a < x < b)$. The cumulative distribution function (cdf) gives the probability as an area. If X is a continuous random variable, the probability density function (pdf), $f(x)$, is used to draw the graph of the probability distribution. The total area under the graph of $f(x)$ is one. The area under the graph of $f(x)$ and between values a and b gives the probability $P(a < x < b)$.

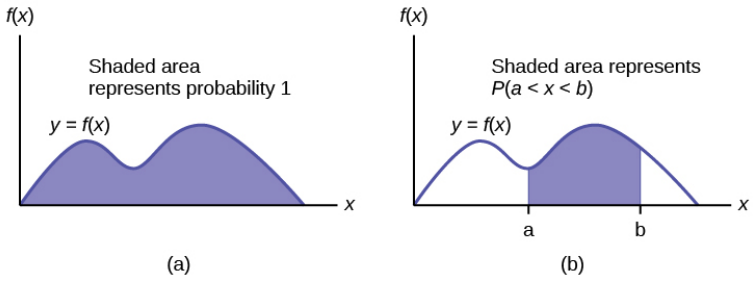


Figure 5.35

The cumulative distribution function (cdf) of X is defined by $P(X \leq x)$. It is a function of x that gives the probability that the random variable is less than or equal to x.

5.2 The Uniform Distribution

If X has a uniform distribution where $a < x < b$ or $a \leq x \leq b$, then X takes on values between a and b (may include a and b). All values x are equally likely. We write $X \sim U(a, b)$. The mean of X is $\mu = \frac{a+b}{2}$. The standard deviation of X is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$. The probability density function of X is $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$. The cumulative distribution function of X is $P(X \leq x) = \frac{x-a}{b-a}$. X is continuous.

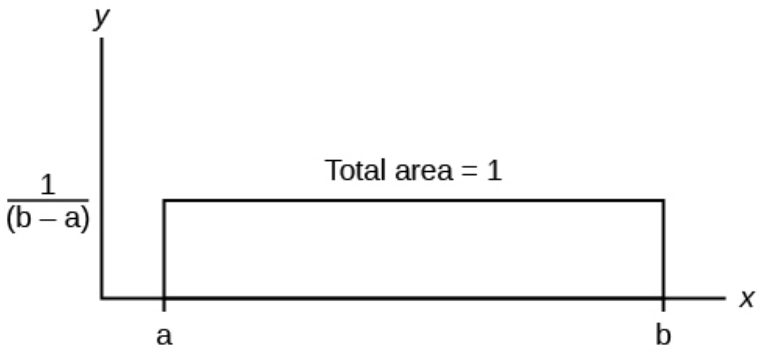


Figure 5.36

The probability $P(c < X < d)$ may be found by computing the area under $f(x)$, between c and d. Since the corresponding area is a rectangle, the area may be found simply by multiplying the width and the height.

5.3 The Exponential Distribution (Optional)

If X has an **exponential distribution** with mean μ, then the **decay parameter** is $m = \frac{1}{\mu}$, and we write $X \sim Exp(m)$ where $x \geq 0$ and $m > 0$. The probability density function of X is $f(x) = me^{-mx}$ (or equivalently $f(x) = \frac{1}{\mu}e^{-x/\mu}$. The cumulative distribution function of X is $P(X \leq x) = 1 - e^{-mx}$.

The exponential distribution has the **memoryless property**, which says that future probabilities do not depend on any past information. Mathematically, it says that $P(X > x + k | X > x) = P(X > k)$.

If T represents the waiting time between events, and if $T \sim Exp(\lambda)$, then the number of events X per unit time follows the Poisson distribution with mean λ. The probability density function of X is $P(X = k) = \frac{\lambda^k e^{-k}}{k!}$. This may be computed using a TI-83, 83+, 84, 84+ calculator with the command poissonpdf(λ, k). The cumulative distribution function $P(X \leq k)$ may be computed using the TI-83, 83+,84, 84+ calculator with the command poissoncdf(λ, k).

Formula Review

5.1 Continuous Probability Functions

Probability density function (pdf) $f(x)$:

- $f(x) \geq 0$
- The total area under the curve $f(x)$ is one.

Cumulative distribution function (cdf): $P(X \leq x)$

5.2 The Uniform Distribution

X = a real number between a and b (in some instances, X can take on the values a and b). a = smallest X, b = largest X

$X \sim U(a, b)$

The mean is $\mu = \frac{a+b}{2}$.

The standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$.

Probability density function: $f(x) = \frac{1}{b-a}$ for $a \leq X \leq b$

Area to the left of x: $P(X < x) = (x - a)\left(\frac{1}{b-a}\right)$

Area to the right of x: $P(X > x) = (b - x)\left(\frac{1}{b-a}\right)$

Area between c and d: $P(c < x < d)$ = (base)(height) = $(d - c)\left(\frac{1}{b-a}\right)$

Uniform: $X \sim U(a, b)$ where $a < x < b$

- pdf: $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$
- cdf: $P(X \leq x) = \frac{x-a}{b-a}$
- mean $\mu = \frac{a+b}{2}$
- standard deviation $\sigma = \sqrt{\frac{(b-a)^2}{12}}$
- $P(c < X < d) = (d - c)(\frac{1}{b-a})$

5.3 The Exponential Distribution (Optional)

Exponential: $X \sim Exp(m)$ where m = the decay parameter

- pdf: $f(x) = me^{(-mx)}$ where $x \geq 0$ and $m > 0$
- cdf: $P(X \leq x) = 1 - e^{(-mx)}$
- mean $\mu = \frac{1}{m}$
- standard deviation $\sigma = \mu$
- percentile k: $k = \frac{ln(1-AreaToTheLeftOfk)}{(-m)}$
- Additionally
  - $P(X > x) = e^{(-mx)}$
  - $P(a < X < b) = e^{(-ma)} - e^{(-mb)}$
- Memoryless property: $P(X > x + k | X > x) = P(X > k)$
- Poisson probability: $P(X = k) = \frac{\lambda^k e^{-k}}{k!}$ with mean λ
- $k! = k*(k-1)*(k-2)*(k-3)*...3*2*1$

Practice

5.1 Continuous Probability Functions

1. Which type of distribution does the graph illustrate?

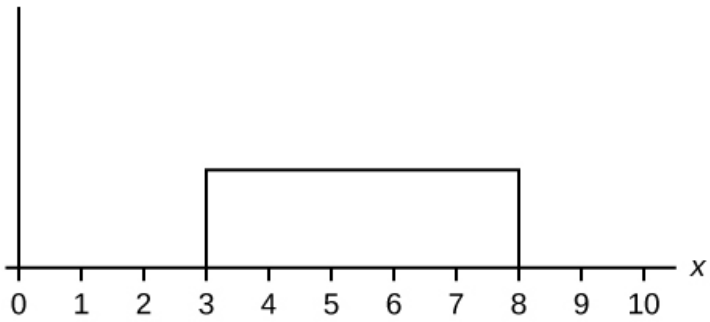


Figure 5.37

2. Which type of distribution does the graph illustrate?

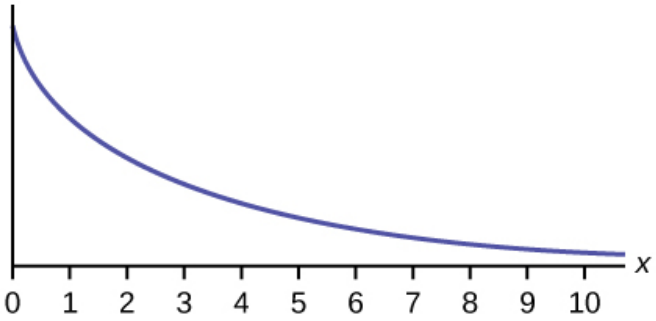


Figure 5.38

3. Which type of distribution does the graph illustrate?

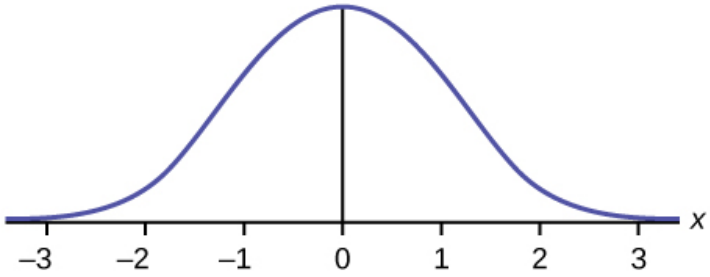


Figure 5.39

4. What does the shaded area represent? $P(___ < x < ___)$

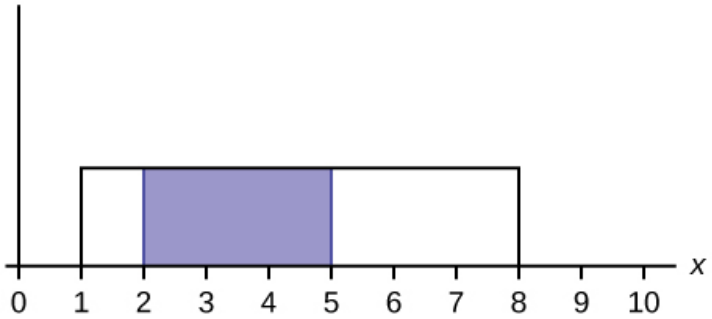


Figure 5.40

5. What does the shaded area represent? $P(___ < x < ___)$

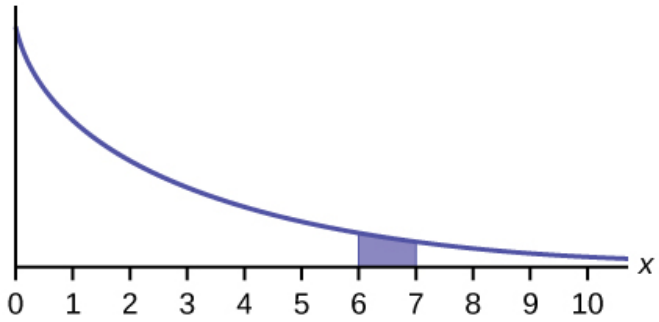


Figure 5.41

6. For a continuous probablity distribution, $0 \leq x \leq 15$. What is $P(x > 15)$?
7. What is the area under $f(x)$ if the function is a continuous probability density function?
8. For a continuous probability distribution, $0 \leq x \leq 10$. What is $P(x = 7)$?
9. A **continuous** probability function is restricted to the portion between $x = 0$ and 7. What is $P(x = 10)$?
10. $f(x)$ for a continuous probability function is $\frac{1}{5}$, and the function is restricted to $0 \leq x \leq 5$. What is $P(x < 0)$?
11. $f(x)$, a continuous probability function, is equal to $\frac{1}{12}$, and the function is restricted to $0 \leq x \leq 12$. What is $P(0 < x < 12)$?
12. Find the probability that x falls in the shaded area.

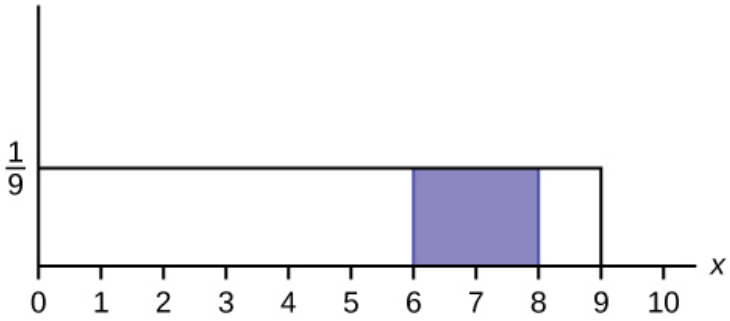


Figure 5.42

13. Find the probability that x falls in the shaded area.

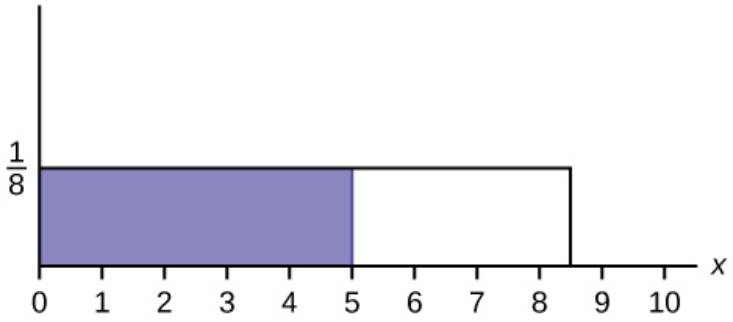


Figure 5.43

14. Find the probability that x falls in the shaded area.

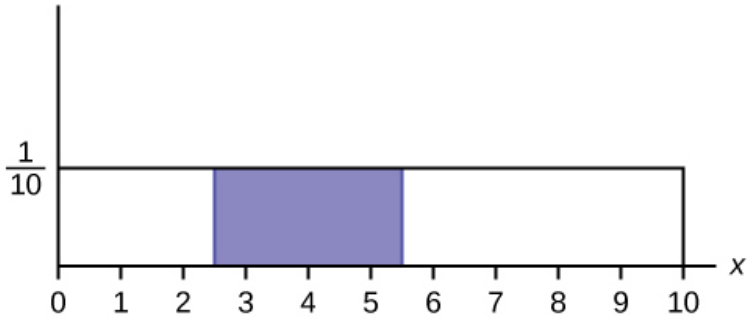


Figure 5.44

15. $f(x)$, a continuous probability function, is equal to $\frac{1}{3}$ and the function is restricted to $1 \leq x \leq 4$. Describe $P\left(x > \frac{3}{2}\right)$.

5.2 The Uniform Distribution

Use the following information to answer the next 10 questions. The data that follow are the square footage (in 1,000 feet squared) of 28 homes:

1.5	2.4	3.6	2.6	1.6	2.4	2.0
3.5	2.5	1.8	2.4	2.5	3.5	4.0
2.6	1.6	2.2	1.8	3.8	2.5	1.5
2.8	1.8	4.5	1.9	1.9	3.1	1.6

Table 5.4

The sample mean = 2.50 and the sample standard deviation = 0.8302.

The distribution can be written as $X \sim U(1.5, 4.5)$.

16. What type of distribution is this?

17. In this distribution, outcomes are equally likely. What does this mean?

18. What is the height of $f(x)$ for the continuous probability distribution?

19. What are the constraints for the values of x?

20. Graph $P(2 < x < 3)$.

21. What is $P(2 < x < 3)$?

22. What is $P(x < 3.5 \mid x < 4)$?

23. What is $P(x = 1.5)$?

24. What is the 90^{th} percentile of square footage for homes?

25. Find the probability that a randomly selected home has more than 3,000 square feet given that you already know the house has more than 2,000 square feet.

Use the following information to answer the next eight exercises. A distribution is given as $X \sim U(0, 12)$.

26. What is a? What does it represent?

27. What is b? What does it represent?

28. What is the probability density function?

29. What is the theoretical mean?

30. What is the theoretical standard deviation?

31. Draw the graph of the distribution for $P(x > 9)$.

32. Find $P(x > 9)$.

33. Find the 40^{th} percentile.

Use the following information to answer the next 12 exercises. The age of cars in the staff parking lot of a suburban college is uniformly distributed from six months (0.5 years) to 9.5 years.

34. What is being measured here?

35. In words, define the random variable X.

36. Are the data discrete or continuous?

37. The interval of values for x is _____.

38. The distribution for X is _____.

39. Write the probability density function.

40. Graph the probability distribution.

a. Sketch the graph of the probability distribution.

Figure 5.45

b. Identify the following values:
   i. Lowest value for $\overline{x}$: _____
   ii. Highest value for $\overline{x}$: _____
   iii. Height of the rectangle: _____
   iv. Label for x-axis (words): _____
   v. Label for y-axis (words): _____

41. Find the average age of the cars in the lot.

42. Find the probability that a randomly chosen car in the lot was less than four years old.

a. Sketch the graph, and shade the area of interest.

Figure 5.46

b. Find the probability. $P(x < 4) =$ _____

43. Considering only the cars less than 7.5 years old, find the probability that a randomly chosen car in the lot was less than four years old.

a. Sketch the graph, shade the area of interest.

Figure 5.47

b. Find the probability. $P(x < 4 \mid x < 7.5) =$ _____

44. What has changed in the previous two problems that made the solutions different?

45. Find the third quartile of ages of cars in the lot. This means you will have to find the value such that $\frac{3}{4}$, or 75 percent, of the cars are at most (less than or equal to) that age.

a. Sketch the graph, and shade the area of interest.

Figure 5.48

b. Find the value k such that $P(x < k) = 0.75$.
c. The third quartile is _____

5.3 The Exponential Distribution (Optional)

Use the following information to answer the next 10 exercises. A customer service representative must spend different amounts of time with each customer to resolve various concerns. The amount of time spent with each customer can be modeled by the following distribution: $X \sim Exp(0.2)$

46. What type of distribution is this?

47. Are outcomes equally likely in this distribution? Why or why not?

48. What is m? What does it represent?

49. What is the mean?

50. What is the standard deviation?

51. State the probability density function.

52. Graph the distribution.

53. Find $P(2 < x < 10)$.

54. Find $P(x > 6)$.

55. Find the 70^{th} percentile.

Use the following information to answer the next eight exercises. A distribution is given as $X \sim Exp(0.75)$.

56. What is m?

57. What is the probability density function?

58. What is the cumulative distribution function?

59. Draw the distribution.

60. Find $P(x < 4)$.

61. Find the 30^{th} percentile.

62. Find the median.

63. Which is larger, the mean or the median?

Use the following information to answer the next eight exercises. Carbon-14 is a radioactive element with a half-life of about 5,730 years. Carbon-14 is said to decay exponentially. The decay rate is 0.000121. We start with one gram of carbon-14. We are interested in the time (years) it takes to decay carbon-14.

64. What is being measured here?

65. Are the data discrete or continuous?

66. In words, define the random variable X.

67. What is the decay rate (m)?

68. The distribution for X is _____.

69. Find the amount (percent of one gram) of carbon-14 lasting less than 5,730 years. The question means that you need to find $P(x < 5{,}730)$.

a. Sketch the graph, and shade the area of interest.

Figure 5.49

b. Find the probability. $P(x < 5{,}730)$ = _____

70. Find the percentage of carbon-14 lasting longer than 10,000 years.

a. Sketch the graph, and shade the area of interest.

Figure 5.50

b. Find the probability. $P(x > 10{,}000)$ = _____

71. Thirty percent of carbon-14 will decay within how many years?

a. Sketch the graph, and shade the area of interest.

Figure 5.51

b. Find the value k such that $P(x < k) = 0.30$.

Homework

5.1 Continuous Probability Functions

For each probability and percentile problem, draw the picture.

72. Consider the following experiment. You are one of 100 people enlisted to take part in a study to determine percentage of nurses in America with an R.N. (registered nurse) degree. You ask nurses if they have an R.N. degree. The nurses answer yes or no. You then calculate the percentage of nurses with an R.N. degree. You give that percentage to your supervisor.
   a. What part of the experiment will yield discrete data?
   b. What part of the experiment will yield continuous data?

73. When age is rounded to the nearest year, do the data stay continuous, or do they become discrete? Why?

5.2 The Uniform Distribution

For each probability and percentile problem, draw the picture.

74. Births are approximately uniformly distributed between the 52 weeks of the year. They can be said to follow a uniform distribution from one to 53 (spread of 52 weeks).
   a. $X \sim$ _____
   b. Graph the probability distribution.
   c. $f(x) =$ _____
   d. $\mu =$ _____
   e. $\sigma =$ _____
   f. Find the probability that a person is born at the exact moment week 19 starts. That is, find $P(x = 19) =$ _____
   g. $P(2 < x < 31) =$ _____
   h. Find the probability that a person is born after week 40.
   i. $P(12 < x \mid x < 28) =$ _____
   j. Find the 70^{th} percentile.
   k. Find the minimum for the upper quarter.

75. A random number generator picks a number from one to nine in a uniform manner.
   a. $X \sim$ _____
   b. Graph the probability distribution.
   c. $f(x) =$ _____
   d. $\mu =$ _____
   e. $\sigma =$ _____
   f. $P(3.5 < x < 7.25) =$ _____
   g. $P(x > 5.67)$
   h. $P(x > 5 \mid x > 3) =$ _____
   i. Find the 90^{th} percentile.

76. According to a study by Dr. John McDougall of his live-in weight loss program, the people who follow his program lose between six and 15 pounds a month until they approach trim body weight. Let's suppose that the weight loss is uniformly distributed. We are interested in the weight loss of a randomly selected individual following the program for one month.

a. Define the random variable. X = _____
b. X ~ _____
c. Graph the probability distribution.
d. $f(x)$ = _____
e. μ = _____
f. σ = _____
g. Find the probability that the individual lost more than 10 pounds in a month.
h. Suppose it is known that the individual lost more than 10 pounds in a month. Find the probability that he lost less than 12 pounds in the month.
i. $P(7 < x < 13 \mid x > 9)$ = _____. State this result in a probability question, similarly to Parts g and h, draw the picture, and find the probability.

77. A subway train arrives every eight minutes during rush hour. We are interested in the length of time a commuter must wait for a train to arrive. The time follows a uniform distribution.

a. Define the random variable. X = _____
b. X ~ _____
c. Graph the probability distribution.
d. $f(x)$ = _____
e. μ = _____
f. σ = _____
g. Find the probability that the commuter waits less than one minute.
h. Find the probability that the commuter waits between three and four minutes.
i. Sixty percent of commuters wait more than how long for the train? State this result in a probability question, similarly to Parts g and h, draw the picture, and find the probability.

78. The age of a first grader on September 1 at Garden Elementary School is uniformly distributed from 5.8 to 6.8 years. We randomly select one first grader from the class.

a. Define the random variable. X = _____
b. X ~ _____
c. Graph the probability distribution.
d. $f(x)$ = _____
e. μ = _____
f. σ = _____
g. Find the probability that she is over 6.5 years old.
h. Find the probability that she is between four and six years old.
i. Find the 70^{th} percentile for the age of first graders on September 1 at Garden Elementary School.

Use the following information to answer the next three exercises. The Sky Train from the terminal to the rental–car and long–term parking center is supposed to arrive every eight minutes. The waiting times for the train are known to follow a uniform distribution.

79. What is the average waiting time (in minutes)?

a. zero
b. two
c. three
d. four

80. Find the 30^{th} percentile for the waiting times (in minutes).

a. two
b. 2.4
c. 2.75
d. three

81. The probability of waiting more than seven minutes given a person has waited more than four minutes is?

a. 0.125
b. 0.25
c. 0.5
d. 0.75

82. The time (in minutes) until the next bus departs a major bus depot follows a distribution with $f(x) = \frac{1}{20}$ where x goes from 25 to 45 minutes.
a. Define the random variable. $X =$ _____
b. $X \sim$ _____
c. Graph the probability distribution.
d. The distribution is _____ (name of distribution). It is _____ (discrete or continuous).
e. $\mu =$ _____
f. $\sigma =$ _____
g. Find the probability that the time is at most 30 minutes. Sketch and label a graph of the distribution. Shade the area of interest. Write the answer in a probability statement.
h. Find the probability that the time is between 30 and 40 minutes. Sketch and label a graph of the distribution. Shade the area of interest. Write the answer in a probability statement.
i. $P(25 < x < 55) =$ _____. State this result in a probability statement, similarly to Parts g and h, draw the picture, and find the probability.
j. Find the 90^{th} percentile. This means that 90 percent of the time, the time is less than _____ minutes.
k. Find the 75^{th} percentile. In a complete sentence, state what this means. (See Part j.)
l. Find the probability that the time is more than 40 minutes given (or knowing that) it is at least 30 minutes.

83. Suppose that the value of a stock varies each day from $16 to $25 with a uniform distribution.
a. Find the probability that the value of the stock is more than $19.
b. Find the probability that the value of the stock between $19 and $22.
c. Find the upper quartile — 25 percent of all days the stock is above what value? Draw the graph.
d. Given that the stock is greater than $18, find the probability that the stock is more than $21.

84. A fireworks show is designed so that the time between fireworks is between one and five seconds, and follows a uniform distribution.
a. Find the average time between fireworks.
b. Find the probability that the time between fireworks is greater than four seconds.

85. The number of miles driven by a truck driver falls between 300 and 700, and follows a uniform distribution.
a. Find the probability that the truck driver goes more than 650 miles in a day.
b. Find the probability that the truck driver goes between 400 and 650 miles in a day.
c. At least how many miles does the truck driver travel on the 10 percent of days with the highest mileage?

5.3 The Exponential Distribution (Optional)

86. Suppose that the length of long-distance phone calls, measured in minutes, is known to have an exponential distribution with the average length of a call equal to eight minutes.
a. Define the random variable. $X =$ _____.
b. Is X continuous or discrete?
c. $X \sim$ _____
d. $\mu =$ _____
e. $\sigma =$ _____
f. Draw a graph of the probability distribution. Label the axes.
g. Find the probability that a phone call lasts less than nine minutes.
h. Find the probability that a phone call lasts more than nine minutes.
i. Find the probability that a phone call lasts between seven and nine minutes.
j. If 25 phone calls are made one after another, on average, what would you expect the total to be? Why?

87. Suppose that the useful life of a particular car battery, measured in months, decays with parameter 0.025. We are interested in the life of the battery.
a. Define the random variable. $X =$ _____.
b. Is X continuous or discrete?

c. $X \sim$ _____
d. On average, how long would you expect one car battery to last?
e. On average, how long would you expect nine car batteries to last, if they are used one after another?
f. Find the probability that a car battery lasts more than 36 months.
g. Seventy percent of the batteries last at least how long?

88. The percent of persons (ages five and older) in each state who speak a language at home other than English is approximately exponentially distributed with a mean of 9.848. Suppose we randomly pick a state.
a. Define the random variable. X = _____.
b. Is X continuous or discrete?
c. $X \sim$ _____
d. μ = _____
e. σ = _____
f. Draw a graph of the probability distribution. Label the axes.
g. Find the probability that percentage is less than 12.
h. Find the probability that percentage is between eight and 14.
i. The percent of all individuals living in the United States who speak a language at home other than English is 13.8.
   i. Why is this number different from 9.848 percent?
   ii. What would make this number higher than 9.848 percent?

89. The time (in years) **after** reaching age 60 that it takes an individual to retire is approximately exponentially distributed with a mean of about five years. Suppose we randomly pick one retired individual. We are interested in the time after age 60 to retirement.
a. Define the random variable. X = _____.
b. Is X continuous or discrete?
c. $X \sim$ = _____
d. μ = _____
e. σ = _____
f. Draw a graph of the probability distribution. Label the axes.
g. Find the probability that the person retired after age 70.
h. Do more people retire before age 65 or after age 65?
i. In a room of 1,000 people over age 80, how many do you expect will **not** have retired yet?

90. The cost of all maintenance for a car during its first year is approximately exponentially distributed with a mean of $150.
a. Define the random variable. X = _____.
b. $X \sim$ = _____
c. μ = _____
d. σ = _____
e. Draw a graph of the probability distribution. Label the axes.
f. Find the probability that a car required over $300 for maintenance during its first year.

Use the following information to answer the next three exercises. The average lifetime of a certain new cell phone is three years. The manufacturer will replace any cell phone failing within two years of the date of purchase. The lifetime of these cell phones is known to follow an exponential distribution.

91. What is the decay rate?
a. 0.3333
b. 0.5000
c. 2
d. 3

92. What is the probability that a phone will fail within two years of the date of purchase?
a. 0.8647
b. 0.4866
c. 0.2212

d. 0.9997

93. What is the median lifetime of these phones (in years)?
a. 0.1941
b. 1.3863
c. 2.0794
d. 5.5452

94. Let $X \sim Exp(0.1)$.
a. decay rate = _____
b. μ = _____
c. Graph the probability distribution function.
d. On the graph, shade the area corresponding to $P(x < 6)$, and find the probability.
e. Sketch a new graph, shade the area corresponding to $P(3 < x < 6)$, and find the probability.
f. Sketch a new graph, shade the area corresponding to $P(x < 7)$, and find the probability.
g. Sketch a new graph, shade the area corresponding to the 40^{th} percentile and find the value.
h. Find the average value of x.

95. Suppose that the longevity of a light bulb is exponential with a mean lifetime of eight years.
a. Find the probability that a light bulb lasts less than one year.
b. Find the probability that a light bulb lasts between six and 10 years.
c. Seventy percent of all light bulbs last at least how long?
d. A company decides to offer a warranty to give refunds to light bulbs whose lifetime is among the lowest two percent of all bulbs. To the nearest month, what should be the cutoff lifetime for the warranty to take place?
e. If a light bulb has lasted seven years, what is the probability that it fails within the 8^{th} year?

96. At a 911 call center, calls come in at an average rate of one call every two minutes. Assume that the time that elapses from one call to the next has the exponential distribution.
a. On average, how much time occurs between five consecutive calls?
b. Find the probability that after a call is received, it takes more than three minutes for the next call to occur.
c. Ninety-percent of all calls occur within how many minutes of the previous call?
d. Suppose that two minutes have elapsed since the last call. Find the probability that the next call will occur within the next minute.
e. Find the probability that fewer than 20 calls occur within an hour.

97. In major league baseball, a no-hitter is a game in which a pitcher, or pitchers, doesn't give up any hits throughout the game. No-hitters occur at a rate of about three per season. Assume that the duration of time between no-hitters is exponential.
a. What is the probability that an entire season elapses with a single no-hitter?
b. If an entire season elapses without any no-hitters, what is the probability that there are no no-hitters in the following season?
c. What is the probability that there are more than three no-hitters in a single season?

98. During the years 1998–2012, a total of 29 earthquakes of magnitude greater than 6.5 occurred in Papua New Guinea. Assume that the time spent waiting between earthquakes is exponential. Assume that the current year is 2013
a. What is the probability that the next earthquake occurs within the next three months?
b. Given that six months has passed without an earthquake in Papua New Guinea, what is the probability that the next three months will be free of earthquakes?
c. What is the probability of zero earthquakes occurring in 2014?
d. What is the probability that at least two earthquakes will occur in 2014?

99. According to the American Red Cross, about one out of nine people in the United States have type B blood. Suppose the blood types of people arriving at a blood drive are independent. In this case, the number of type B blood types that arrive roughly follows the Poisson distribution.
a. If 100 people arrive, how many on average would be expected to have type B blood?
b. What is the probability that more than 10 people out of these 100 have type B blood?
c. What is the probability that more than 20 people arrive before a person with type B blood is found?

100. A website experiences traffic during normal working hours at a rate of 12 visits per hour. Assume that the duration between visits has the exponential distribution.
   a. Find the probability that the duration between two successive visits to the website is more than 10 minutes.
   b. The top 25 percent of durations between visits are at least how long?
   c. Suppose that 20 minutes have passed since the last visit to the website. What is the probability that the next visit will occur within the next five minutes?
   d. Find the probability that fewer than seven visits occur within a one-hour period.

101. At an urgent care facility, patients arrive at an average rate of one patient every seven minutes. Assume that the duration between arrivals is exponentially distributed.
   a. Find the probability that the time between two successive visits to the urgent care facility is less than two minutes.
   b. Find the probability that the time between two successive visits to the urgent care facility is more than 15 minutes.
   c. If 10 minutes have passed since the last arrival, what is the probability that the next person will arrive within the next five minutes?
   d. Find the probability that more than eight patients arrive during a half-hour period.

References

5.2 The Uniform Distribution

McDougall, J. A. (1995). *The McDougall program for maximum weight loss*. New York: Plume

5.3 The Exponential Distribution (Optional)

Baseball-Reference.com. (2013). *No-hitter*. Retrieved from http://www.baseball-reference.com/bullpen/No-hitter

U.S. Census Bureau. (n.d.). Retrieved from https://www.census.gov/

World Earthquakes. (2013). *Earthquake data for Papua New Guinea*. Retrieved from http://www.world-earthquakes.com/

Zhou, Rick. (2013). *Exponential distribution lecture slides*. Retrieved from www.public.iastate.edu/~riczw/stat330s11/lecture/lec13.pdf

Solutions

1. Uniform distribution

3. Normal distribution

5. $P(6 < x < 7)$

7. one

9. zero

11. one

13. 0.625

15. The probability is equal to the area from $x = \frac{3}{2}$ to $x = 4$ above the x-axis and up to $f(x) = \frac{1}{3}$.

17. It means that the value of x is just as likely to be any number between 1.5 and 4.5.

19. $1.5 \leq x \leq 4.5$

21. 0.3333

23. zero

25. 0.6

27. b is 12, and it represents the highest value of x.

29. six

31.

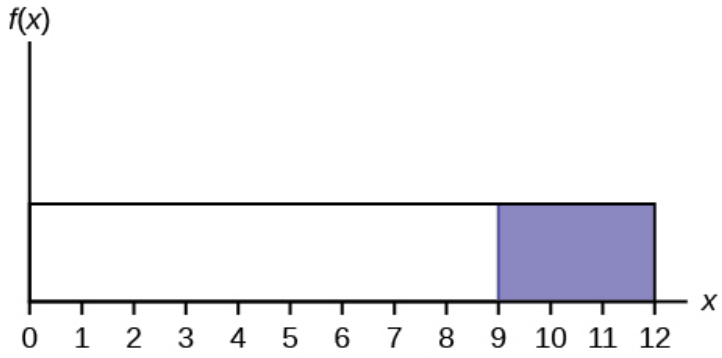


Figure 5.52

33. 4.8

35. X = The age (in years) of cars in the staff parking lot

37. 0.5 to 9.5

39. $f(x) = \frac{1}{9}$ where x is between 0.5 and 9.5, inclusive.

41. $\mu = 5$

43. a. Check student's solution.
b. $\frac{3.5}{7}$

45. a. Check student's solution
b. $k = 7.25$
c. 7.25

47. No, outcomes are not equally likely. In this distribution, more people require a little bit of time, and fewer people require a lot of time, so it is more likely that someone will require less time.

49. five

51. $f(x) = 0.2e^{-0.2x}$

53. 0.5350

55. 6.02

57. $f(x) = 0.75e^{-0.75x}$

59.

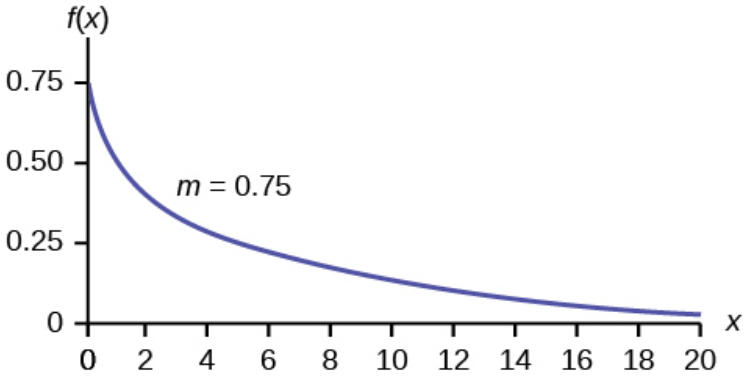


Figure 5.53

61. 0.4756

63. The mean is larger. The mean is $\frac{1}{m} = \frac{1}{0.75} \approx 1.33$, which is greater than 0.9242.

65. continuous

67. $m = 0.000121$

69.
a. Check student's solution
b. $P(x < 5{,}730) = 0.5001$

71.
a. Check student's solution
b. $k = 2947.73$

73. Age is a measurement, regardless of the accuracy used.

75.
a. $X \sim U(1, 9)$
b. Check student's solution
c. $f(x) = \frac{1}{8}$ where $1 \leq x \leq 9$
d. five
e. 2.3
f. $\frac{15}{32}$
g. $\frac{333}{800}$
h. $\frac{2}{3}$
i. 8.2

77.
a. X represents the length of time a commuter must wait for a train to arrive on the Red Line.
b. $X \sim U(0, 8)$
c. $f(x) = \frac{1}{8}$ where $\leq x \leq 8$
d. four
e. 2.31
f. $\frac{1}{8}$
g. $\frac{1}{8}$
h. 3.2

79. d

81. b

83. a. The probability density function of X is $\frac{1}{25-16} = \frac{1}{9}$.
$P(X > 19) = (25 - 19)\left(\frac{1}{9}\right) = \frac{6}{9} = \frac{2}{3}$.

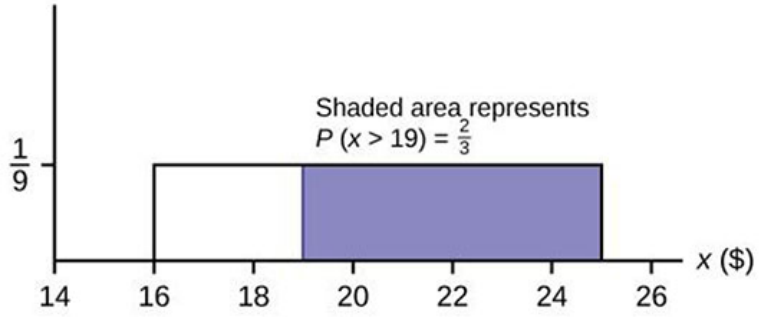


Figure 5.54

b. $P(19 < X < 22) = (22 - 19)\left(\frac{1}{9}\right) = \frac{3}{9} = \frac{1}{3}$.

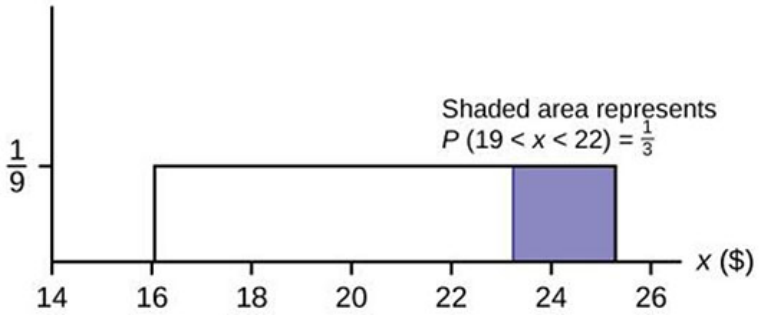


Figure 5.55

c. The area must be 0.25, and $0.25 = (\text{width})\left(\frac{1}{9}\right)$, so width = (0.25)(9) = 2.25. Thus, the value is 25 – 2.25 = 22.75.

d. This is a conditional probability question. P(x > 21 | x > 18). You can do this two ways:
- Draw the graph where a is now 18 and b is still 25. The height is $\frac{1}{(25-18)} = \frac{1}{7}$
So, $P(x > 21 \mid x > 18) = (25 - 21)\left(\frac{1}{7}\right) = 4/7$.
- Use the formula: $P(x > 21 \mid x > 18) = \frac{P(x>21 \text{ AND } x>18)}{P(x>18)}$
$= \frac{P(x>21)}{P(x>18)} = \frac{(25-21)}{(25-18)} = \frac{4}{7}$.

85. a. $P(X > 650) = \frac{700-650}{700-300} = \frac{50}{400} = \frac{1}{8} = 0.125$.

b. $P(400 < X < 650) = \frac{650-400}{700-300} = \frac{250}{400} = 0.625$

c. $0.10 = \frac{\text{width}}{700-300}$, so width = 400(0.10) = 40. Since 700 – 40 = 660, the drivers travel at least 660 miles on the farthest 10 percent of days.

87. a. X = the useful life of a particular car battery, measured in months.
b. X is continuous.
c. $X \sim Exp(0.025)$
d. 40 months
e. 360 months
f. 0.4066
g. 14.27

89. a. X = the time (in years) after reaching age 60 that it takes an individual to retire
b. X is continuous.
c. $X \sim Exp\left(\frac{1}{5}\right)$
d. five
e. five

f. Check student's solution.
g. 0.1353
h. before
i. 18.3

91. a

93. c

95. Let T = the life time of a light bulb.

The decay parameter is $m = 1/8$, and $T \sim \text{Exp}(1/8)$. The cumulative distribution function is $P(T < t) = 1 - e^{-\frac{t}{8}}$

a. Therefore, $P(T < 1) = 1 - e^{-\frac{1}{8}} \approx 0.1175$.

b. We want to find $P(6 < t < 10)$.
To do this, $P(6 < t < 10) - P(t < 6)$
$= = \left(1 - e^{-\frac{1}{8}*10}\right) - \left(1 - e^{-\frac{1}{8}*6}\right) \approx 0.7135 - 0.5276 = 0.1859$

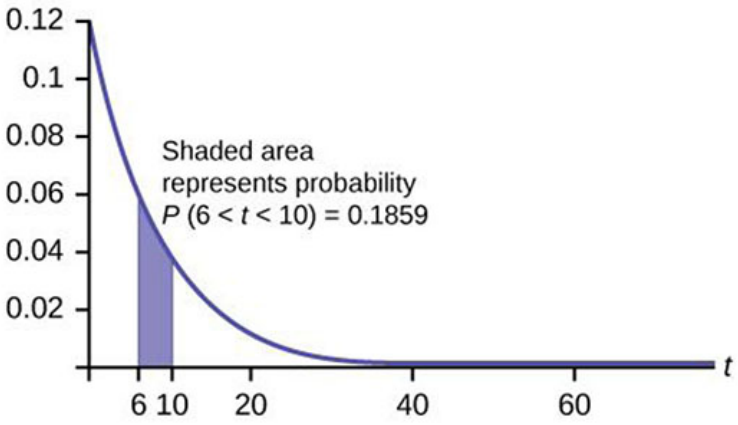


Figure 5.56

c. We want to find $0.70 = P(T > t) = 1 - \left(1 - e^{-\frac{t}{8}}\right) = e^{-\frac{t}{8}}$.
Solving for t, $e^{-\frac{t}{8}} = 0.70$, so $-\frac{t}{8} = ln(0.70)$, and $t = -8ln(0.70) \approx 2.85$ years
Or use $t = \frac{ln(\text{area_to_the_right})}{(-m)} = \frac{ln(0.70)}{-\frac{1}{8}} \approx 2.85$ years.

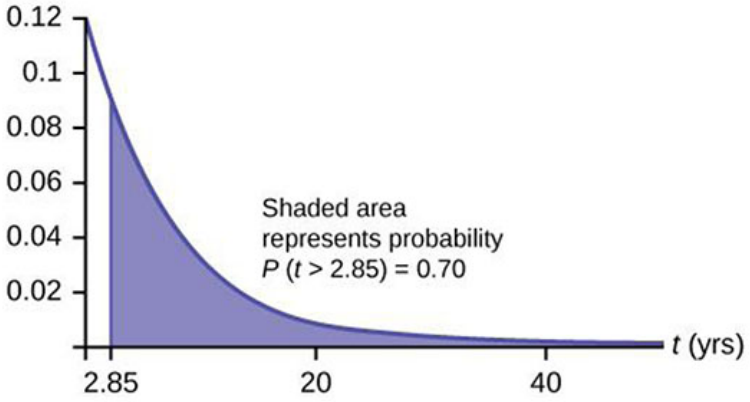


Figure 5.57

d. We want to find $0.02 = P(T < t) = 1 - e^{-\frac{t}{8}}$.
Solving for t, $e^{-\frac{t}{8}} = 0.98$, so $-\frac{t}{8} = ln(0.98)$, and $t = -8ln(0.98) \approx 0.1616$ years, or roughly two months.
The warranty should cover light bulbs that last less than 2 months.

Or use $\frac{\ln(\text{area_to_the_right})}{(-m)} = \frac{\ln(1-0.2)}{-\frac{1}{8}} = 0.1616$.

e. We must find $P(T < 8 \mid T > 7)$.
Notice that by the rule of complement events, $P(T < 8 \mid T > 7) = 1 - P(T > 8 \mid T > 7)$.
By the memoryless property ($P(X > r + t \mid X > r) = P(X > t)$).
So $P(T > 8 \mid T > 7) = P(T > 1) = 1 - \left(1 - e^{-\frac{1}{8}}\right) = e^{-\frac{1}{8}} \approx 0.8825$
Therefore, $P(T < 8 \mid T > 7) = 1 - 0.8825 = 0.1175$.

97. Let X = the number of no-hitters throughout a season. Since the duration of time between no-hitters is exponential, the number of no-hitters per season is Poisson with mean $\lambda = 3$.
Therefore, $(X = 0) = \frac{3^0 e^{-3}}{0!} = e^{-3} \approx 0.0498$

You could let T = duration of time between no-hitters. Since the time is exponential and there are three no-hitters per season, then the time between no-hitters is $\frac{1}{3}$ season. For the exponential, $\mu = \frac{1}{3}$.
Therefore, $m = \frac{1}{\mu} = 3$ and $T \sim Exp(3)$.

a. The desired probability is $P(T > 1) = 1 - P(T < 1) = 1 - (1 - e^{-3}) = e^{-3} \approx 0.0498$.
b. Let T = duration of time between no-hitters. We find $P(T > 2 \mid T > 1)$, and by the **memoryless property** this is simply $P(T > 1)$, which we found to be 0.0498 in part a.
c. Let X = the number of no-hitters is a season. Assume that X is Poisson with mean $\lambda = 3$. Then $P(X > 3) = 1 - P(X \leq 3) = 0.3528$.

99. a. $\frac{100}{9} = 11.11$
b. $P(X > 10) = 1 - P(X \leq 10) = 1 - \text{Poissoncdf}(11.11, 10) \approx 0.5532$.
c. The number of people with Type B blood encountered roughly follows the Poisson distribution, so the number of people X who arrive between successive Type B arrivals is roughly exponential with mean $\mu = 9$ and $m = \frac{1}{9}$.
The cumulative distribution function of X is $P(X < x) = 1 - e^{-\frac{x}{9}}$. Thus hus, $P(X > 20) = 1 - P(X \leq 20) = 1 - \left(1 - e^{-\frac{20}{9}}\right) \approx 0.1084$.

Note

We could also deduce that each person arriving has a $\frac{8}{9}$ chance of not having type B blood. So the probability that none of the first 20 people arrive have type B blood is $\left(\frac{8}{9}\right)^{20} \approx 0.0948$. (The geometric distribution is more appropriate than the exponential because the number of people between type B people is discrete instead of continuous.)

101. Let T = duration (in minutes) between successive visits. Since patients arrive at a rate of one patient every seven minutes, $\mu = 7$ and the decay constant is $m = \frac{1}{7}$. The cdf is $P(T < t) = 1 - e^{\frac{t}{7}}$

a. $P(T < 2) = 1 - 1 - e^{-\frac{2}{7}} \approx 0.2485$.
b. $P(T > 15) = 1 - P(T < 15) = 1 - \left(1 - e^{-\frac{15}{7}}\right) \approx e^{-\frac{15}{7}} \approx 0.1173$.
c. $P(T > 15 \mid T > 10) = P(T > 5) = 1 - \left(1 - e^{-\frac{5}{7}}\right) = e^{-\frac{5}{7}} \approx 0.4895$.
d. Let X = # of patients arriving during a half-hour period. Then X has the Poisson distribution with a mean of $\frac{30}{7}$, $X \sim \text{Poisson}\left(\frac{30}{7}\right)$. Find $P(X > 8) = 1 - P(X \leq 8) \approx 0.0311$.

6 THE NORMAL DISTRIBUTION

Figure 6.1 If you ask enough people about their shoe size, you will find that your graphed data is shaped like a bell curve and can be described as normally distributed. (credit: Ömer Ünlü)

Chapter Objectives

By the end of this chapter, the student should be able to do the following:

- Recognize the normal probability distribution and apply it appropriately
- Recognize the standard normal probability distribution and apply it appropriately
- Compare normal probabilities by converting to the standard normal distribution

Introduction

The normal, a continuous distribution, is the most important of all the distributions. It is widely used and even more widely abused. Its graph is bell-shaped. You see the bell curve in almost all disciplines, including psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may use the **normal distribution** to help determine your grade. Most IQ scores are normally distributed. Often, real-estate prices fit a normal distribution. The normal distribution is extremely important, but it cannot be applied to everything in the real world.

In this chapter, you will study the normal distribution, the standard normal distribution, and applications associated with them.

The normal distribution has two parameters: —the mean (μ) and the standard deviation (σ). If X is a quantity to be measured that has a normal distribution with mean (μ) and standard deviation (σ), we designate this by writing

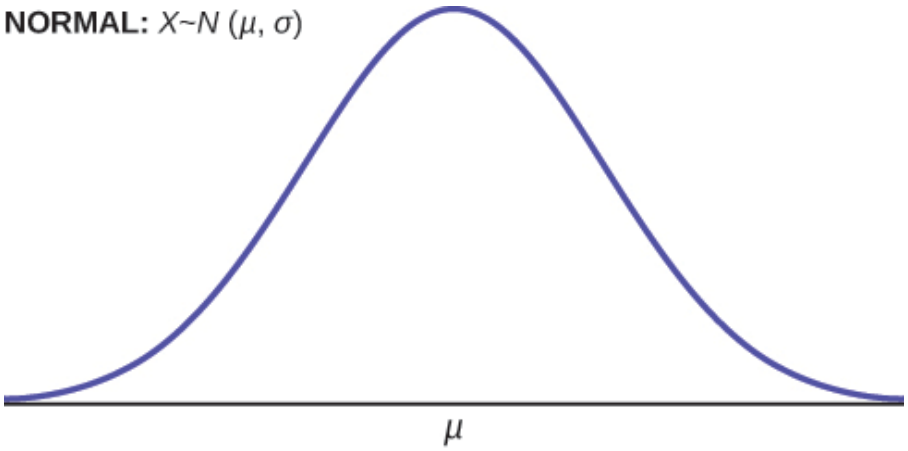


Figure 6.2

The curve is symmetric about a vertical line drawn through the mean, μ. In theory, the mean is the same as the median, because the graph is symmetric about μ. With a normal distribution, the mean, median, and mode all lie at the same point. The normal distribution depends only on the mean and the standard deviation. The location of the mean simply indicates the location of the line of symmetry, in a normal distribution. Since the area under the curve must equal one, a change in the standard deviation, σ, causes a change in the shape of the curve; the curve becomes fatter or skinnier depending on σ. A change in μ causes the graph to shift to the left or right. The location of the mean simply indicates the location of the line of symmetry, in a normal distribution. This means there are an infinite number of normal probability distributions. One distribution of special interest is called the **standard normal distribution**.

COLLABORATIVE EXERCISE

Your instructor will record the heights of both men and women in your class, separately. Draw histograms of your data. Then draw a smooth curve through each histogram. Is each curve somewhat bell-shaped? Do you think that if you had recorded 200 data values for men and 200 for women that the curves would look bell-shaped? Calculate the mean for each data set. Write the means on the *x*-axis of the appropriate graph below the peak. Shade the approximate area that represents the probability that one randomly chosen male is taller than 72 inches. Shade the approximate area that represents the probability that one randomly chosen female is shorter than 60 inches. If the total area under each curve is one, does either probability appear to be more than 0.5?

6.1 The Standard Normal Distribution

The standardized normal distribution is a type of normal distribution, with a mean of 0 and standard deviation of 1. It represents a distribution of standardized scores, called ***z*-scores**, as opposed to raw scores (the actual data values). A ***z*-score** indicates the number of standard deviation a score falls above or below the mean. *Z*-scores allow for comparison of scores, occurring in different data sets, with different means and standard deviations. It would not make sense to compare apples and oranges. Likewise, it does not make sense to compare scores from two different samples that have different means and standard deviations. *Z*-scores can be looked up in a *Z*-Table of Standard Normal Distribution, in order to find the area under the standard normal curve, between a score and the mean, between two scores, or above or below a score. The standard normal distribution allows us to interpret standardized scores and provides us with one table that we may use, in order to compute areas under the normal curve, for an infinite number of data sets, no matter what the mean or standard deviation.

A *z*-score is calculated as $z = \frac{x-\mu}{\sigma}$. The score itself can be found by using algebra and solving for *x*. Multiplying both sides of the equation by σ gives: $(z)(\sigma) = x - \mu$. Adding μ to both sides of the equation gives $\mu + (z)(\sigma) = x$.

Suppose we have a data set with a mean of 5 and standard deviation of 2. We want to determine the number of standard deviations the score of 11 falls above the mean. We can find this answer (or *z*-score) by writing

$$z = \frac{11-5}{2} = 3 \quad \textbf{6.1}$$

or

$$5 + (z)(2) = 11, \tag{6.2}$$

we can solve for z.

$$2z = 6$$
$$z = 3 \tag{6.3}$$

We have determined that the score of 11 falls 3 standard deviations above the mean of 5.

With a standard normal distribution, we indicate the distribution by writing $Z \sim N(0, 1)$ which shows the normal distribution has a mean of 0 and standard deviation of 1. This notation simply indicates that a standard normal distribution is being used.

Z-Scores

As described previously, if X is a normally distributed random variable and $X \sim N(\mu, \sigma)$, then the z-score is

$$z = \frac{x - \mu}{\sigma}.$$

The z-score tells you how many standard deviations the value x is above, to the right of, or below, to the left of, the mean, μ. Values of x that are larger than the mean have positive z-scores, and values of x that are smaller than the mean have negative z-scores. If x equals the mean, then x has a z-score of zero.

When determining the z-score for an x-value, for a normal distribution, with a given mean and standard deviation, the notation above for a normal distribution, will be given.

EXAMPLE 6.1

Suppose $X \sim N(5, 6)$. This equation says that X is a normally distributed random variable with mean $\mu = 5$ and standard deviation $\sigma = 6$. Suppose $x = 17$. Then,

$$z = \frac{x - \mu}{\sigma} = \frac{17 - 5}{6} = 2.$$

This means that $x = 17$ is **two standard deviations** (2σ) above, or to the right, of the mean $\mu = 5$.

Notice that $5 + (2)(6) = 17$. The pattern is $\mu + z\sigma = x$.

Now suppose $x = 1$. Then, $z = \frac{x - \mu}{\sigma} = \frac{1 - 5}{6} = -0.67$, rounded to two decimal places.

This means that $x = 1$ is 0.67 standard deviations (-0.67σ) below or to the left of the mean $\mu = 5$. This z-score shows that $x = 1$ is less than 1 standard deviation below the mean of 5. Therefore, the score doesn't fall very far below the mean.

Summarizing, when z is positive, x is above or to the right of μ, and when z is negative, x is to the left of or below μ. Or, when z is positive, x is greater than μ, and when z is negative, x is less than μ. The absolute value of z indicates how far the score is from the mean, in either direction.

TRY IT 6.1

What is the z-score of x, when $x = 1$ and $X \sim N(12, 3)$?

EXAMPLE 6.2

Some doctors believe that a person can lose five pounds, on average, in a month by reducing his or her fat intake and by consistently exercising. Suppose weight loss has a normal distribution. Let X = the amount of weight lost, in pounds, by a person in a month. Use a standard deviation of two pounds. $X \sim N(5, 2)$. Fill in the blanks.

Problem

a. Suppose a person *lost* 10 pounds in a month. The z-score when $x = 10$ pounds is $z = 2.5$ (verify). This z-score tells you that $x = 10$ is _____ standard deviations to the _____ (right or left) of the mean ____ (What is the mean?).

Solution

a. This *z*-score tells you that $x = 10$ is **2.5** standard deviations to the *right* of the mean *five*.

Problem

b. Suppose a person *gained* three pounds, a negative weight loss. Then $z =$ _____. This *z*-score tells you that $x = –3$ is _____ standard deviations to the _____ (right or left) of the mean.

Solution

b. $z =$ **–4**. This *z*-score tells you that $x = –3$ is *four* standard deviations to the *left* of the mean.

Problem

c. Suppose the random variables X and Y have the following normal distributions: $X \sim N(5, 6)$ and $Y \sim N(2, 1)$. If $x = 17$, then $z = 2$. This was previously shown. If $y = 4$, what is z?

Solution

c. $z = \frac{y-\mu}{\sigma} = \frac{4-2}{1} = 2$, where $\mu = 2$ and $\sigma = 1$.

The *z*-score for $y = 4$ is $z = 2$. This means that four is $z = 2$ standard deviations to the right of the mean. Therefore, $x = 17$ and $y = 4$ are both two of *their own* standard deviations to the right of *their* respective means.

The z-score allows us to compare data that are scaled differently. To better understand the concept, suppose $X \sim N(5, 6)$ represents weight gains for one group of people who are trying to gain weight in a six-week period and $Y \sim N(2, 1)$ measures the same weight gain for a second group of people. A negative weight gain would be a weight loss. Since $x = 17$ and $y = 4$ are each two standard deviations to the right of their means, they represent the same, standardized weight gain *relative to their means*.

TRY IT 6.2

Fill in the blanks.

Jerome averages 16 points a game with a standard deviation of four points. $X \sim N(16, 4)$. Suppose Jerome scores 10 points in a game. The *z*-score when $x = 10$ is –1.5. This score tells you that $x = 10$ is ____ standard deviations to the _____ (right or left) of the mean_____ (What is the mean?).

The Empirical Rule

If X is a random variable and has a normal distribution with mean μ and standard deviation σ, then the **Empirical Rule** states the following:

- About 68 percent of the x values lie between $–1\sigma$ and $+1\sigma$ of the mean μ (within one standard deviation of the mean).
- About 95 percent of the x values lie between $–2\sigma$ and $+2\sigma$ of the mean μ (within two standard deviations of the mean).
- About 99.7 percent of the x values lie between $–3\sigma$ and $+3\sigma$ of the mean μ (within three standard deviations of the mean). Notice that almost all the x values lie within three standard deviations of the mean.
- The *z*-scores for $+1\sigma$ and $–1\sigma$ are +1 and –1, respectively.
- The *z*-scores for $+2\sigma$ and $–2\sigma$ are +2 and –2, respectively.
- The *z*-scores for $+3\sigma$ and $–3\sigma$ are +3 and –3, respectively.

So, in other words, this is that about 68 percent of the values lie between *z*-scores of –1 and 1, about 95% of the values lie between *z*-scores of –2 and 2, and about 99.7 percent of the values lie between *z*-scores of -3 and 3. These facts can be checked, by looking up the mean to *z* area in a *z*-table for each positive *z*-score and multiplying by 2.

The empirical rule is also known as the *68–95–99.7 rule*.

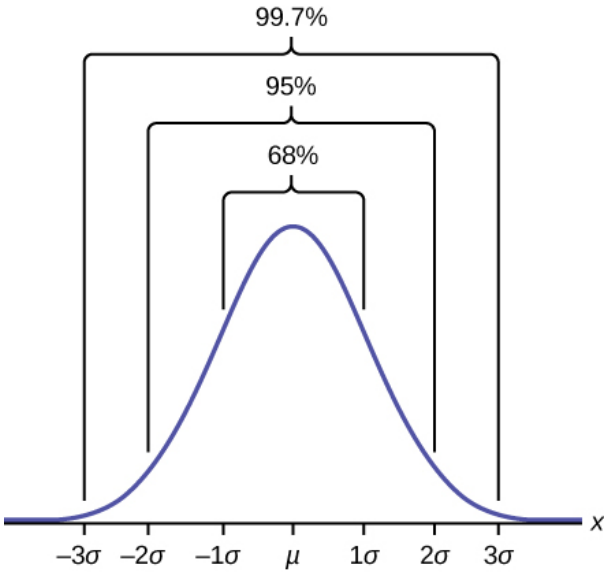


Figure 6.3

EXAMPLE 6.3

The mean height of 15-to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let X = the height of a 15-to 18-year-old male from Chile in 2009–2010. Then $X \sim N(170, 6.28)$.

Problem

a. Suppose a 15-to 18-year-old male from Chile was 168 cm tall in 2009–2010. The z-score when $x = 168$ cm is $z =$ _____. This z-score tells you that $x = 168$ is _____ standard deviations to the _____ (right or left) of the mean _____ (What is the mean?).

Solution

a. –0.32, 0.32, left, 170

Problem

b. Suppose that the height of a 15-to 18-year-old male from Chile in 2009–2010 has a z-score of $z = 1.27$. What is the male's height? The z-score ($z = 1.27$) tells you that the male's height is _____ standard deviations to the _____ (right or left) of the mean.

Solution

b. 177.98 cm, 1.27, right

TRY IT 6.3

Use the information in Example 6.3 to answer the following questions:

a. Suppose a 15-to 18-year-old male from Chile was 176 cm tall from 2009–2010. The z-score when $x = 176$ cm is $z =$ _____. This z-score tells you that $x = 176$ cm is _____ standard deviations to the _____ (right or left) of the mean _____ (What is the mean?).

b. Suppose that the height of a 15-to 18-year-old male from Chile in 2009–2010 has a z-score of $z = –2$. What is the male's height? The z-score ($z = –2$) tells you that the male's height is _____ standard deviations to the _____ (right or left) of the mean.

EXAMPLE 6.4

Problem

From 1984 to 1985, the mean height of 15-to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let Y = the height of 15-to 18-year-old males from 1984–1985, and y = the height of one male from this group. Then $Y \sim N(172.36, 6.34)$.

The mean height of 15-to 18-year-old males from Chile in 2009–2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let X = the height of a 15-to 18-year-old male from Chile in 2009–2010, and x = the height of one male from this group. Then $X \sim N(170, 6.28)$.

Find the z-scores for $x = 160.58$ cm and $y = 162.85$ cm. Interpret each z-score. What can you say about $x = 160.58$ cm and $y = 162.85$ cm as they compare to their respective means and standard deviations?

Solution

The z-score for $x = 160.58$ cm is $z = -1.5$.
The z-score for $y = 162.85$ cm is $z = -1.5$.
Both $x = 160.58$ and $y = 162.85$ deviate the same number of standard deviations from their respective means and in the same direction.

TRY IT 6.4

In 2012, 1,664,479 students took the SAT exam. The distribution of scores in the verbal section of the SAT had a mean $\mu = 496$ and a standard deviation $\sigma = 114$. Let X = a SAT exam verbal section score in 2012. Then, $X \sim N(496, 114)$.

Find the z-scores for $x_1 = 325$ and $x_2 = 366.21$. Interpret each z-score. What can you say about $x_1 = 325$ and $x_2 = 366.21$, as they compare to their respective means and standard deviations?

EXAMPLE 6.5

Suppose x has a normal distribution with mean 50 and standard deviation 6.

- About 68 percent of the x values lie within one standard deviation of the mean. Therefore, about 68 percent of the x values lie between $-1\sigma = (-1)(6) = -6$ and $1\sigma = (1)(6) = 6$ of the mean 50. The values 50 – 6 = 44 and 50 + 6 = 56 are within one standard deviation from the mean 50. The z-scores are –1 and +1 for 44 and 56, respectively.
- About 95 percent of the x values lie within two standard deviations of the mean. Therefore, about 95 percent of the x values lie between $-2\sigma = (-2)(6) = -12$ and $2\sigma = (2)(6) = 12$. The values 50 – 12 = 38 and 50 + 12 = 62 are within two standard deviations from the mean 50. The z-scores are –2 and +2 for 38 and 62, respectively.
- About 99.7 percent of the x values lie within three standard deviations of the mean. Therefore, about 95 percent of the x values lie between $-3\sigma = (-3)(6) = -18$ and $3\sigma = (3)(6) = 18$ of the mean 50. The values 50 – 18 = 32 and 50 + 18 = 68 are within three standard deviations from the mean 50. The z-scores are –3 and +3 for 32 and 68, respectively.

TRY IT 6.5

Suppose X has a normal distribution with mean 25 and standard deviation five. Between what values of x do 68 percent of the values lie?

EXAMPLE 6.6

Problem

From 1984–1985, the mean height of 15-to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let Y = the height of 15-to 18-year-old males in 1984–1985. Then $Y \sim N(172.36, 6.34)$.

a. About 68 percent of the *y* values lie between what two values? These values are _____. The *z*-scores are _____, respectively.
b. About 95 percent of the *y* values lie between what two values? These values are _____. The *z*-scores are _____ respectively.
c. About 99.7 percent of the *y* values lie between what two values? These values are _____. The *z*-scores are _____, respectively.

Solution

a. About 68 percent of the values lie between 166.02 cm and 178.7 cm. The *z*-scores are –1 and 1.
b. About 95 percent of the values lie between 159.68 cm and 185.04 cm. The *z*-scores are –2 and 2.
c. About 99.7 percent of the values lie between 153.34 cm and 191.38 cm. The *z*-scores are –3 and 3.

TRY IT 6.6

The scores on a college entrance exam have an approximate normal distribution with mean, $\mu = 52$ points and a standard deviation, $\sigma = 11$ points.

a. About 68 percent of the *y* values lie between what two values? These values are _____. The *z*-scores are _____, respectively.
b. About 95 percent of the *y* values lie between what two values? These values are _____. The *z*-scores are _____, respectively.
c. About 99.7 percent of the *y* values lie between what two values? These values are _____. The *z*-scores are _____, respectively.

6.2 Using the Normal Distribution

The shaded area in the following graph indicates the area to the left of *x*. This area could represent the percentage of students scoring less than a particular grade on a final exam. This area is represented by the probability $P(X < x)$. Normal tables, computers, and calculators are used to provide or calculate the probability $P(X < x)$.

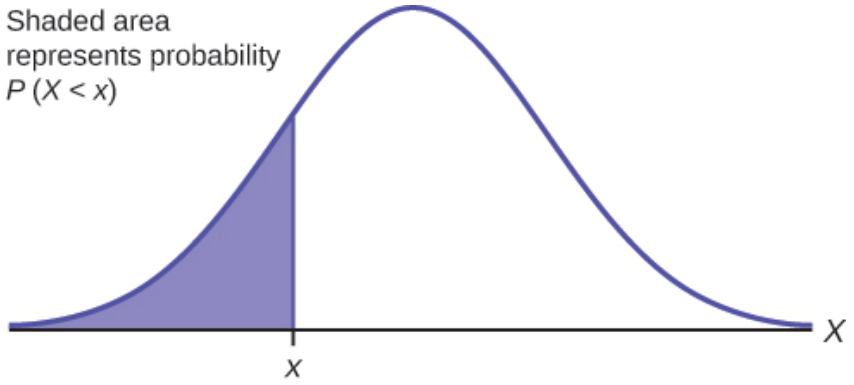


Figure 6.4

The area to the right is then $P(X > x) = 1 – P(X < x)$. Remember, $P(X < x)$ = *Area to the left* of the vertical line through *x*. $P(X < x) = 1 – P(X < x)$ = *Area to the right* of the vertical line through *x*. $P(X < x)$ is the same as $P(X \leq x)$ and $P(X > x)$ is the same as $P(X \geq x)$ for continuous distributions.

Suppose the graph above were to represent the percentage of students scoring less than 75 on a final exam, with this probability equal to 0.39. This would also indicate that the percentage of students scoring higher than 75 was equal to 1 minus 0.39 or 0.61.

Calculations of Probabilities

Probabilities are calculated using technology. There are instructions given as necessary for the TI-83+ and TI-84 calculators.

NOTE

To calculate the probability, use the probability tables provided in Appendix H Tables without the use of technology. The tables include instructions for how to use them.

The probability is represented by the area under the normal curve. To find the probability, calculate the *z*-score and look up the *z*-score in the z-table under the *z*-column. Most *z*-tables show the area under the normal curve to the left of *z*. Others show the mean to *z* area. The method used will be indicated on the table.

We will discuss the *z*-table that represents the area under the normal curve to the left of *z*. Once you have located the *z*-score, locate the corresponding area. This will be the area under the normal curve, to the left of the *z*-score. This area can be used to find the area to the right of the *z*-score, or by subtracting from 1 or the total area under the normal curve. These areas can also be used to determine the area between two *z*-scores.

EXAMPLE 6.7

If the area to the left is 0.0228, then the area to the right is 1 – 0.0228 = 0.9772.

TRY IT 6.7

If the area to the left of *x* is 0.012, then what is the area to the right?

EXAMPLE 6.8

The final exam scores in a statistics class were normally distributed with a mean of 63 and a standard deviation of five.

Problem

a. Find the probability that a randomly selected student scored more than 65 on the exam.

Solution

a. Let X = a score on the final exam. $X \sim N(63, 5)$, where $\mu = 63$ and $\sigma = 5$.

Draw a graph.

Calculate the *z*-score:

$$z = \frac{x - \mu}{\sigma} = \frac{65 - 63}{5} = \frac{2}{5} = .40 \quad \textbf{6.4}$$

The *z*-table shows that the area to the left of *z* is 0.6554. Subtracting this area from 1 gives 0.3446.

Then, find $P(x > 65)$.

$$P(x > 65) = 0.3446$$

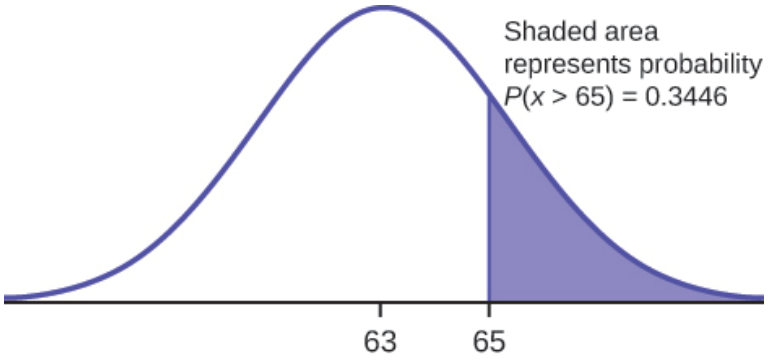


Figure 6.5

The probability that any student selected at random scores more than 65 is 0.3446.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Go into 2nd DISTR.
After pressing 2nd DISTR, press 2:normalcdf.

The syntax for the instructions is as follows:

normalcdf(lower value, upper value, mean, standard deviation) For this problem: normalcdf(65,1E99,63,5) = 0.3446. You get 1E99 (= 10^{99}) by pressing 1, the EE key—a 2nd key—and then 99. Or, you can enter 10^99 instead. The number 10^{99} is way out in the right tail of the normal curve. We are calculating the area between 65 and 10^{99}. In some instances, the lower number of the area might be –1E99 (= -10^{99}). The number -10^{99} is way out in the left tail of the normal curve. We chose the exponent of 99 because this produces such a large number that we can reasonably expect all of the values under the curve to fall below it. This is an arbitrary value and one that works well, for our purpose.

Historical Note

The TI probability program calculates a *z*-score and then the probability from the *z*-score. Before technology, the *z*-score was looked up in a standard normal probability table, also known as a Z-table—the math involved to find probability is cumbersome. In this example, a standard normal table with area to the left of the *z*-score was used. You calculate the *z*-score and look up the area to the left. The probability is the area to the right.

Problem

b. Find the probability that a randomly selected student scored less than 85.

Solution

b. Draw a graph.

Then find $P(x < 85)$, and shade the graph.

Using a computer or calculator, find $P(x < 85) = 1$.

normalcdf(0,85,63,5) = 1 (rounds to one)

The probability that one student scores less than 85 is approximately one, or 100 percent.

Problem

c. Find the 90^{th} percentile, —that is, find the score *k* that has 90 percent of the scores below *k* and 10 percent of the scores above *k*.

Solution

c. Find the 90th percentile. For each problem or part of a problem, draw a new graph. Draw the *x*-axis. Shade the area that corresponds to the 90th percentile. This time, we are looking for a score that corresponds to a given area under the curve.

Let *k* = the 90th percentile. The variable *k* is located on the *x*-axis. $P(x < k)$ is the area to the left of *k*. The 90th percentile *k* separates the exam scores into those that are the same or lower than *k* and those that are the same or higher. Ninety percent of the test scores are the same or lower than *k*, and 10 percent are the same or higher. The variable *k* is often called a **critical value**.

We know the mean, standard deviation, and area under the normal curve. We need to find the *z*-score that corresponds to the area of 0.9 and then substitute it with the mean and standard deviation, into our *z*-score formula. The *z*-table shows a *z*-score of approximately 1.28, for an area under the normal curve to the left of *z* (larger portion) of approximately 0.9. Thus, we can write the following:

$$1.28 = \frac{x - 63}{5}$$

Multiplying each side of the equation by 5 gives

$$6.4 = x - 63$$

Adding 63 to both sides of the equation gives

$$69.4 = \text{x}.$$

Thus, our score, *k*, is 69.4.

$$k = 69.4$$

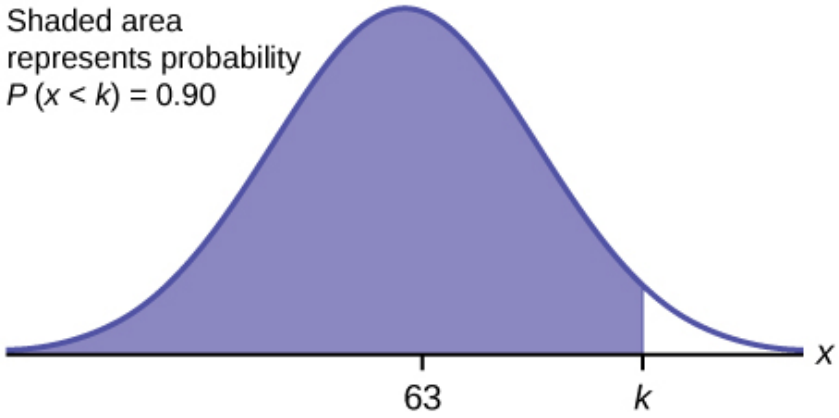


Figure 6.6

The 90th percentile is 69.4. This means that 90 percent of the test scores fall at or below 69.4 and 10 percent fall at or above. To get this answer on the calculator, follow this next step:

USING THE TI-83, 83+, 84, 84+ CALCULATOR

`invNorm` in 2nd DISTR. invNorm(area to the left, mean, standard deviation)
For this problem, invNorm(0.90,63,5) = 69.4

Problem

d. Find the 70th percentile, —that is, find the score *k* such that 70 percent of scores are below *k* and 30 percent of the scores are above *k*.

Solution

d. Find the 70th percentile.

Draw a new graph and label it appropriately. $k = 65.6$

The 70th percentile is 65.6. This means that 70 percent of the test scores fall at or below 65.5 and 30 percent fall at or above.

invNorm(0.70,63,5) = 65.6

TRY IT 6.8

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three.

Find the probability that a randomly selected golfer scored less than 65.

EXAMPLE 6.9

A personal computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking, and a myriad of other things. Suppose that the average number of hours a household personal computer is used for entertainment is two hours per day. Assume the times for entertainment are normally distributed and the standard deviation for the times is half an hour.

Problem

a. Find the probability that a household personal computer is used for entertainment between 1.8 and 2.75 hours per day.

Solution

a. Let X = the amount of time, in hours, a household personal computer is used for entertainment. $X \sim N(2, 0.5)$ where $\mu = 2$ and $\sigma = 0.5$.

Find $P(1.8 < x < 2.75)$.

First, calculate the z-scores for each x-value.

$$z = \frac{1.8-2}{0.5} = \frac{-0.2}{0.5} = -0.40$$

$$z = \frac{2.75-2}{0.5} = \frac{0.75}{0.5} = 1.5$$

Now, use the Z-table to locate the area under the normal curve to the left of each of these z-scores.

The area to the left of the z-score of −0.40 is 0.3446. The area to the left of the z-score of 1.5 is 0.9332. The area between these scores will be the difference in the two areas, or $0.9332 - 0.3446$, which equals 0.5886.

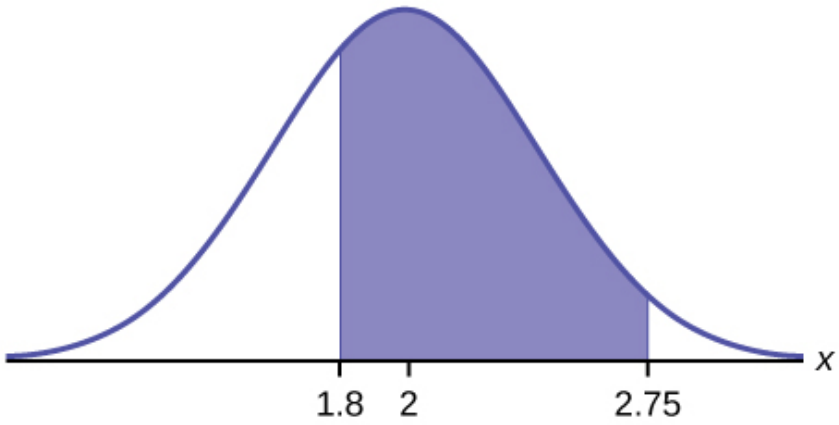


Figure 6.7

normalcdf(1.8,2.75,2,0.5) = 0.5886

The probability that a household personal computer is used between 1.8 and 2.75 hours per day for entertainment is 0.5886.

Problem

b. Find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment.

Solution

b. To find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment, **find the 25th percentile,** k, where $P(x < k) = 0.25$.

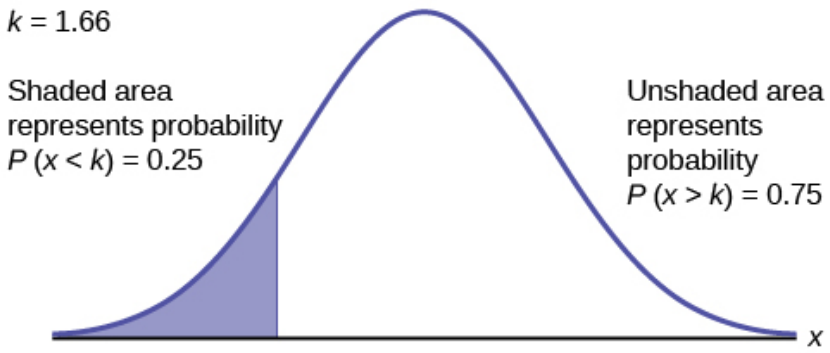


Figure 6.8

invNorm(0.25,2,0.5) = 1.66

We use invNorm because we are looking for the k-value.

The maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment is 1.66 hours.

TRY IT 6.9

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three. Find the probability that a golfer scored between 66 and 70.

EXAMPLE 6.10

In the United States smartphone users between the ages of 13 and 55+ approximately follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years, respectively.

Problem

a. Determine the probability that a random smartphone user in the age range 13 to 55+ is between 23 and 64.7 years old.

Solution

a. normalcdf(23,64.7,36.9,13.9) = 0.8186

The z-scores are calculated as

$z = \frac{23-36.9}{13.9} = \frac{-13.9}{13.9} = -1$
$z = \frac{64.7-36.9}{13.9} = \frac{27.8}{13.9} = 2$

The *Z*-table shows the area to the left of a z-score with an absolute value of 1 to be 0.1587. It shows the area to the left of a *z*-score of 2 to be 0.9772. The difference in the two areas is 0.8185.

This is slightly different than the area given by the calculator, due to rounding.

Problem

b. Determine the probability that a randomly selected smartphone user in the age range 13 to 55+ is at most 50.8 years old.

Solution

b. normalcdf(-10^{99},50.8,36.9,13.9) = 0.8413

Problem

c. Find the 80th percentile of this distribution, and interpret it in a complete sentence.

Solution

c.

invNorm(0.80,36.9,13.9) = 48.6
The 80th percentile is 48.6 years.
80 percent of the smartphone users in the age range 13–55+ are 48.6 years old or less.

TRY IT 6.10

Use the information in Example 6.10 to answer the following questions:

a. Find the 30th percentile, and interpret it in a complete sentence.
b. What is the probability that the age of a randomly selected smartphone user in the range 13 to 55+ is less than 27 years old?

EXAMPLE 6.11

In the United States the ages 13 to 55+ of smartphone users approximately follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years, respectively. Using this information, answer the following questions. —Round answers to one decimal place.

Problem

a. Calculate the interquartile range (*IQR*).

Solution

a.

IQR = $Q_3 - Q_1$
Calculate Q_3 = 75th percentile and Q_1 = 25th percentile.
Recall that we can use invNorm to find the *k*-value. We can use this to find the quartile values.
invNorm(0.75,36.9,13.9) = Q_3 = 46.2754
invNorm(0.25,36.9,13.9) = Q_1 = 27.5246
IQR = $Q_3 - Q_1$ = 18.8

Problem

b. Forty percent of the ages that range from 13 to 55+ are at least what age?

Solution

b.

Find k where $P(x \geq k) = 0.40$. *At least* translates to *greater than or equal to*.
0.40 = the area to the right
The area to the left = 1 – 0.40 = 0.60.
The area to the left of k = 0.60
invNorm(0.60,36.9,13.9) = 40.4215
k = 40.4.
Forty percent of the ages that range from 13 to 55+ are at least 40.4 years.

TRY IT 6.11

Two thousand students took an exam. The scores on the exam have an approximate normal distribution with a mean $\mu = 81$ points and standard deviation $\sigma = 15$ points.

a. Calculate the first- and third-quartile scores for this exam.
b. The middle 50 percent of the exam scores are between what two values?

EXAMPLE 6.12

A citrus farmer who grows mandarin oranges finds that the diameters of mandarin oranges harvested on his farm follow a normal distribution with a mean diameter of 5.85 cm and a standard deviation of 0.24 cm.

Problem

a. Find the probability that a randomly selected mandarin orange from this farm has a diameter larger than 6.0 cm. Sketch the graph.

Solution

a. normalcdf(6,10^99,5.85,0.24) = 0.2660

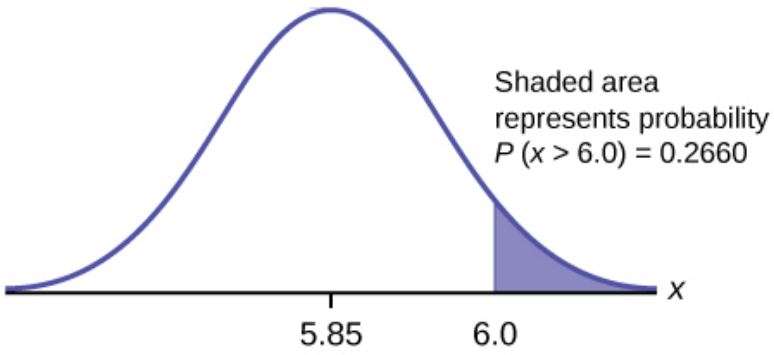


Figure 6.9

Problem

b. The middle 20 percent of mandarin oranges from this farm have diameters between _____ and _____.

Solution

b.

1 – 0.20 = 0.80. Outside of the middle 20 percent will be 80 percent of the values.

The tails of the graph of the normal distribution each have an area of 0.40.
Find k_1, the 40th percentile, and k_2, the 60th percentile (0.40 + 0.20 = 0.60). This leaves the middle 20 percent, in the middle of the distribution.
k_1 = invNorm(0.40,5.85,0.24) = 5.79 cm
k_2 = invNorm(0.60,5.85,0.24) = 5.91 cm

So, the middle 20 percent of mandarin oranges have diameters between 5.79 cm and 5.91 cm.

Problem
c. Find the 90th percentile for the diameters of mandarin oranges, and interpret it in a complete sentence.

Solution
c. 6.16, Ninety percent of the diameter of the mandarin oranges is at most 6.16 cm.

TRY IT 6.12

Using the information from Example 6.12, answer the following:

a. The middle 45 percent of mandarin oranges from this farm are between _____ and _____.
b. Find the 16th percentile, and interpret it in a complete sentence.

6.3 Normal Distribution—Lap Times

Stats Lab

Normal Distribution (Lap Times)

Student Learning Outcome

- The student will compare and contrast empirical data and a theoretical distribution to determine if Terry Vogel's lap times fit a continuous distribution.

Directions

Round the relative frequencies and probabilities to four decimal places. Carry all other decimal answers to two places.

Collect the Data

Use the data from Appendix C Data Sets. Use a stratified sampling method by lap— races 1 to 20— and a random number generator to pick six lap times from each stratum. Record the lap times below for laps two to seven.

_____	_____	_____	_____	_____	_____
_____	_____	_____	_____	_____	_____
_____	_____	_____	_____	_____	_____
_____	_____	_____	_____	_____	_____
_____	_____	_____	_____	_____	_____
_____	_____	_____	_____	_____	_____

Table 6.1

Construct a histogram. Make five to six intervals. Sketch the graph using a ruler and pencil. Scale the axes.

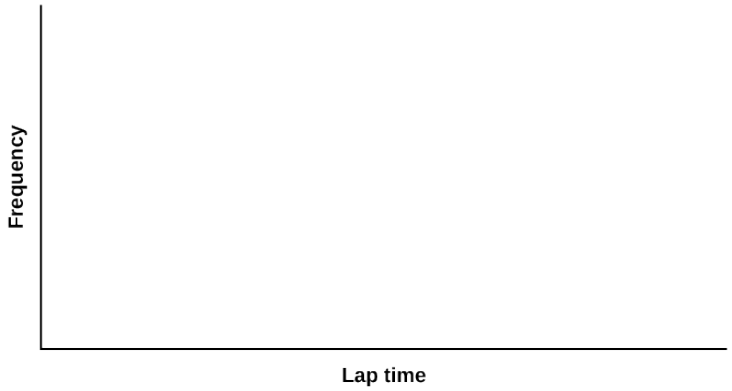


Figure 6.10

Calculate the following:

a. $\overline{x}$ = _____
b. s = _____

Draw a smooth curve through the tops of the bars of the histogram. Write one to two complete sentences to describe the general shape of the curve. (Keep it simple. Does the graph go straight across, does it have a V-shape, does it have a hump in the middle or at either end, and so on?)

Analyze the Distribution

Using your sample mean, sample standard deviation, and histogram to help, what is the approximate theoretical distribution of the data?

- X ~ _____(_____,_____)
- How does the histogram help you arrive at the approximate distribution?

Describe the Data

Use the data you collected to complete the following statements.

- The IQR goes from _____ to _____.
- IQR = _____. ($IQR = Q_3 - Q_1$)
- The 15th percentile is _____.
- The 85th percentile is _____.
- The median is _____.
- The empirical probability that a randomly chosen lap time is more than 130 seconds is _____.
- Explain the meaning of the 85th percentile of this data.

Theoretical Distribution

Using the theoretical distribution, complete the following statements. You should use a normal approximation based on your sample data.

- The IQR goes from _____ to _____.
- IQR = _____.
- The 15th percentile is _____.
- The 85th percentile is _____.
- The median is _____.
- The probability that a randomly chosen lap time is more than 130 seconds is _____.
- Explain the meaning of the 85th percentile of this distribution.

Discussion Questions

Do the data from the section titled Collect the Data give a close approximation to the theoretical distribution in the section titled Analyze the Distribution? In complete sentences and comparing the result in the sections titled Describe the Data and Theoretical Distribution, explain why or why not.

6.4 Normal Distribution—Pinkie Length

Stats Lab

Normal Distribution (Pinkie Length)

Student Learning Outcomes

- The student will compare empirical data and a theoretical distribution to determine if data from the experiment follow a continuous distribution.

Collect the Data

Measure the length of your pinkie finger, in centimeters.

1. Randomly survey 30 adults for their pinkie finger lengths. Round the lengths to the nearest 0.5 cm.

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

Table 6.2

2. Construct a histogram. Make five to six intervals. Sketch the graph using a ruler and pencil. Scale the axes.

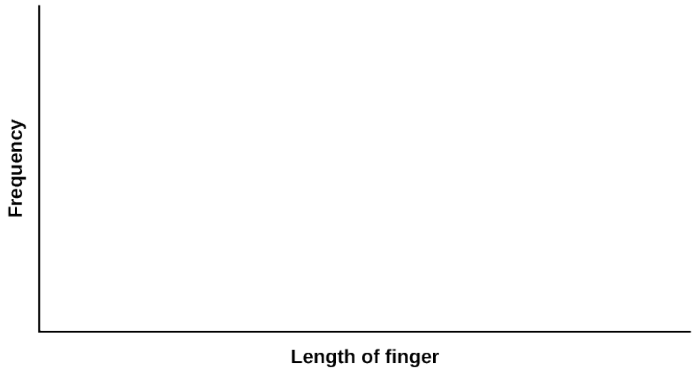


Figure 6.11

3. Calculate the following:
   a. $\overline{x}$ = _____
   b. s = _____
4. Draw a smooth curve through the top of the bars of the histogram. Write one to two complete sentences to describe the general shape of the curve. Keep it simple. Does the graph go straight across, does it have a V-shape, does it have a hump in the middle or at either end, and so on?

Analyze the Distribution

Using your sample mean, sample standard deviation, and histogram, what was the approximate theoretical distribution of the data you collected?

- X ~ _____ (_____, _____)

- How does the histogram help you arrive at the approximate distribution?

Describe the Data

Using the data you collected complete the following statements. Hint—Order the data.

Remember

($IQR = Q_3 - Q_1$)

- IQR = _____
- The 15^{th} percentile is _____.
- The 85^{th} percentile is _____.
- Median is _____.
- What is the theoretical probability that a randomly chosen pinkie length is more than 6.5 cm?
- Explain the meaning of the 85^{th} percentile of these data.

Theoretical Distribution

Using the theoretical distribution, complete the following statements. Use a normal approximation based on the sample mean and standard deviation.

- IQR = _____
- The 15^{th} percentile is _____.
- The 85^{th} percentile is _____.
- Median is _____.
- What is the theoretical probability that a randomly chosen pinkie length is more than 6.5 cm?
- Explain the meaning of the 85^{th} percentile of these data.

Discussion Questions

Do the data you collected give a close approximation to the theoretical distribution? In complete sentences and comparing the results in the sections titled Describe the Data and Theoretical Distribution, explain why or why not.

Key Terms

normal distribution a continuous random variable (RV) where μ is the mean of the distribution and σ is the standard deviation; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called the **standard normal distribution**.

standard normal distribution a continuous random variable (RV) $X \sim N(0, 1)$; when X follows the standard normal distribution, it is often noted as $Z \sim N(0, 1)$.

z-score the linear transformation of the form $z = \frac{x - \mu}{\sigma}$; if this transformation is applied to any normal distribution $X \sim N(\mu, \sigma)$, the result is the standard normal distribution $Z \sim N(0, 1)$;
If this transformation is applied to any specific value x of the RV with mean μ and standard deviation σ, the result is called the z-score of x. The z-score allows us to compare data that are normally distributed but scaled differently.

Chapter Review

6.1 The Standard Normal Distribution

A z-score is a standardized value. Its distribution is the standard normal, $Z \sim N(0, 1)$. The mean of the z-scores is zero and the standard deviation is one. If z is the z-score for a value x from the normal distribution $N(\mu, \sigma)$, then z tells you how many standard deviations x is above—greater than—or below—less than—μ.

6.2 Using the Normal Distribution

The normal distribution, which is continuous, is the most important of all the probability distributions. Its graph is bell-shaped. This bell-shaped curve is used in almost all disciplines. Since it is a continuous distribution, the total area under the curve is one. The parameters of the normal are the mean μ and the standard deviation σ. A special normal distribution, called the standard normal distribution, is the distribution of z-scores. Its mean is zero, and its standard deviation is one.

Formula Review

Introduction

$X \sim N(\mu, \sigma)$

μ = the mean, σ = the standard deviation

6.1 The Standard Normal Distribution

$Z \sim N(0, 1)$

z = a standardized value (z-score)

mean = 0, standard deviation = 1

To find the k^{th} percentile of X when the z-score is known, $k = \mu + (z)\sigma$

z-score: $z = \frac{x - \mu}{\sigma}$

Z = the random variable for z-scores

6.2 Using the Normal Distribution

Normal Distribution: $X \sim N(\mu, \sigma)$, where μ is the mean and σ is the standard deviation

Standard Normal Distribution: $Z \sim N(0, 1)$.

Calculator function for probability: normalcdf (lower x value of the area, upper x value of the area, mean, standard deviation)

Calculator function for the k^{th} percentile: k = invNorm (area to the left of k, mean, standard deviation)

Practice

6.1 The Standard Normal Distribution

1. A bottle of water contains 12.05 fluid ounces with a standard deviation of 0.01 ounces. Define the random variable X in words. X = _____.
2. A normal distribution has a mean of 61 and a standard deviation of 15. What is the median?
3. $X \sim N(1, 2)$
   σ = _____
4. A company manufactures rubber balls. The mean diameter of a ball is 12 cm with a standard deviation of 0.2 cm. Define the random variable X in words. X = _____.
5. $X \sim N(-4, 1)$
   What is the median?
6. $X \sim N(3, 5)$
   σ = _____
7. $X \sim N(-2, 1)$
   μ = _____

8. What does a *z*-score measure?

9. What does standardizing a normal distribution do to the mean?

10. Is $X \sim N(0, 1)$ a standardized normal distribution? Why or why not?

11. What is the *z*-score of $x = 12$, if it is two standard deviations to the right of the mean?

12. What is the *z*-score of $x = 9$, if it is 1.5 standard deviations to the left of the mean?

13. What is the *z*-score of $x = –2$, if it is 2.78 standard deviations to the right of the mean?

14. What is the *z*-score of $x = 7$, if it is 0.133 standard deviations to the left of the mean?

15. Suppose $X \sim N(2, 6)$. What value of *x* has a *z*-score of three?

16. Suppose $X \sim N(8, 1)$. What value of *x* has a *z*-score of –2.25?

17. Suppose $X \sim N(9, 5)$. What value of *x* has a *z*-score of –0.5?

18. Suppose $X \sim N(2, 3)$. What value of *x* has a *z*-score of –0.67?

19. Suppose $X \sim N(4, 2)$. What value of *x* is 1.5 standard deviations to the left of the mean?

20. Suppose $X \sim N(4, 2)$. What value of *x* is two standard deviations to the right of the mean?

21. Suppose $X \sim N(8, 9)$. What value of *x* is 0.67 standard deviations to the left of the mean?

22. Suppose $X \sim N(–1, 2)$. What is the *z*-score of $x = 2$?

23. Suppose $X \sim N(12, 6)$. What is the *z*-score of $x = 2$?

24. Suppose $X \sim N(9, 3)$. What is the *z*-score of $x = 9$?

25. Suppose a normal distribution has a mean of six and a standard deviation of 1.5. What is the *z*-score of $x = 5.5$?

26. In a normal distribution, $x = 5$ and $z = –1.25$. This tells you that $x = 5$ is ____ standard deviations to the ____ (right or left) of the mean.

27. In a normal distribution, $x = 3$ and $z = 0.67$. This tells you that $x = 3$ is ____ standard deviations to the ____ (right or left) of the mean.

28. In a normal distribution, $x = –2$ and $z = 6$. This tells you that $x = –2$ is ____ standard deviations to the ____ (right or left) of the mean.

29. In a normal distribution, $x = –5$ and $z = –3.14$. This tells you that $x = –5$ is ____ standard deviations to the ____ (right or left) of the mean.

30. In a normal distribution, $x = 6$ and $z = –1.7$. This tells you that $x = 6$ is ____ standard deviations to the ____ (right or left) of the mean.

31. About what percent of *x* values from a normal distribution lie within one standard deviation, left and right, of the mean of that distribution?

32. About what percent of the *x* values from a normal distribution lie within two standard deviations, left and right, of the mean of that distribution?

33. About what percent of *x* values lie between the second and third standard deviations, both sides?

34. Suppose $X \sim N(15, 3)$. Between what *x* values does 68.27 percent of the data lie? The range of *x* values is centered at the mean of the distribution (i.e., 15).

35. Suppose $X \sim N(–3, 1)$. Between what *x* values does 95.45 percent of the data lie? The range of *x* values is centered at the mean of the distribution (i.e., –3).

36. Suppose $X \sim N(–3, 1)$. Between what *x* values does 34.14 percent of the data lie?

37. About what percent of *x* values lie between the mean and three standard deviations?

38. About what percent of *x* values lie between the mean and one standard deviation?

39. About what percent of *x* values lie between the first and second standard deviations from the mean, both sides?

40. About what percent of *x* values lie between the first and third standard deviations, both sides?

Use the following information to answer the next two exercises: The life of Sunshine CD players is normally distributed with mean of 4.1 years and a standard deviation of 1.3 years. A CD player is guaranteed for three years. We are interested in the length of time a CD player lasts.

41. Define the random variable X in words. X = _____.

42. $X \sim$ _____(_____, _____)

6.2 Using the Normal Distribution

43. How would you represent the area to the left of one in a probability statement?

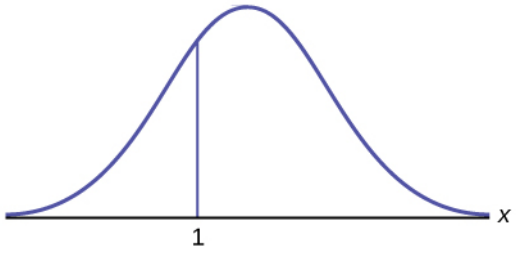


Figure 6.12

44. What is the area to the right of one?

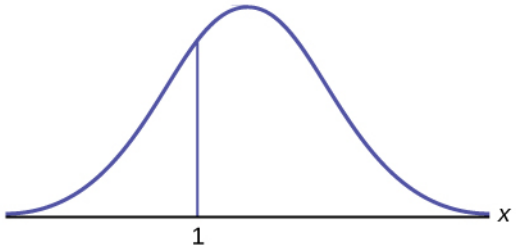


Figure 6.13

45. Is $P(x < 1)$ equal to $P(x \leq 1)$? Why or why not?

46. How would you represent the area to the left of three in a probability statement?

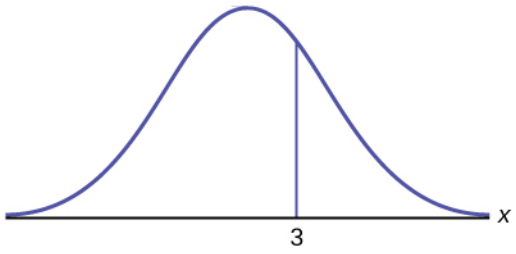


Figure 6.14

47. What is the area to the right of three?

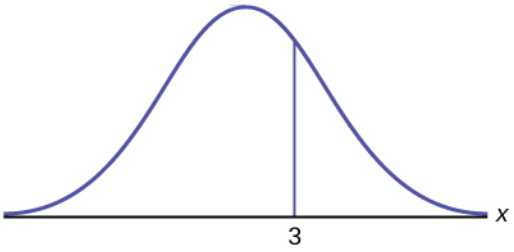


Figure 6.15

48. If the area to the left of x in a normal distribution is 0.123, what is the area to the right of x?

49. If the area to the right of x in a normal distribution is 0.543, what is the area to the left of x?

Use the following information to answer the next four exercises:

$X \sim N(54, 8)$

50. Find the probability that $x > 56$.

51. Find the probability that $x < 30$.

52. Find the 80^{th} percentile.

53. Find the 60^{th} percentile.

54. $X \sim N(6, 2)$
Find the probability that x is between three and nine.

55. $X \sim N(-3, 4)$
Find the probability that x is between one and four.

56. $X \sim N(4, 5)$
Find the maximum of x in the bottom quartile.

57. *Use the following information to answer the next three exercises:* The life of Sunshine CD players is normally distributed with a mean of 4.1 years and a standard deviation of 1.3 years. A CD player is guaranteed for three years. We are interested in the length of time a CD player lasts. Find the probability that a CD player will break down during the guarantee period.

a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the probability.

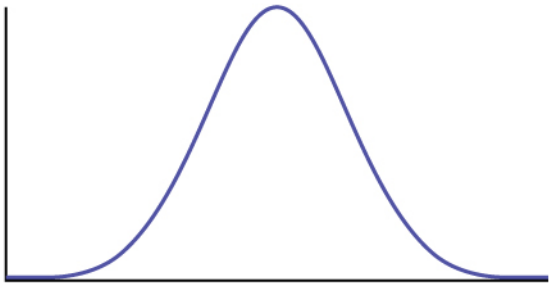


Figure 6.16

b. $P(0 < x <$ _____$) =$ _____. Use zero for the minimum value of x.

58. Find the probability that a CD player will last between 2.8 and 6 years.

a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the probability.

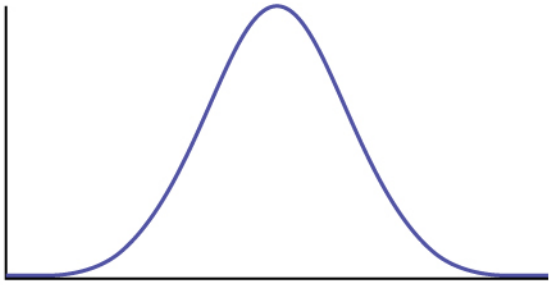


Figure 6.17

b. $P($_____ $< x <$ _____$) =$ _____

59. Find the 70th percentile of the distribution for the time a CD player lasts.

a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the lower 70 percent.

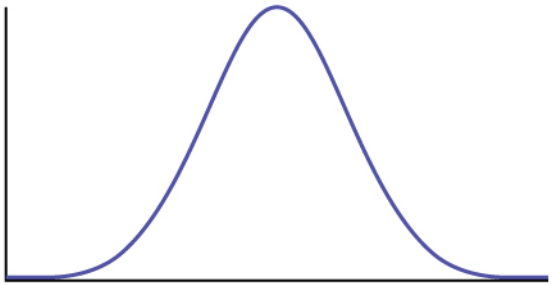


Figure 6.18

b. $P(x < k) =$ _____. Therefore, $k =$ _____.

Homework

6.1 The Standard Normal Distribution

Use the following information to answer the next two exercises: The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.

60. What is the median recovery time?

a. 2.7
b. 5.3
c. 7.4
d. 2.1

61. What is the *z*-score for a patient who takes 10 days to recover?

a. 1.5
b. 0.2
c. 2.2
d. 7.3

62. The length of time it takes to find a parking space at 9 a.m. follows a normal distribution with a mean of five minutes and a standard deviation of two minutes. If the mean is significantly greater than the standard deviation, which of the following statements is true?

I. The data cannot follow the uniform distribution.
II. The data cannot follow the exponential distribution.
III. The data cannot follow the normal distribution.

a. I only
b. II only
c. III only

d. I, II, and III

63. The heights of the 430 basketball players were listed on team rosters at the start of the 2005–2006 season. The heights of basketball players have an approximate normal distribution with a mean, $\mu = 79$ inches, and a standard deviation, $\sigma = 3.89$ inches. For each of the following heights, calculate the z-score and interpret it using complete sentences:
   a. 77 inches
   b. 85 inches
   c. If a player reported his height had a z-score of 3.5, would you believe him? Explain your answer.

64. The systolic blood pressure, given in millimeters, of males has an approximately normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 14$. Systolic blood pressure for males follows a normal distribution.
   a. Calculate the z-scores for the male systolic blood pressures 100 and 150 millimeters.
   b. If a male friend of yours said he thought his systolic blood pressure was 2.5 standard deviations below the mean, and that he believed his blood pressure was between 100 and 150 millimeters, what would you say to him?

65. Kyle's doctor told him that the z-score for his systolic blood pressure is 1.75. Which of the following is the best interpretation of this standardized score? The systolic blood pressure, given in millimeters, of males has an approximately normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 14$. If $X =$ a systolic blood pressure score, then $X \sim N(125, 14)$.
   a. Which answer(s) is/are correct?
      i. Kyle's systolic blood pressure is 175.
      ii. Kyle's systolic blood pressure is 1.75 times the average blood pressure of men his age.
      iii. Kyle's systolic blood pressure is 1.75 above the average systolic blood pressure of men his age.
      iv. Kyles's systolic blood pressure is 1.75 standard deviations above the average systolic blood pressure for men.
   b. Calculate Kyle's blood pressure.

66. Height and weight are two measurements used to track a child's development. The World Health Organization measures child development by comparing the weights of children who are the same height and same gender. In 2009, weights for all 80 cm girls in the reference population had a mean $\mu = 10.2$ kg and standard deviation $\sigma = 0.8$ kg. Weights are normally distributed. $X \sim N(10.2, 0.8)$. Calculate the z-scores that correspond to the following weights and interpret them:
   a. 11 kg
   b. 7.9 kg
   c. 12.2 kg

67. In 2005, 1,475,623 students heading to college took the SAT exam. The distribution of scores in the math section of the SAT follows a normal distribution with mean $\mu = 520$ and standard deviation $\sigma = 115$.
   a. Calculate the z-score for an SAT score of 720. Interpret it using a complete sentence.
   b. What math SAT score is 1.5 standard deviations above the mean? What can you say about this SAT score?
   c. For 2012, the SAT math test had a mean of 514 and standard deviation 117. The ACT math test is an alternative to the SAT math test, and is approximately normally distributed with mean 21 and standard deviation 5.3. If one person took the SAT math test and scored 700 and a second person took the ACT math test and scored 30, who did better with respect to the test that each person took?

6.2 Using the Normal Distribution

Use the following information to answer the next two exercises: The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.

68. What is the probability of spending more than two days in recovery?
   a. 0.0580
   b. 0.8447
   c. 0.0553
   d. 0.9420

69. The 90^{th} percentile for recovery times is –

a. 8.89
b. 7.07
c. 7.99
d. 4.32

Use the following information to answer the next three exercises: The length of time it takes to find a parking space at 9 a.m. follows a normal distribution with a mean of five minutes and a standard deviation of two minutes.

70. Based on the given information and numerically justified, would you be surprised if it took less than one minute to find a parking space?
a. Yes
b. No
c. Unable to determine

71. Find the probability that it takes at least eight minutes to find a parking space.
a. 0.0001
b. 0.9270
c. 0.1862
d. 0.0668

72. Seventy percent of the time, it takes more than how many minutes to find a parking space?
a. 1.24
b. 2.41
c. 3.95
d. 6.05

73. According to a study done by De Anza students, the height for Asian adult males is normally distributed with an average of 66 inches and a standard deviation of 2.5 inches. Suppose one Asian adult male is randomly chosen. Let X = height of the individual.
a. X ~ _____ (_____ ,_____)
b. Find the probability that the person is between 65 and 69 inches. Include a sketch of the graph, and write a probability statement.
c. Would you expect to meet many Asian adult males taller than 72 inches? Explain why or why not, and numerically justify your answer.
d. The middle 40 percent of heights fall between what two values? Sketch the graph, and write the probability statement.

74. IQ is normally distributed with a mean of 100 and a standard deviation of 15. Suppose one individual is randomly chosen. Let X = IQ of an individual.
a. X ~ _____ (_____, _____)
b. Find the probability that the person has an IQ greater than 120. Include a sketch of the graph, and write a probability statement.
c. MENSA is an organization whose members have the top 2 percent of all IQs. Find the minimum IQ needed to qualify for the MENSA organization. Sketch the graph, and write the probability statement.
d. The middle 50 percent of IQs fall between what two values? Sketch the graph, and write the probability statement.

75. The percent of fat calories that a person in the United States consumes each day is normally distributed with a mean of about 36 and a standard deviation of 10. Suppose that one individual is randomly chosen. Let X = percentage of fat calories.
a. X ~ _____ (_____, _____)
b. Find the probability that the percentage of fat calories a person consumes is more than 40. Graph the situation. Shade in the area to be determined.
c. Find the maximum number for the lower quarter of percent of fat calories. Sketch the graph and write the probability statement.

76. Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet.

a. If X = distance in feet for a fly ball, then X ~ _____ (_____, _____)
b. If one fly ball is randomly chosen from this distribution, what is the probability that this ball traveled less than 220 feet? Sketch the graph. Scale the horizontal axis X. Shade the region corresponding to the probability. Find the probability.
c. Find the 80^{th} percentile of the distribution of fly balls. Sketch the graph, and write the probability statement.

77. In China, four-year-olds average three hours a day unsupervised. Most of the unsupervised children live in rural areas, considered safe. Suppose that the standard deviation is 1.5 hours and the amount of time spent alone is normally distributed. We randomly select one Chinese four-year-old living in a rural area. We are interested in the amount of time that child spends alone per day.
a. In words, define the random variable X.
b. X ~ _____ (_____, _____)
c. Find the probability that the child spends less than one hour per day unsupervised. Sketch the graph, and write the probability statement.
d. What percentage of the children spend more than 10 hours per day unsupervised?
e. Seventy percent of the children spend at least how long per day unsupervised?

78. In the 1992 presidential election, Alaska's 40 election districts averaged 1,956.8 votes per district for a candidate. The standard deviation was 572.3. There are only 40 election districts in Alaska. The distribution of the votes per district for the candidate was bell-shaped. Let X = number of votes for the candidate for an election district.
a. State the approximate distribution of X.
b. Is 1,956.8 a population mean or a sample mean? How do you know?
c. Find the probability that a randomly selected district had fewer than 1,600 votes for the candidate. Sketch the graph, and write the probability statement.
d. Find the probability that a randomly selected district had between 1,800 and 2,000 votes for the candidate.
e. Find the third quartile for votes for the candidate.

79. Suppose that the duration of a particular type of criminal trial is known to be normally distributed with a mean of 21 days and a standard deviation of seven days.
a. In words, define the random variable X.
b. X ~ _____ (_____, _____)
c. If one of the trials is randomly chosen, find the probability that it lasted at least 24 days. Sketch the graph and write the probability statement.
d. Sixty percent of all trials of this type are completed within how many days?

80. Terri Vogel, an amateur motorcycle racer, averages 129.71 seconds per 2.5-mile lap, in a seven-lap race, with a standard deviation of 2.28 seconds. The distribution of her race times is normally distributed. We are interested in one of her randomly selected laps.
a. In words, define the random variable X.
b. X ~ _____ (_____, _____)
c. Find the percent of her laps that are completed in less than 130 seconds.
d. The fastest 3 percent of her laps are under _____.
e. The middle 80 percent of her laps are from _____ seconds to _____ seconds.

81. Thuy Dau, Ngoc Bui, Sam Su, and Lan Voung conducted a survey as to how long customers at Lucky claimed to wait in the checkout line until their turn. Let X = time in line. Table 6.3 displays the ordered real data, in minutes.

0.50	4.25	5	6	7.25
1.75	4.25	5.25	6	7.25
2	4.25	5.25	6.25	7.25
2.25	4.25	5.5	6.25	7.75

Table 6.3

2.25	4.5	5.5	6.5	8
2.5	4.75	5.5	6.5	8.25
2.75	4.75	5.75	6.5	9.5
3.25	4.75	5.75	6.75	9.5
3.75	5	6	6.75	9.75
3.75	5	6	6.75	10.75

Table 6.3

a. Calculate the sample mean and the sample standard deviation.
b. Construct a histogram.
c. Draw a smooth curve through the midpoints of the tops of the bars.
d. In words, describe the shape of your histogram and smooth curve.
e. Let the sample mean approximate μ and the sample standard deviation approximate σ. The distribution of X can then be approximated by $X \sim$ _____ (_____, _____)
f. Use the distribution in part e to calculate the probability that a person will wait fewer than 6.1 minutes.
g. Determine the cumulative relative frequency for waiting less than 6.1 minutes.
h. Why aren't the answers to part f and part g exactly the same?
i. Why are the answers to part f and part g as close as they are?
j. If only 10 customers were surveyed rather than 50, do you think the answers to part f and part g would have been closer together or farther apart? Explain your conclusion.

82. Suppose that Ricardo and Anita attend different colleges. Ricardo's GPA is the same as the average GPA at his school. Anita's GPA is 0.70 standard deviations above her school average. In complete sentences, explain why each of the following statements may be false:
a. Ricardo's actual GPA is lower than Anita's actual GPA.
b. Ricardo is not passing because his z-score is zero.
c. Anita is in the 70^{th} percentile of students at her college.

83. Table 6.4 shows a sample of the maximum capacity—maximum number of spectators—of sports stadiums. The table does not include horse-racing or motor-racing stadiums.

40,000	40,000	45,050	45,500	46,249	48,134
49,133	50,071	50,096	50,466	50,832	51,100
51,500	51,900	52,000	52,132	52,200	52,530
52,692	53,864	54,000	55,000	55,000	55,000
55,000	55,000	55,000	55,082	57,000	58,008
59,680	60,000	60,000	60,492	60,580	62,380
62,872	64,035	65,000	65,050	65,647	66,000
66,161	67,428	68,349	68,976	69,372	70,107

Table 6.4

70,585	71,594	72,000	72,922	73,379	74,500
75,025	76,212	78,000	80,000	80,000	82,300

Table 6.4

a. Calculate the sample mean and the sample standard deviation for the maximum capacity of sports stadiums.
b. Construct a histogram.
c. Draw a smooth curve through the midpoints of the tops of the bars of the histogram.
d. In words, describe the shape of your histogram and smooth curve.
e. Let the sample mean approximate μ and the sample standard deviation approximate σ. The distribution of X can then be approximated by $X \sim$ _____ (_____, _____).
f. Use the distribution in part e to calculate the probability that the maximum capacity of sports stadiums is less than 67,000 spectators.
g. Determine the cumulative relative frequency that the maximum capacity of sports stadiums is less than 67,000 spectators. Hint—Order the data and count the sports stadiums that have a maximum capacity less than 67,000. Divide by the total number of sports stadiums in the sample.
h. Why aren't the answers to part f and part g exactly the same?

84. The length of a pregnancy of a certain female animal is normally distributed with a mean of 280 days and a standard deviation of 13 days. The father was not present from 240 to 306 days before the birth of the offspring, so the pregnancy would have been less than 240 days or more than 306 days long, if he was the father. What is the probability that he was NOT the father? What is the probability that he could be the father? Calculate the z-scores first, and then use those to calculate the probability.

85. A NUMMI assembly line, which has been operating since 1984, has built an average of 6,000 cars and trucks a week. Generally, 10 percent of the cars were defective coming off the assembly line. Suppose we draw a random sample of $n = 100$ cars. Let X represent the number of defective cars in the sample. What can we say about X in regard to the 68–95–99.7 empirical rule—one standard deviation, two standard deviations, and three standard deviations from the mean being referred to? Assume a normal distribution for the defective cars in the sample.

86. We flip a coin 100 times ($n = 100$) and note that it only comes up heads 20 percent ($p = 0.20$) of the time. The mean and standard deviation for the number of times the coin lands on heads is $\mu = 20$ and $\sigma = 4$—verify the mean and standard deviation. Solve the following:
a. There is about a 68 percent chance that the number of heads will be somewhere between ___ and ___.
b. There is about a ____chance that the number of heads will be somewhere between 12 and 28.
c. There is about a ____ chance that the number of heads will be somewhere between eight and 32.

87. A child playing a carnival game will be a winner one out of five times. If 190 games are played, what is the probability that there are
a. somewhere between 34 and 54 wins
b. somewhere between 54 and 64 wins
c. more than 64 wins

88. A social media site provides a variety of statistics on its website that detail the growth and popularity of the site. On average, 28 percent of 18- to 34-year-olds check their social media profiles before getting out of bed in the morning. Suppose this percentage follows a normal distribution with a standard deviation of five percent.
a. Find the probability that the percentage of 18- to 34-year-olds who check the social media website before getting out of bed in the morning is at least 30.
b. Find the 95th percentile, and express it in a sentence.

References

6.1 The Standard Normal Distribution

CollegeBoard. (2012). 2012 *College-bound seniors: Total group profile report*. Retrieved from http://media.collegeboard.com/digitalServices/pdf/research/TotalGroup-2012.pdf

Joyce, C. A., Janssen, S., & Liu, M. L. (2010). *The world almanac and book of facts, 2010*. New York, NY: World Almanac Books.

London School of Hygiene and Tropical Medicine. (2009). Calculation of z-scores. Retrieved from http://conflict.lshtm.ac.uk/page_125.htm

National Center for Education Statistics. (2009). *ACT score averages and standard deviations, by sex and race/ethnicity, and percentage of ACT test takers, by selected composite score ranges and planned fields of study: Selected years, 1995 through 2009*. Retrieved from http://nces.ed.gov/programs/digest/d09/tables/dt09_147.asp

NBA.com. (2013). NBA Media Ventures. Retrieved from http://www.nba.com

StatCrunch. (2010). *Blood pressure of males and females*. Retrieved from http://www.statcrunch.com/5.0/viewreport.php?reportid=11960

The Mercury News. (n.d.). Retrieved from http://www.mercurynews.com/

Wikipedia. (2013). List of stadiums by capacity - Wikipedia. Retrieved from https://en.wikipedia.org/wiki/List_of_stadiums_by_capacity

6.2 Using the Normal Distribution

Chicago Public Media & Ira Glass. (2013). 403: *NUMMI*. Retrieved from http://www.thisamericanlife.org/radioarchives/episode/403/nummi

lauramitchell347. (2012, Dec. 28). Smart phone users, by the numbers. *Visually*. Retrieved from http://visual.ly/smart-phone-users-numbers

Statistics Brain Research Institute. (2013). *Facebook company statistics – statistic brain*. Retrieved from http://www.statisticbrain.com/facebook-statistics/

Wikipedia (2013). *Naegele's rule*. Retrieved from http://en.wikipedia.org/wiki/Naegele's_rule

Win at the Lottery. (2013). *Scratch-off lottery ticket playing tips*. Retrieved from www.winatthelottery.com/public/department40.cfm

Solutions

1. ounces of water in a bottle

3. 2

5. –4

7. –2

9. The mean becomes zero.

11. $z = 2$

13. $z = 2.78$

15. $x = 20$

17. $x = 6.5$

19. $x = 1$

21. $x = 1.97$

23. $z = -1.67$

25. $z \approx -0.33$

27. 0.67, right

29. 3.14, left

31. about 68 percent

33. about 4 percent

35. between –5 and –1

37. about 50 percent

39. about 27 percent

41. The lifetime of a Sunshine CD player measured in years

43. $P(x < 1)$

45. Yes, because they are the same in a continuous distribution: $P(x = 1) = 0$

47. $1 - P(x < 3)$ or $P(x > 3)$

49. 1 – 0.543 = 0.457

51. 0.0013

53. 56.03

55. 0.1186

57.
a. Check student's solution
b. 3, 0.1979

59.
a. Check student's solution
b. 0.70, 4.78 years

61. c

63.
a. Use the *z*-score formula. $z = -0.5141$. The height of 77 inches is 0.5141 standard deviations below the mean. An NBA player whose height is 77 inches is shorter than average.
b. Use the *z*-score formula. $z = 1.5424$. The height 85 inches is 1.5424 standard deviations above the mean. An NBA player whose height is 85 inches is taller than average.
c. Height = 79 + 3.5(3.89) = 90.67 inches, which is over 7.7 feet tall. There are very few NBA players this tall; so, the answer is no, not likely.

65.
a. iv
b. Kyle's blood pressure is equal to 125 + (1.75)(14) = 149.5.

67. Let X = an SAT math score and Y = an ACT math score.
a. $X = 720$ $\frac{720 - 520}{15} = 1.74$ The exam score of 720 is 1.74 standard deviations above the mean of 520.
b. $z = 1.5$
The math SAT score is 520 + 1.5(115) ≈ 692.5. The exam score of 692.5 is 1.5 standard deviations above the mean of 520.
c. $\frac{X - \mu}{\sigma} = \frac{700 - 514}{117} \approx 1.59$, the *z*-score for the SAT. $\frac{Y - \mu}{\sigma} = \frac{30 - 21}{5.3} \approx 1.70$, the *z*-scores for the ACT. With respect to the test they took, the person who took the ACT did better—has the higher *z*-score).

69. c

71. d

73. a. $X \sim N(66, 2.5)$
b. 0.5404
c. No, the probability that an Asian male is over 72 inches tall is 0.0082.

75. a. $X \sim N(36, 10)$
b. The probability that a person consumes more than 40 percent of their calories as fat is 0.3446.
c. Approximately 25 percent of people consume less than 29.26 percent of their calories as fat.

77. a. X = number of hours that a Chinese four-year-old in a rural area is unsupervised during the day.
b. $X \sim N(3, 1.5)$
c. The probability that the child spends less than one hour a day unsupervised is 0.0918.
d. The probability that a child spends over 10 hours a day unsupervised is less than 0.0001.
e. 2.21 hours

79. a. X = the distribution of the number of days a particular type of criminal trial will take
b. $X \sim N(21, 7)$
c. The probability that a randomly selected trial will last more than 24 days is 0.3336.
d. 22.77

81. a. mean = 5.51, $s = 2.15$
b. Check student's solution.
c. Check student's solution.
d. Check student's solution.
e. $X \sim N(5.51, 2.15)$
f. 0.6029
g. The cumulative frequency for less than 6.1 minutes is 0.64.
h. The answers to part f and part g are not exactly the same, because the normal distribution is only an approximation to the real one.
i. The answers to part f and part g are close, because a normal distribution is an excellent approximation when the sample size is greater than 30.
j. The approximation would have been less accurate, because the smaller sample size means that the data does not fit a normal curve as well.

83. 1. mean = 60,136
$s = 10{,}468$
2. Answers will vary
3. Answers will vary
4. Answers will vary
5. $X \sim N(60136, 10468)$
6. 0.7440
7. The cumulative relative frequency is 43/60 = 0.717.
8. The answers for part f and part g are not the same because the normal distribution is only an approximation.

85. $n = 100$; $p = 0.1$; $q = 0.9$
$\mu = np = (100)(0.10) = 10$
$\sigma = \sqrt{npq} = \sqrt{(100)(0.1)(0.9)} = 3$

i. $z = \pm 1$: $x_1 = \mu + z\sigma = 10 + 1(3) = 13$ and $x2 = \mu - z\sigma = 10 - 1(3) = 7$. 68 percent of the defective cars will fall between seven and 13
ii. $z = \pm 2$: $x_1 = \mu + z\sigma = 10 + 2(3) = 16$ and $x2 = \mu - z\sigma = 10 - 2(3) = 4$. 95 percent of the defective cars will fall between four and 16
iii. $z = \pm 3$: $x_1 = \mu + z\sigma = 10 + 3(3) = 19$ and $x2 = \mu - z\sigma = 10 - 3(3) = 1$. 99.7 percent of the defective cars will fall between one and 19

87. $n = 190$; $p = \frac{1}{5} = 0.2$; $q = 0.8$

$\mu = np = (190)(0.2) = 38$

$\sigma = \sqrt{npq} = \sqrt{(190)(0.2)(0.8)} = 5.5136$

a. For this problem: $P(34 < x < 54)$ = normalcdf(34,54,48,5.5136) = 0.7641
b. For this problem: $P(54 < x < 64)$ = normalcdf(54,64,48,5.5136) = 0.0018
c. For this problem: $P(x > 64)$ = normalcdf(64,10^{99},48,5.5136) = 0.0000012 (approximately 0)

7 THE CENTRAL LIMIT THEOREM

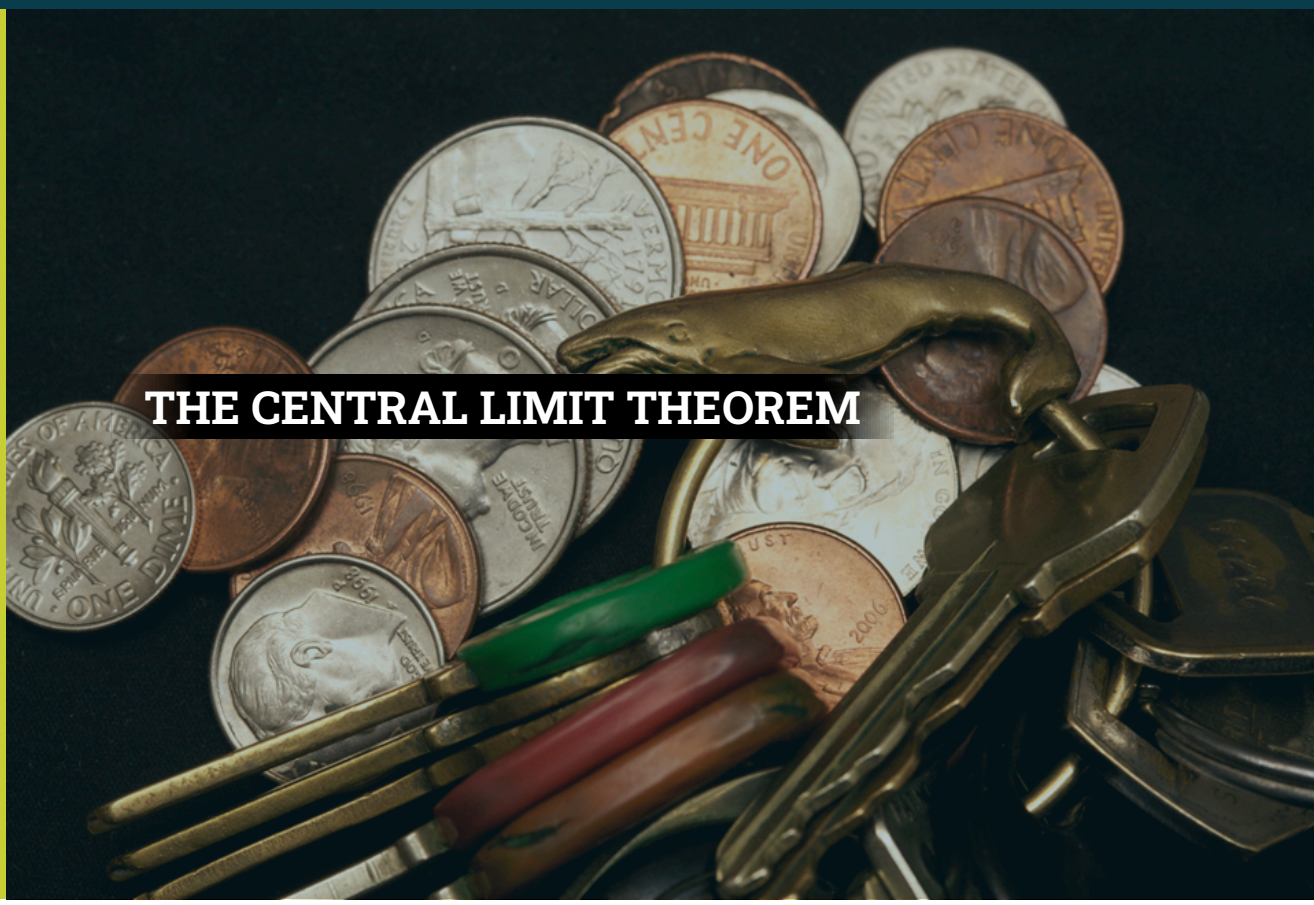


Figure 7.1 If you want to figure out the distribution of the change people carry in their pockets, using the central limit theorem and assuming your sample is large enough, you will find that the distribution is normal and bell-shaped. (credit: John Lodder)

Chapter Objectives

By the end of this chapter, the student should be able to do the following:

- Recognize central limit theorem problems
- Classify continuous word problems by their distributions
- Apply and interpret the central limit theorem for means
- Apply and interpret the central limit theorem for sums

Introduction

Why are we so concerned with means? Two reasons are they give us a middle ground for comparison, and they are easy to calculate. In this chapter, you will study means and the **central limit theorem**.

The **central limit theorem** (clt) is one of the most powerful and useful ideas in all of statistics. There are two alternative forms of the theorem, and both alternatives are concerned with drawing a finite samples size n from a population with a known mean, μ, and a known standard deviation, σ. The first alternative says that if we collect samples of size n with a *large enough* n, calculate each sample's mean, and create a histogram of those means, then the resulting histogram will tend to have an approximate normal bell shape. The second alternative says that if we again collect samples of size n that are *large enough*, calculate the sum of each sample and create a histogram, then the resulting histogram will again tend to have a normal bell shape. The central limit theorem for sample means is more discussed in the world of statistics, but it is important to note that taking each sample's sum and graphing the sums will also result in a normal histogram. There are instances where one wishes to calculate the sum of a sample, as opposed to its mean.

In either case, it does not matter what the distribution of the original population is, or whether you even need to know it. The important fact is that the distributions of sample means and the sums tend to follow the normal distribution.

The size of the sample, *n*, that is required in order to be *large enough* depends on the original population from which the samples are drawn (the sample size should be at least 30 or the data should come from a normal distribution). If the original population is far from normal, then more observations are needed for the sample means or sums to be normal. **Sampling is done with replacement**.

COLLABORATIVE EXERCISE

Suppose eight of you roll one fair die ten times, seven of you roll two fair dice ten times, nine of you roll five fair dice ten times, and 11 of you roll ten fair dice ten times.

Each time a person rolls more than one die, he or she calculates the sample **mean** of the faces showing. For example, one person might roll five fair dice and get 2, 2, 3, 4, and 6 on one roll.

The mean is $\frac{2+2+3+4+6}{5}$ = 3.4. The 3.4 is one mean when five fair dice are rolled. This same person would roll the five dice nine more times and calculate nine more means for a total of ten means.

Your instructor will pass out the dice to several people. Roll your dice ten times. For each roll, record the faces, and find the mean. Round to the nearest 0.5.

Your instructor (and possibly you) will produce one graph (it might be a histogram) for one die, one graph for two dice, one graph for five dice, and one graph for ten dice. Because the *mean* when you roll one die is just the face on the die, what distribution do these **means** appear to be representing?

Draw the graph for the means using two dice. Do the sample means show any kind of pattern?

Draw the graph for the means using five dice. Do you see any pattern emerging?

Finally, draw the graph for the means using ten dice. Do you see any pattern to the graph? What can you conclude as you increase the number of dice?

As the number of dice rolled increases from one to two to five to ten, the following is happening:

1. The mean of the sample means remains approximately the same.
2. The spread of the sample means (the standard deviation of the sample means) gets smaller.
3. The graph appears steeper and thinner.

You have just demonstrated the central limit theorem (clt).

The central limit theorem tells you that as you increase the number of dice, **the sample means tend toward a normal distribution (the sampling distribution)**.

7.1 The Central Limit Theorem for Sample Means (Averages)

Suppose X is a random variable with a distribution that may be known or unknown (it can be any distribution). Using a subscript that matches the random variable, suppose

a. μ_X = the mean of X
b. σ_X = the standard deviation of X

If you draw random samples of size n, then as n increases, the random variable $\overline{X}$, which consists of sample means, tends to be **normally distributed** and

$$\overline{X} \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right) \quad 7.1$$

The **central limit theorem** for sample means says that if you keep drawing larger and larger samples (such as rolling one, two, five, and finally, ten dice) and **calculating their means**, the sample means form their own **normal distribution** (the sampling distribution). The normal distribution has the same mean as the original distribution and a variance that equals the original variance divided by the sample size. The variable n is the number of values that are averaged together, not the number of times the experiment is done.

To put it more formally, if you draw random samples of size n, the distribution of the random variable $\overline{X}$, which consists of sample means, is called the **sampling distribution of the mean**. The sampling distribution of the mean approaches a

normal distribution as n, the **sample size**, increases.

The random variable $\overline{X}$ has a different z-score associated with it from that of the random variable X. The mean $\overline{x}$ is the value of $\overline{X}$ in one sample.

$$z = \frac{\overline{x} - \mu_X}{\left(\frac{\sigma_X}{\sqrt{n}}\right)},$$

μ_X is the average of both X and $\overline{X}$.

$\sigma\overline{x} = \frac{\sigma X}{\sqrt{n}}$ = standard deviation of $\overline{X}$ and is called the **standard error of the mean**.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

To find probabilities for means on the calculator, follow these steps.

2nd DISTR
2:normalcdf

$$normalcdf\left(lower\ value\ of\ the\ area,\ upper\ value\ of\ the\ area,\ mean,\ \frac{standard\ deviation}{\sqrt{sample\ size}}\right)$$

where

- *mean* is the mean of the original distribution
- *standard deviation* is the standard deviation of the original distribution
- *sample size* = n

EXAMPLE 7.1

A distribution has a mean of 90 and a standard deviation of 15. Samples of size $n = 25$ are drawn randomly from the population.

Problem

a. Find the probability that the **sample mean** is between 85 and 92.

Solution

a. Let X = one value from the original unknown population. The probability question asks you to find a probability for the **sample mean**.

Let $\overline{X}$ = the mean of a sample of size 25. Because $\mu_X = 90$, $\sigma_X = 15$, and $n = 25$,

$$\overline{X} \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$$

Find $P(85 < \overline{x} < 92)$. Draw a graph.

$P(85 < \overline{x} < 92) = 0.6997$

The probability that the sample mean is between 85 and 92 is 0.6997.

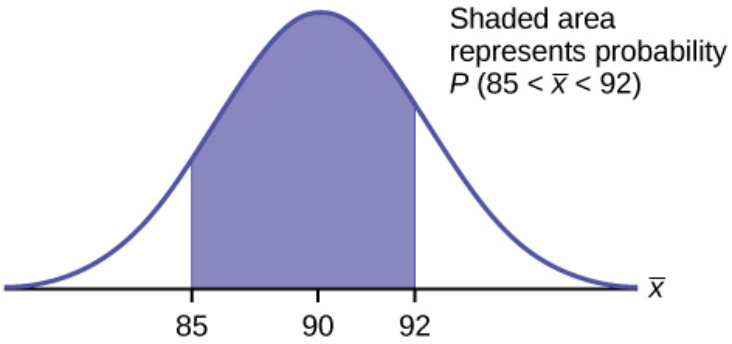


Figure 7.2

Find $P(85 < \bar{x} < 92)$. Draw a graph.

$$P(85 < \bar{x} < 92) = 0.6997 \quad \textbf{7.2}$$

USING THE TI-83, 83+, 84, 84+ CALCULATOR

`normalcdf`(lower value, upper value, mean, standard error of the mean)

The parameter list is abbreviated (lower value, upper value, μ, $\frac{\sigma}{\sqrt{n}}$).

`normalcdf`$(85,92,90,\frac{15}{\sqrt{25}}) = 0.6997$

Problem

b. Find the value that is two standard deviations above the expected value, 90, of the sample mean.

Solution

b. To find the value that is two standard deviations above the expected value 90, use the following formula

$$\text{value} = \mu_x + (\#\text{ofSTDEVs})\left(\frac{\sigma_x}{\sqrt{n}}\right)$$

$$\text{value} = 90 + 2\left(\frac{15}{\sqrt{25}}\right) = 96.$$

The value that is two standard deviations above the expected value is 96.

The standard error of the mean is $\frac{\sigma x}{\sqrt{n}} = \frac{15}{\sqrt{25}} = 3$. Recall that the standard error of the mean is a description of how far (on average) that the sample mean will be from the population mean in repeated simple random samples of size n.

TRY IT 7.1

An unknown distribution has a mean of 45 and a standard deviation of eight. Samples of size $n = 30$ are drawn randomly from the population. Find the probability that the sample mean is between 42 and 50.

EXAMPLE 7.2

Problem

The length of time, in hours, it takes a group of people, 40 years old and older, to play one soccer match is normally distributed with a **mean of 2 hours** and a **standard deviation of 0.5 hours**. A **sample of size $n = 50$** is drawn randomly

from the population. Find the probability that the **sample mean** is between 1.8 hours and 2.3 hours.

Solution

Let X = the time, in hours, it takes to play one soccer match.

The probability question asks you to find a probability for the **sample mean time,** *in hours*, it takes to play one soccer match.

Let $\overline{X}$ = the **mean** time, in hours, it takes to play one soccer match.

If μ_X = _____, σ_X = _____, and n = _____, then $X \sim N$(_____, _____) by the **central limit theorem for means**.

$\mu_X = 2$, $\sigma_X = 0.5$, $n = 50$, and $X \sim N\left(2, \frac{0.5}{\sqrt{50}}\right)$

Find $P(1.8 < \overline{x} < 2.3)$. Draw a graph.

$$P(1.8 < \overline{x} < 2.3) = 0.9977$$

`normalcdf`

$$\left(1.8, 2.3, 2, \frac{.5}{\sqrt{50}}\right) = 0.9977$$

The probability that the mean time is between 1.8 hours and 2.3 hours is 0.9977.

TRY IT 7.2

The length of time taken on the SAT exam for a group of students is normally distributed with a mean of 2.5 hours and a standard deviation of 0.25 hours. A sample size of $n = 60$ is drawn randomly from the population. Find the probability that the sample mean is between two hours and three hours.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

To find percentiles for means on the calculator, follow these steps.

2nd DIStR
3:invNorm

$$k = \text{invNorm}\left(\text{area to the left of } k,\ \text{mean},\ \frac{\textit{standard deviation}}{\sqrt{\textit{sample size}}}\right)$$

where

- k = the k^{th} percentile
- *mean* is the mean of the original distribution
- *standard deviation* is the standard deviation of the original distribution
- *sample size* = n

EXAMPLE 7.3

Problem

In a recent study reported Oct. 29, 2012, the mean age of tablet users is 34 years. Suppose the standard deviation is 15 years. Take a sample of size $n = 100$.

a. What are the mean and standard deviation for the sample mean ages of tablet users?

b. What does the distribution look like?
c. Find the probability that the sample mean age is more than 30 years (the reported mean age of tablet users in this particular study).
d. Find the 95th percentile for the sample mean age (to one decimal place).

Solution

a. Because the sample mean tends to target the population mean, we have $\mu_{\bar{X}} = \mu = 34$. The sample standard deviation is given by $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{100}} = \frac{15}{10} = 1.5$.
b. The central limit theorem states that for large sample sizes (n), the sampling distribution will be approximately normal.
c. The probability that the sample mean age is more than 30 is given by $P(X > 30) = \texttt{normalcdf}(30,E99,34,1.5) = 0.9962$.
d. Let k = the 95th percentile.
$k = \text{invNorm}\left(0.95,34,\frac{15}{\sqrt{100}}\right) = 36.5$

TRY IT 7.3

A gaming marketing gap for men between the ages of 30 to 40 has been identified. You are researching a startup game targeted at the 35-year-old demographic. Your idea is to develop a strategy game that can be played by men from their late 20s through their late 30s. Based on the article's data, industry research shows that the average strategy player is 28 years old with a standard deviation of 4.8 years. You take a sample of 100 randomly selected gamers. If your target market is 29- to 35-year-olds, should you continue with your development strategy?

EXAMPLE 7.4

Problem

The mean number of minutes for app engagement by a tablet user is 8.2 minutes. Suppose the standard deviation is one minute. Take a sample of 60.

a. What are the mean and standard deviation for the sample mean number of app engagement minutes by a tablet user?
b. What is the standard error of the mean?
c. Find the 90th percentile for the sample mean time for app engagement for a tablet user. Interpret this value in a complete sentence.
d. Find the probability that the sample mean is between eight minutes and 8.5 minutes.

Solution

a. $\mu_{\bar{x}} = \mu = 8.2\ \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{60}} = 0.13$
b. This allows us to calculate the probability of sample means of a particular distance from the mean, in repeated samples of size 60.
c. Let k = the 90th percentile.
$k = \texttt{invNorm}\left(0.90,8.2,\frac{1}{\sqrt{60}}\right) = 8.37$. This values indicates that 90 percent of the average app engagement time for table users is less than 8.37 minutes.
d. $P(8 < \bar{x} < 8.5) = \texttt{normalcdf}\left(8,8.5,8.2,\frac{1}{\sqrt{60}}\right) = 0.9293$

TRY IT 7.4

Cans of a cola beverage claim to contain 16 ounces. The amounts in a sample are measured and the statistics are n =

34, $\bar{x}$ = 16.01 ounces. If the cans are filled so that μ = 16.00 ounces (as labeled) and σ = 0.143 ounces, find the probability that a sample of 34 cans will have an average amount greater than 16.01 ounces. Do the results suggest that cans are filled with an amount greater than 16 ounces?

7.2 The Central Limit Theorem for Sums (Optional)

Suppose X is a random variable with a distribution that may be **known or unknown** (it can be any distribution) and suppose:

a. μ_X = the mean of X
b. σ_X = the standard deviation of X

If you draw random samples of size n, then as n increases, the random variable ΣX consisting of sums tends to be **normally distributed** and $\Sigma X \sim N[(n)(\mu_X), (\sqrt{n})(\sigma_X)]$.

The central limit theorem for sums says that if you keep drawing larger and larger samples and taking their sums, the sums form their own normal distribution (the sampling distribution), which approaches a normal distribution as the sample size increases. *The normal distribution has a mean equal to the original mean multiplied by the sample size and a standard deviation equal to the original standard deviation multiplied by the square root of the sample size.*

The random variable ΣX has the following z-score associated with it:

a. Σx is one sum.
b. $z = \frac{\Sigma x - (n)(\mu_X)}{(\sqrt{n})(\sigma_X)}$
   i. $(n)(\mu_X)$ = mean of ΣX
   ii. $(\sqrt{n})(\sigma_X)$ = standard deviation of ΣX

USING THE TI-83, 83+, 84, 84+ CALCULATOR

To find probabilities for sums on the calculator, follow these steps:

2nd DISTR
2:normalcdf
normalcdf(lower value of the area, upper value of the area, (n)(mean), $(\sqrt{n})$(standard deviation))

where,

- *mean* is the mean of the original distribution,
- *standard deviation* is the standard deviation of the original distribution, and
- *sample size* = n.

EXAMPLE 7.5

An unknown distribution has a mean of 90 and a standard deviation of 15. A sample of size 80 is drawn randomly from the population.

Problem

a. Find the probability that the sum of the 80 values (or the total of the 80 values) is more than 7,500.
b. Find the sum that is 1.5 standard deviations above the mean of the sums.

Solution

Let X = one value from the original unknown population. The probability question asks you to find a probability for **the sum (or total of) 80 values**.

ΣX = the sum or total of 80 values. Because μ_X = 90, σ_X = 15, and n = 80, $\Sigma X \sim N[(80)(90), (\sqrt{80})(15)]$

- mean of the sums = $(n)(\mu_X)$ = (80)(90) = 7200

- standard deviation of the sums = $(\sqrt{n})(\sigma_X) = (\sqrt{80})(15)$
- sum of 80 values = $\Sigma x = 7500$

a. Find $P(\Sigma x > 7500)$

$P(\Sigma x > 7500) = 0.0127$

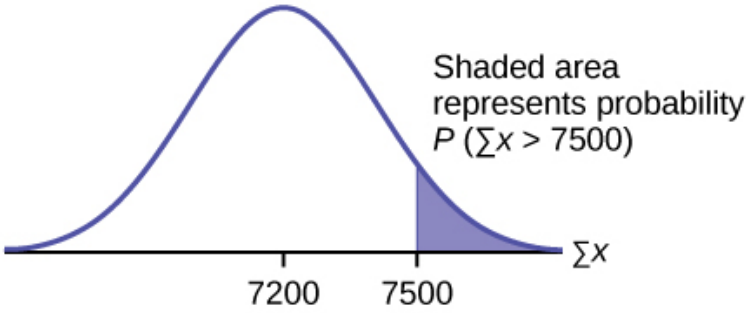


Figure 7.3

USING THE TI-83, 83+, 84, 84+ CALCULATOR

`normalcdf`(lower value, upper value, mean of sums, `stdev` of sums)

The parameter list is abbreviated(lower, upper, $(n)(\mu_X)$, $(\sqrt{n})(\sigma_X)$)

`normalcdf` (7500,1E99,(80)(90),$(\sqrt{80})$(15)) = 0.0127

Reminder

1E99 = 10^{99}.

Press the EE key for E.

b. Find Σx where $z = 1.5$.

$\Sigma x = (n)(\mu_X) + (z)(\sqrt{n})(\sigma_X) = (80)(90) + (1.5)(\sqrt{80})(15) = 7401.2$

TRY IT 7.5

An unknown distribution has a mean of 45 and a standard deviation of 8. A sample size of 50 is drawn randomly from the population. Find the probability that the sum of the 50 values is more than 2,400.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

To find percentiles for sums on the calculator, follow these steps:

2^{nd} DIStR
3:invNorm
k = invNorm (area to the left of k, (n)(mean), $(\sqrt{n})$(standard deviation))

where,

- k is the k^{th} percentile,
- *mean* is the mean of the original distribution,
- *standard deviation* is the standard deviation of the original distribution, and
- *sample size* = n.

EXAMPLE 7.6

Problem

In a recent study reported Oct. 29, 2012, the mean age of tablet users is 34 years. Suppose the standard deviation is 15 years. The sample size is 50.

a. What are the mean and standard deviation for the sum of the ages of tablet users? What is the distribution?
b. Find the probability that the sum of the ages is between 1,500 and 1,800 years.
c. Find the 80th percentile for the sum of the 50 ages.

Solution

a. $\mu_{\Sigma x} = n\mu_x = 50(34) = 1{,}700$ and $\sigma_{\Sigma x} = \sqrt{n}\sigma_x = (\sqrt{50})(15) = 106.01$
The distribution is normal for sums by the central limit theorem.
b. $P(1500 < \Sigma x < 1800)$ = `normalcdf` (1500, 1800, (50)(34), $(\sqrt{50})(15)$) = 0.7974
c. Let k = the 80th percentile.
k = `invNorm`(0.80,(50)(34),$(\sqrt{50})(15)$) = 1789.3

TRY IT 7.6

In a recent study reported Oct.29, 2012, the mean age of tablet users is 35 years. Suppose the standard deviation is 10 years. The sample size is 39.

a. What are the mean and standard deviation for the sum of the ages of tablet users? What is the distribution?
b. Find the probability that the sum of the ages is between 1,400 and 1,500 years.
c. Find the 90th percentile for the sum of the 39 ages.

EXAMPLE 7.7

Problem

The mean number of minutes for app engagement by a tablet user is 8.2 minutes. Suppose the standard deviation is one minute. Take a sample size of 70.

a. What are the mean and standard deviation for the sums?
b. Find the 95th percentile for the sum of the sample. Interpret this value in a complete sentence.
c. Find the probability that the sum of the sample is at least 10 hours.

Solution

a. $\mu_{\Sigma x} = n\mu_x = 70(8.2) = 574$ minutes and $\sigma_{\Sigma x} = \left(\sqrt{n}\right)(\sigma_x) = (\sqrt{70})(1) = 8.37$ minutes
b. Let k = the 95th percentile.
k = invNorm (0.95,(70)(8.2),$(\sqrt{70})(1)$) = 587.76 minutes
Ninety-five percent of the app engagement times are at most 587.76 minutes.
c. 10 hours = 600 minutes
$P(\Sigma x \geq 600)$ = `normalcdf`(600,E99,(70)(8.2),$(\sqrt{70})(1)$) = 0.0009

TRY IT 7.7

The mean number of minutes for app engagement by a tablet user is 8.2 minutes. Suppose the standard deviation is one minute. Take a sample size of 70.

a. What is the probability that the sum of the sample is between seven hours and 10 hours? What does this mean in context of the problem?
b. Find the 84th and 16th percentiles for the sum of the sample. Interpret these values in context.

7.3 Using the Central Limit Theorem

It is important for you to understand when to use the **central limit theorem**. If you are being asked to find the probability of the mean, use the clt for the means. If you are being asked to find the probability of a sum or total, use the clt for sums. This also applies to percentiles for means and sums.

NOTE

If you are being asked to find the probability of an **individual** value, do **not** use the clt. *Use the distribution of its random variable.*

Examples of the Central Limit Theorem

Law of Large Numbers

The **law of large numbers** says that if you take samples of larger and larger sizes from any population, then the mean $\overline{x}$ of the samples tends to get closer and closer to μ. From the central limit theorem, we know that as n gets larger and larger, the sample means follow a normal distribution. The larger n gets, the smaller the standard deviation gets. (Remember that the standard deviation for $\overline{X}$ is $\frac{\sigma}{\sqrt{n}}$.) This means that the sample mean $\overline{x}$ must be close to the population mean μ. We can say that μ is the value that the sample means approach as n gets larger. The central limit theorem illustrates the law of large numbers.

Central Limit Theorem for the Mean and Sum Examples

EXAMPLE 7.8

A study involving stress is conducted among the students on a college campus. *The stress scores follow a* ***uniform distribution*** with the lowest stress score equal to one and the highest equal to five. Using a sample of 75 students, find:

a. the probability that the *mean stress score* for the 75 students is less than 2
b. the 90th percentile for the *mean stress score* for the 75 students
c. the probability that the *total of the 75 stress scores* is less than 200
d. the 90th percentile for the *total stress score* for the 75 students

Let X = one stress score.

Problems (a) and (b) ask you to find a probability or a percentile for a **mean**. Problems (c) and (d) ask you to find a probability or a percentile for a **total or sum**. The sample size, n, is equal to 75.

Because the individual stress scores follow a uniform distribution, $X \sim U(1, 5)$ where $a = 1$ and $b = 5$ (see Continuous Random Variables for an explanation of a uniform distribution),

$$\mu_X = \frac{a+b}{2} = \frac{1+5}{2} = 3 \quad \textbf{7.3}$$

$$\sigma_X = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(5-1)^2}{12}} = 1.15 \quad \textbf{7.4}$$

In the formula above, the denominator is understood to be 12, regardless of the endpoints of the uniform distribution.

For problems (a) and (b), let $\overline{X}$ = the mean stress score for the 75 students. Then,

$$\overline{X} \sim N\left(3, \frac{1.15}{\sqrt{75}}\right) \text{ where } n = 75. \quad \textbf{7.5}$$

Problem

a. Find $P(\overline{x} < 2)$. Draw the graph.

Solution

a. $P(\overline{x} < 2) = 0$

The probability that the mean stress score is less than 2 is about zero.

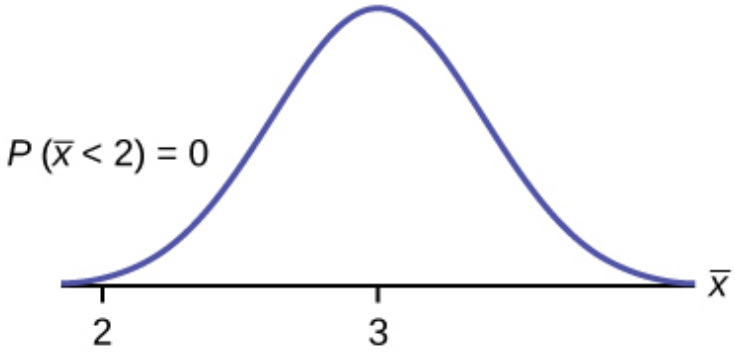


Figure 7.4

normalcdf $\left(1,2,3,\frac{1.15}{\sqrt{75}}\right) = 0$

Reminder

The smallest stress score is one.

Problem

b. Find the 90th percentile for the mean of 75 stress scores. Draw a graph.

Solution

b. Let k = the 90th precentile.

Find k, where $P(\overline{x} < k) = 0.90$.

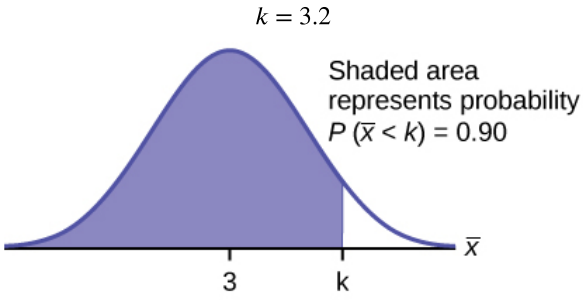


Figure 7.5

The 90th percentile for the mean of 75 scores is about 3.2. This tells us that 90 percent of all the means of 75 stress scores are at most 3.2, and that 10 percent are at least 3.2.

invNorm $\left(0.90,3,\frac{1.15}{\sqrt{75}}\right) = 3.2$

For problems (c) and (d), let ΣX = the sum of the 75 stress scores. Then, $\Sigma X \sim N[(75)(3),(\sqrt{75})(1.15)]$.

Problem

c. Find $P(\Sigma x < 200)$. Draw the graph.

Solution

c. The mean of the sum of 75 stress scores is (75)(3) = 225.

The standard deviation of the sum of 75 stress scores is $(\sqrt{75})(1.15) = 9.96$.

$$P(\Sigma x < 200) = 0$$

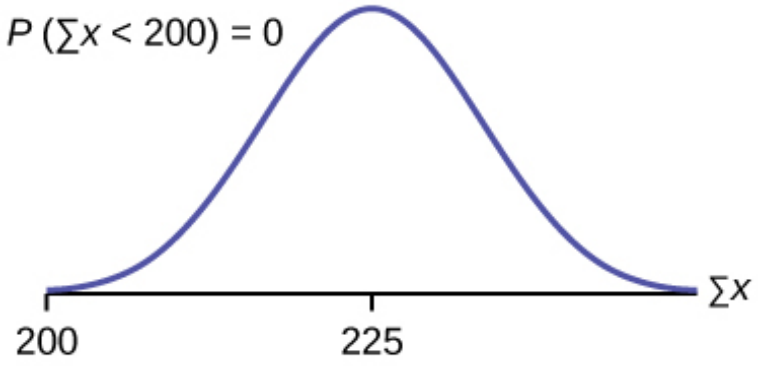


Figure 7.6

The probability that the total of 75 scores is less than 200 is about zero.

`normalcdf` (75,200,(75)(3),$(\sqrt{75})(1.15)$).

Reminder

The smallest total of 75 stress scores is 75, because the smallest single score is one.

Problem

d. Find the 90th percentile for the total of 75 stress scores. Draw a graph.

Solution

d. Let k = the 90th percentile.

Find k where $P(\Sigma x < k) = 0.90$.

$$k = 237.8$$

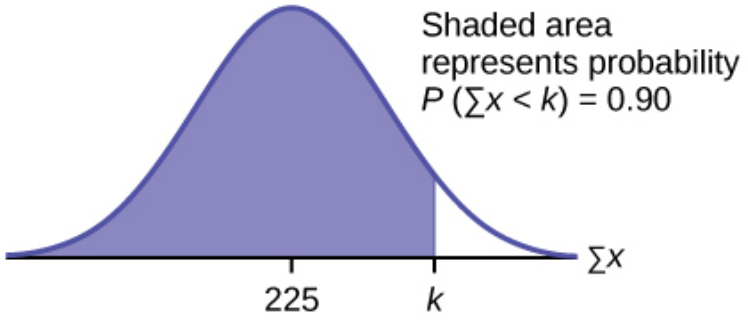


Figure 7.7

The 90th percentile for the sum of 75 scores is about 237.8. This tells us that 90 percent of all the sums of 75 scores are no more than 237.8 and 10 percent are no less than 237.8.

`invNorm`(0.90,(75)(3),$(\sqrt{75})(1.15)$) = 237.8

TRY IT 7.8

Use the information in Example 7.8, but use a sample size of 55 to answer the following questions.

a. Find $P(\bar{x} < 7)$.
b. Find $P(\Sigma x > 170)$.
c. Find the 80th percentile for the mean of 55 scores.
d. Find the 85th percentile for the sum of 55 scores.

EXAMPLE 7.9

Suppose that a market research analyst for a cell phone company conducts a study of their customers who exceed the time allowance included on their basic cell phone contract. The analyst finds that for those people who exceed the time included in their basic contract, the **excess time used** follows an **exponential distribution** with a mean of 22 minutes.

Consider a random sample of 80 customers who exceed the time allowance included in their basic cell phone contract.

Let X = the excess time used by one INDIVIDUAL cell phone customer who exceeds his contracted time allowance.

$X \sim Exp\left(\frac{1}{22}\right)$. From previous chapters, we know that $\mu = 22$ and $\sigma = 22$.

Let $\overline{X}$ = the mean excess time used by a sample of $n = 80$ customers who exceed their contracted time allowance.

$\overline{X} \sim N\left(22, \frac{22}{\sqrt{80}}\right)$ by the central limit theorem for sample means.

Problem

Using the clt to find probability

a. Find the probability that the mean excess time used by the 80 customers in the sample is longer than 20 minutes. This is asking us to find $P(\bar{x} > 20)$. Draw the graph.
b. Suppose that one customer who exceeds the time limit for his cell phone contract is randomly selected. Find the probability that this individual customer's excess time is longer than 20 minutes. This is asking us to find $P(x > 20)$.
c. Explain why the probabilities in parts (a) and (b) are different.

Solution

a. Find: $P(\bar{x} > 20)$

$P(\bar{x} > 20) = 0.79199$ using $\texttt{normalcdf}\left(20, 1\text{E}99, 22, \frac{22}{\sqrt{80}}\right)$

The probability is 0.7919 that the mean excess time used is more than 20 minutes, for a sample of 80 customers who exceed their contracted time allowance.

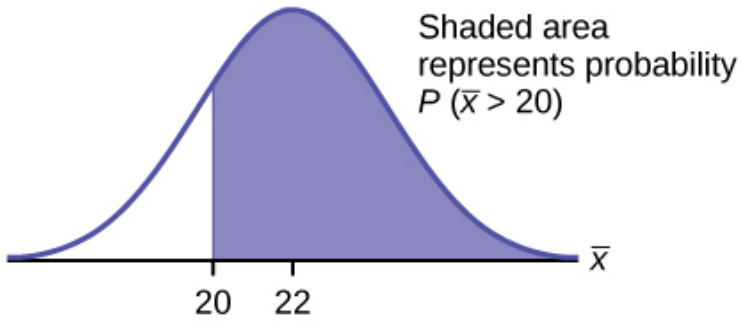


Figure 7.8

Reminder
1E99 = 10^{99} and –1E99 = -10^{99}. Press the EE key for E. Or just use 10^{99} instead of 1E99.

b. Find $P(x > 20)$. Remember to use the exponential distribution for an **individual.** $X \sim Exp\left(\frac{1}{22}\right)$.

$$P(x > 20) = e^{\left(-\left(\frac{1}{22}\right)(20)\right)} \text{ or } e^{(-0.04545(20))} = 0.4029$$

c. 1. $P(x > 20) = 0.4029$, but $P(\overline{x} > 20) = 0.7919$
   2. The probabilities are not equal because we use different distributions to calculate the probability for individuals and for means.
   3. When asked to find the probability of an individual value, use the stated distribution of its random variable; do not use the clt. Use the clt with the normal distribution when you are being asked to find the probability for a mean.

Problem

Using the clt to find percentiles

Find the 95th percentile for the **sample mean excess time** for a sample of 80 customers who exceed their basic contract time allowances. Draw a graph.

Solution

Let k = the 95th percentile. Find k where $P(\overline{x} < k) = 0.95$.

$k = 26.0$ using $\texttt{invNorm}\left(0.95, 22, \frac{22}{\sqrt{80}}\right) = 26.0$

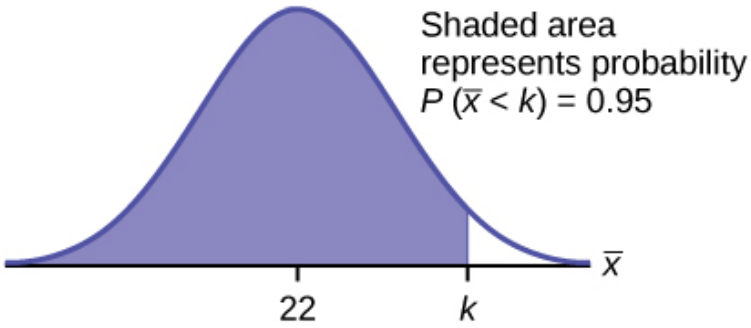


Figure 7.9

The 95th percentile for the *sample mean excess time used* is about 26.0 minutes for a random sample of 80 customers who exceed their contractual allowed time.

95 percent of such samples would have means under 26 minutes; only five percent of such samples would have means above 26 minutes.

TRY IT 7.9

Use the information in Example 7.9, but change the sample size to 144.

a. Find $P(20 < \overline{x} < 30)$.
b. Find $P(\Sigma x \text{ is at least } 3000)$.
c. Find the 75th percentile for the sample mean excess time of 144 customers.
d. Find the 85th percentile for the sum of 144 excess times used by customers.

EXAMPLE 7.10

U.S. scientists studying a certain medical condition discovered that a new person is diagnosed every two minutes, on average. Suppose the standard deviation is 0.5 minutes and the sample size is 100.

Problem

a. Find the median, the first quartile, and the third quartile for the sample mean time of diagnosis in the United States.

b. Find the median, the first quartile, and the third quartile for the sum of sample times of diagnosis in the United States.
c. Find the probability that a diagnosis occurs on average between 1.75 and 1.85 minutes.
d. Find the value that is two standard deviations above the sample mean.
e. Find the *IQR* for the sum of the sample times.

Solution

a. We have $\mu_X = \mu = 2$ and $\sigma_X = \frac{\sigma}{\sqrt{n}} = \frac{0.5}{10} = 0.05$. Therefore,
   1. 50^{th} percentile = $\mu_X = \mu = 2$,
   2. 25^{th} percentile = *invNorm*(0.25,2,0.05) = 1.97, and
   3. 75^{th} percentile = *invNorm*(0.75,2,0.05) = 2.03.

b. We have $\mu_{\Sigma X} = n(\mu_X) = 100(2) = 200$ and $\sigma_{\mu X} = \sqrt{n}(\sigma_X) = 10(0.5) = 5$. Therefore,
   1. 50^{th} percentile = $\mu_{\Sigma X} = n(\mu_X) = 100(2) = 200$,
   2. 25^{th} percentile = invNorm(0.25,200,5) = 196.63, and
   3. 75^{th} percentile = invNorm(0.75,200,5) = 203.37.

c. $P(1.75 < \overline{x} < 1.85)$ = `normalcdf`(1.75,1.85,2,0.05) = 0.0013
d. Using the *z*-score equation, $z = \frac{\overline{x} - \mu_{\overline{x}}}{\sigma_{\overline{x}}}$, and solving for *x*, we get $x = 2(0.05) + 2 = 2.1$.
e. The *IQR* is 75^{th} percentile – 25^{th} percentile = 203.37 – 196.63 = 6.74.

TRY IT 7.10

Based on data from the National Health Survey, women between the ages of 18 and 24 have an average systolic blood pressures (in mm Hg) of 114.8 with a standard deviation of 13.1. Systolic blood pressure for women between the ages of 18 to 24 follows a normal distribution.

a. If one woman from this population is randomly selected, find the probability that her systolic blood pressure is greater than 120.
b. If 40 women from this population are randomly selected, find the probability that their mean systolic blood pressure is greater than 120.
c. If the sample was four women between the ages of 18–24 and we did not know the original distribution, could the central limit theorem be used?

EXAMPLE 7.11

Problem

A study was done about a medical condition that affects a certain group of people. The age range of the people was 14–61. The mean age was 30.9 years with a standard deviation of nine years.

a. In a sample of 25 people, what is the probability that the mean age of the people is less than 35?
b. Is it likely that the mean age of the sample group could be more than 50 years? Interpret the results.
c. In a sample of 49 people, what is the probability that the sum of the ages is no less than 1,600?
d. Is it likely that the sum of the ages of the 49 people are at most 1,595? Interpret the results.
e. Find the 95^{th} percentile for the sample mean age of 65 people. Interpret the results.
f. Find the 90^{th} percentile for the sum of the ages of 65 people. Interpret the results.

Solution

a. $P(\overline{x} < 35)$ = `normalcdf`(-*E*99,35,30.9,1.8) = 0.9886
b. $P(\overline{x} > 50)$ = `normalcdf`(50, *E*99,30.9,1.8) ≈ 0. For this sample group, it is almost impossible for the group's average age to be more than 50. However, it is still possible for an individual in this group to have an age greater than 50.
c. $P(\Sigma x \geq 1{,}600)$ = `normalcdf`(1600,E99,1514.10,63) = 0.0864
d. $P(\Sigma x \leq 1{,}595)$ = `normalcdf`(-E99,1595,1514.10,63) = 0.9005. This means that there is a 90 percent chance that the sum of the ages for the sample group $n = 49$ is at most 1,595.

e. The 95th percentile = `invNorm`(0.95,30.9,1.1) = 32.7. This indicates that 95 percent of the people in the sample of 65 are younger than 32.7 years, on average.
f. The 90th percentile = `invNorm`(0.90,2008.5,72.56) = 2101.5. This indicates that 90 percent of the people in the sample of 65 have a sum of ages less than 2,101.5 years.

TRY IT 7.11

According to data from an aerospace company, the 757 airliner carries 200 passengers and has doors with a mean height of 72 inches. Assume for a certain population of men we have a mean of 69 inches inches and a standard deviation of 2.8 inches.

a. What mean doorway height would allow 95 percent of men to enter the aircraft without bending?
b. Assume that half of the 200 passengers are men. What mean doorway height satisfies the condition that there is a 0.95 probability that this height is greater than the mean height of 100 men?
c. For engineers designing the 757, which result is more relevant: the height from part (a) or part (b)? Why?

HISTORICAL NOTE

Normal Approximation to the Binomial

Historically, being able to compute binomial probabilities was one of the most important applications of the central limit theorem. Binomial probabilities with a small value for n (say, 20) were displayed in a table in a book. To calculate the probabilities with large values of n, you had to use the binomial formula, which could be very complicated. Using the *normal approximation to the binomial* distribution simplified the process. To compute the normal approximation to the binomial distribution, take a simple random sample from a population. You must meet the following conditions for a **binomial distribution**:

- There are a certain number, n, of independent trials.
- The outcomes of any trial are success or failure.
- Each trial has the same probability of a success, p.

Recall that if X is the binomial random variable, then $X \sim B(n, p)$. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five ($np > 5$ and $nq > 5$; the approximation is better if they are both greater than or equal to 10. The product >5 is more or less accepted as the norm here.). This is another accepted rule. So, for whatever value of x we are looking at (the number of successes). We add 0.5 if we are looking for the probability that is less than or equal to that number. We subtract 0.5 if we are looking for the probability that is greater than or equal to that number. Then the binomial can be approximated by the normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$. Remember that $q = 1 - p$. In order to get the best approximation, add 0.5 to x or subtract 0.5 from x (use $x + 0.5$ or $x - 0.5$).

This is another accepted rule. So, for whatever value of x we are looking at (the number of successes). We add 0.5 if we are looking for the probability that is less than or equal to that number. We subtract 0.5 if we are looking for the probability that is greater than or equal to that number. The number 0.5 is called the **continuity correction factor** and is used in the following example.

EXAMPLE 7.12

Suppose in a local kindergarten through 12th grade (K–12) school district, 53 percent of the population favor a charter school for grades K through 5. A simple random sample of 300 is surveyed.

a. Find the probability that *at least 150* favor a charter school.
b. Find the probability that *at most 160* favor a charter school.

c. Find the probability that *more than 155* favor a charter school.
d. Find the probability that *fewer than 147* favor a charter school.
e. Find the probability that *exactly 175* favor a charter school.

Let X = the number that favor a charter school for grades K through 5. $X \sim B(n, p)$ where $n = 300$ and $p = 0.53$. Because $np > 5$ and $nq > 5$, use the normal approximation to the binomial. The formulas for the mean and standard deviation are $\mu = np$ and $\sigma = \sqrt{npq}$. The mean is 159, and the standard deviation is 8.6447. The random variable for the normal distribution is Y. $Y \sim N(159, 8.6447)$. See The Normal Distribution for help with calculator instructions.

For Part (a), you *include 150* so $P(X \geq 150)$ has a normal approximation $P(Y \geq 149.5) = 0.8641$.

`normalcdf`(149.5,10^99,159,8.6447) = 0.8641.

For Part (b), you *include 160* so $P(X \leq 160)$ has a normal approximation $P(Y \leq 160.5) = 0.5689$.

`normalcdf`(0,160.5,159,8.6447) = 0.5689

For Part (c), you *exclude 155* so $P(X > 155)$ has normal approximation $P(y > 155.5) = 0.6572$.

`normalcdf`(155.5,10^99,159,8.6447) = 0.6572.

For Part (d), you *exclude 147* so $P(X < 147)$ has normal approximation $P(Y < 146.5) = 0.0741$.

`normalcdf`(0,146.5,159,8.6447) = 0.0741

For Part (e), $P(X = 175)$ has normal approximation $P(174.5 < Y < 175.5) = 0.0083$.

`normalcdf`(174.5,175.5,159,8.6447) = 0.0083

Because of calculators and computer software that let you calculate binomial probabilities for large values of n easily, it is not necessary to use the the normal approximation to the binomial distribution, provided that you have access to these technology tools. Most school labs have computer software that calculates binomial probabilities. Many students have access to calculators that calculate probabilities for binomial distribution. If you type in *binomial probability distribution calculation* in an internet browser, you can find at least one online calculator for the binomial.

For Example 7.10, the probabilities are calculated using the following binomial distribution: ($n = 300$ and $p = 0.53$). Compare the binomial and normal distribution answers. See Discrete Random Variables for help with calculator instructions for the binomial.

$P(X \geq 150)$:`1 - binomialcdf`(300,0.53,149) = 0.8641

$P(X \leq 160)$:`binomialcdf`(300,0.53,160) = 0.5684

$P(X > 155)$:`1 - binomialcdf`(300,0.53,155) = 0.6576

$P(X < 147)$:`binomialcdf`(300,0.53,146) = 0.0742

$P(X = 175)$:(You use the binomial pdf.)`binomialpdf`(300,0.53,175) = 0.0083

TRY IT 7.12

In a city, 46 percent of the population favors the incumbent, Dawn Morgan, for mayor. A simple random sample of 500 is taken. Using the continuity correction factor, find the probability that at least 250 favor Dawn Morgan for mayor.

7.4 Central Limit Theorem (Pocket Change)

Stats Lab

Central Limit Theorem (Pocket Change)

Student Learning Outcome

- The student will demonstrate and compare properties of the central limit theorem.

Note

This lab works best when sampling from several classes and combining data.

Collect the Data

1. Count the change in your pocket. (Do not include bills.)
2. Randomly survey 30 classmates. Record the values of the change in Table 7.1.

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

Table 7.1

3. Construct a histogram. Make five to six intervals. Sketch the graph using a ruler and pencil. Scale the axes.

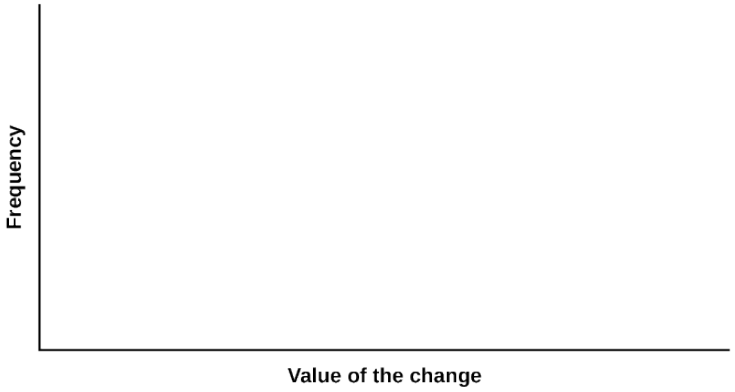


Figure 7.10

4. Calculate the following ($n = 1$, surveying one person at a time):
   a. $\overline{x}$ = _____
   b. s = _____
5. Draw a smooth curve through the tops of the bars of the histogram. Use one to two complete sentences to describe the general shape of the curve.

Collecting Averages of Pairs:

Repeat steps one through five of the section Collect the Data with one exception. Instead of recording the change of 30

classmates, record the average change of 30 pairs.

1. Randomly survey 30 *pairs* of classmates.
2. Record the values of the average of their change in Table 7.2.

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

Table 7.2

3. Construct a histogram. Scale the axes using the same scaling you used for the section titled Collect the Data. Sketch the graph using a ruler and a pencil.

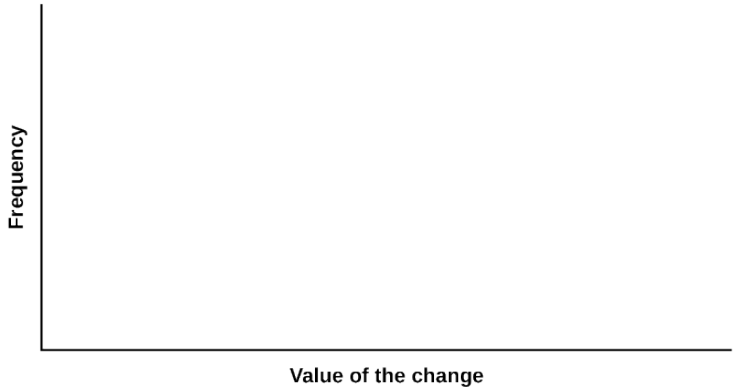


Figure 7.11

4. Calculate the following ($n = 2$, surveying two people at a time):
   a. $\overline{x}$ = _____
   b. s = _____
5. Draw a smooth curve through the tops of the bars of the histogram. Use one to two complete sentences to describe the general shape of the curve.

Collecting Averages of Groups of Five:

Repeat steps one through five (of the section titled Collect the Data), with one exception. Instead of recording the change of 30 classmates, record the average change of 30 groups of five.

1. Randomly survey 30 *groups of five* classmates.
2. Record the values of the averages of their change.

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

Table 7.3

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

Table 7.3

3. Construct a histogram. Scale the axes using the same scaling you used for the section titled Collect the Data. Sketch the graph using a ruler and a pencil.

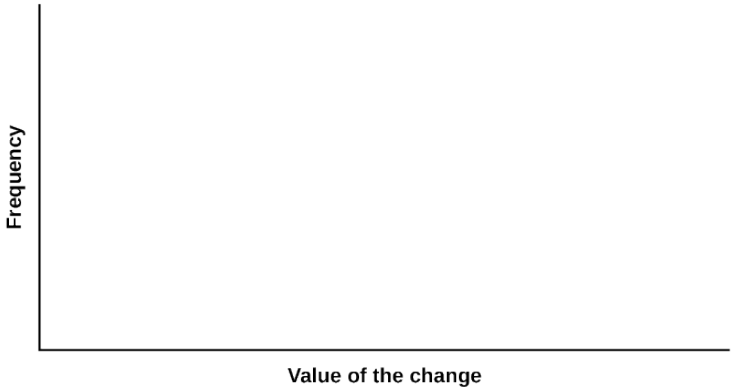


Figure 7.12

4. Calculate the following ($n = 5$, surveying five people at a time):
   a. $\overline{x}$ = _____
   b. s = _____
5. Draw a smooth curve through the tops of the bars of the histogram. Use one to two complete sentences to describe the general shape of the curve.

Discussion Questions

1. Why did the shape of the distribution of the data change, as n changed? Use one to two complete sentences to explain what happened.
2. In the section titled Collect the Data, what was the approximate distribution of the data?
3. X ~ _____(_____, _____)
4. In the section titled Collecting Averages of Groups of Five, what was the approximate distribution of the averages? $\overline{X}$ ~ _____(_____, _____)
5. In one to two complete sentences, explain any differences in your answers to the previous two questions.

7.5 Central Limit Theorem (Cookie Recipes)

Stats Lab

Central Limit Theorem (Cookie Recipes)

Student Learning Outcome

- The student will demonstrate and compare properties of the central limit theorem.

Given

X = length of time (in days) that a cookie recipe lasted at the Olmstead Homestead. (Assume that each of the different recipes makes the same quantity of cookies.)

Recipe #	X		Recipe #	X		Recipe #	X		Recipe #	X
1	1		16	2		31	3		46	2
2	5		17	2		32	4		47	2
3	2		18	4		33	5		48	11
4	5		19	6		34	6		49	5
5	6		20	1		35	6		50	5
6	1		21	6		36	1		51	4
7	2		22	5		37	1		52	6
8	6		23	2		38	2		53	5
9	5		24	5		39	1		54	1
10	2		25	1		40	6		55	1
11	5		26	6		41	1		56	2
12	1		27	4		42	6		57	4
13	1		28	1		43	2		58	3
14	3		29	6		44	6		59	6
15	2		30	2		45	2		60	5

Table 7.4

Calculate the following:

a. μ_X = _____
b. σ_X = _____

Collect the Data

Use a random number generator to randomly select four samples of size $n = 5$ from the given population. Record your samples in Table 7.5. Then, for each sample, calculate the mean to the nearest tenth. Record them in the spaces provided. Record the sample means for the rest of the class.

1. Complete the following table:

	Sample 1	Sample 2	Sample 3	Sample 4	Sample means from other groups:

Table 7.5

	Sample 1	Sample 2	Sample 3	Sample 4	Sample means from other groups:
Means:	$\overline{x}$ = ____	$\overline{x}$ = ____	$\overline{x}$ = ____	$\overline{x}$ = ____	

Table 7.5

2. Calculate the following:
   a. $\overline{x}$ = _____
   b. $s_{\overline{x}}$ = _____
3. Again, use a random number generator to randomly select four samples from the population. This time, make the samples of size n = 10. Record the samples in Table 7.6. As before, for each sample, calculate the mean to the nearest tenth. Record them in the spaces provided. Record the sample means for the rest of the class.

	Sample 1	Sample 2	Sample 3	Sample 4	Sample means from other groups
Means:	$\overline{x}$ = ____	$\overline{x}$ = ____	$\overline{x}$ = ____	$\overline{x}$ = ____	

Table 7.6

4. Calculate the following:
   a. $\overline{x}$ = _____
   b. $s_{\overline{x}}$ = _____
5. For the original population, construct a histogram. Make intervals with a bar width of one day. Sketch the graph using a ruler and pencil. Scale the axes.

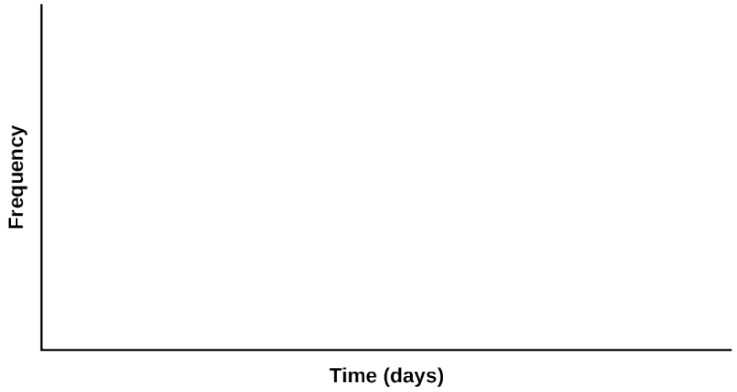


Figure 7.13

6. Draw a smooth curve through the tops of the bars of the histogram. Use one to two complete sentences to describe the general shape of the curve.

Repeat the procedure for n = 5.

1. For the sample of n = 5 days averaged together, construct a histogram of the averages (your means together with

the means of the other groups). Make intervals with bar widths of $\frac{1}{2}$ day. Sketch the graph using a ruler and pencil. Scale the axes.

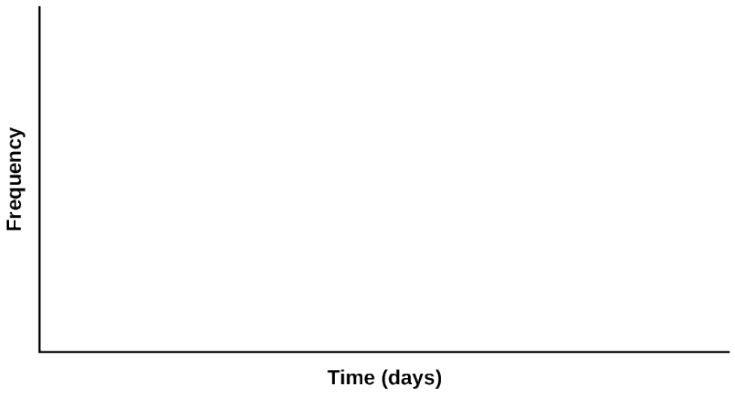


Figure 7.14

2. Draw a smooth curve through the tops of the bars of the histogram. Use one to two complete sentences to describe the general shape of the curve.

Repeat the procedure for *n* = 10.

1. For the sample of *n* = 10 days averaged together, construct a histogram of the averages (your means together with the means of the other groups). Make intervals with bar widths of $\frac{1}{2}$ day. Sketch the graph using a ruler and pencil. Scale the axes.

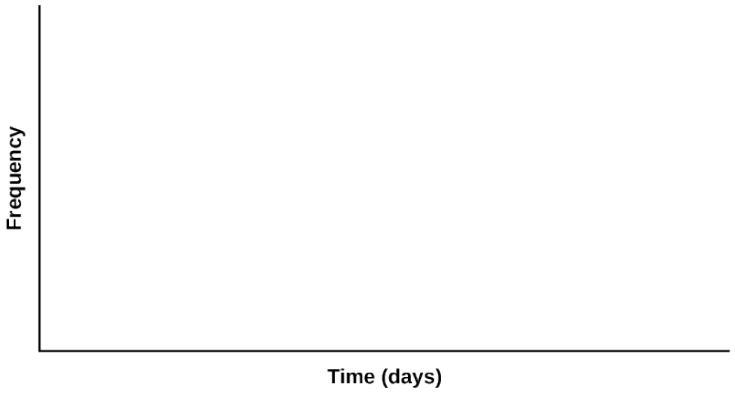


Figure 7.15

2. Draw a smooth curve through the tops of the bars of the histogram. Use one to two complete sentences to describe the general shape of the curve.

Discussion Questions

1. Compare the three histograms you have made, the one for the population and the two for the sample means. In three to five sentences, describe the similarities and differences.
2. State the theoretical (according to the clt) distributions for the sample means.
   a. *n* = 5: $\overline{x}$ ~ _____(_____, _____)
   b. *n* = 10: $\overline{x}$ ~ _____(_____, _____)
3. Are the sample means for *n* = 5 and *n* = 10 *close* to the theoretical mean, μ_x? Explain why or why not.
4. Which of the two distributions of sample means has the smaller standard deviation? Why?
5. As *n* changed, why did the shape of the distribution of the data change? Use one to two complete sentences to explain what happened.

Key Terms

average a number that describes the central tendency of the data; there are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean

central limit theorem given a random variable (RV) with a known mean, μ, and known standard deviation, σ, and sampling with size n, we are interested in two new RVs: the sample mean, $\overline{X}$, and the sample sum, ΣX
If the size (n) of the sample is sufficiently large, then $\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$ and $\Sigma X \sim N(n\mu, (\sqrt{n})(\sigma))$. If the size ($n$) of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean, and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean

exponential distribution a continuous random variable (RV) that appears when we are interested in the intervals of time between a random events; for example, the length of time between emergency arrivals at a hospital, notation: $X \sim Exp(m)$
The mean is $\mu = \frac{1}{m}$ and the standard deviation is $\sigma = \frac{1}{m}$. The probability density function is $f(x) = me^{-mx}$, $x \geq 0$, and the cumulative distribution function is $P(X \leq x) = 1 - e^{-mx}$

mean a number that measures the central tendency; a common name for mean is *average*; the term *mean* is a shortened form of *arithmetic mean;*.
by definition, the mean for a sample (denoted by $\overline{x}$) is $\overline{x} = \frac{\text{sum of all values in the sample}}{\text{number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{sum of all values in the population}}{\text{number of values in the population}}$.

normal distribution a continuous random variable (RV) with probability density function (pdf) $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, where μ is the mean of the distribution and σ is the standard deviation; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called a **standard normal distribution**

sampling distribution given simple random samples of size n from a given population with a measured characteristic such as mean, proportion, or standard deviation for each sample, the probability distribution of all the measured characteristics is called a sampling distribution.

standard error of the mean the standard deviation of the distribution of the sample means, or $\frac{\sigma}{\sqrt{n}}$

uniform distribution a continuous random variable (RV) that has equally likely outcomes over the domain $a < x < b$; often referred as the **rectangular distribution** because the graph of the pdf has the form of a rectangle
Notation: $X \sim U(a, b)$. The mean is $\mu = \frac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$. The probability density function is $f(x) = \frac{1}{b-a}$ for $a < x < b$ or $a \leq x \leq b$. The cumulative distribution is $P(X \leq x) = \frac{x-a}{b-a}$

Chapter Review

7.1 The Central Limit Theorem for Sample Means (Averages)

In a population whose distribution may be known or unknown, if the size (n) of the sample is sufficiently large, the distribution of the sample means will be approximately normal. The mean of the sample means will equal the population mean. The standard deviation of the distribution of the sample means, called the standard error of the mean, is equal to the population standard deviation divided by the square root of the sample size (n).

7.2 The Central Limit Theorem for Sums (Optional)

The central limit theorem tells us that for a population with any distribution, the distribution of the sums for the sample means approaches a normal distribution as the sample size increases. In other words, if the sample size is large enough, the distribution of the sums can be approximated by a normal distribution, even if the original population is not normally distributed. Additionally, if the original population has a mean of μ_X and a standard deviation of σ_X, the mean of the sums is $n\mu_X$ and the standard deviation is $(\sqrt{n})(\sigma_X)$, where n is the sample size.

7.3 Using the Central Limit Theorem

The central limit theorem can be used to illustrate the law of large numbers. The law of large numbers states that the larger the sample size you take from a population, the closer the sample mean, $\overline{x}$, gets to μ.

Formula Review

7.1 The Central Limit Theorem for Sample Means (Averages)

Central limit theorem for sample means: $\overline{X} \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$

Mean $\overline{X}$: μ_X

Central limit theorem for sample means z-score and standard error of the mean: $z = \frac{\overline{x} - \mu_X}{\left(\frac{\sigma_X}{\sqrt{n}}\right)}$

Standard error of the mean (standard deviation ($\overline{X}$)): $\frac{\sigma_X}{\sqrt{n}}$

7.2 The Central Limit Theorem for Sums (Optional)

Central limit theorem for sums: $\sum X \sim N[(n)(\mu_X), (\sqrt{n})(\sigma_X)]$

Mean for sums ($\sum X$): $(n)(\mu_X)$

Central limit theorem for sums z-score and standard deviation for sums: z for the sample mean $= \frac{\sum x - (n)(\mu_X)}{(\sqrt{n})(\sigma_X)}$

Standard deviation for sums ($\sum X$): $(\sqrt{n})(\sigma_X)$

Practice

7.1 The Central Limit Theorem for Sample Means (Averages)

Use the following information to answer the next six exercises: Yoonie is a personnel manager in a large corporation. Each month she must review 16 of the employees. From past experience, she has found that the reviews take her approximately four hours each to do with a population standard deviation of 1.2 hours. Let X be the random variable representing the time it takes her to complete one review. Assume X is normally distributed. Let $\overline{X}$ be the random variable representing the mean time to complete the 16 reviews. Assume that the 16 reviews represent a random set of reviews.

1. What is the mean, standard deviation, and sample size?

2. Complete the distributions.
   a. $X \sim$ _____(_____, _____)
   b. $\overline{X} \sim$ _____(_____, _____)

3. Find the probability that *one* review will take Yoonie from 3.5 to 4.25 hours. Sketch the graph, labeling and scaling the horizontal axis. Shade the region corresponding to the probability.
   a.

Figure 7.16

   b. P(_____ < x < _____) = _____

4. Find the probability that the *mean* of a month's reviews will take Yoonie from 3.5 to 4.25 hrs. Sketch the graph, labeling and scaling the horizontal axis. Shade the region corresponding to the probability.

a.

Figure 7.17

b. P(_____) = _____

5. What causes the probabilities in Exercise 7.3 and Exercise 7.4 to be different?
6. Find the 95th percentile for the mean time to complete one month's reviews. Sketch the graph.

a.

Figure 7.18

b. The 95th percentile =_____

7.2 The Central Limit Theorem for Sums (Optional)

Use the following information to answer the next four exercises: An unknown distribution has a mean of 80 and a standard deviation of 12. A sample size of 95 is drawn randomly from the population.

7. Find the probability that the sum of the 95 values is greater than 7,650.
8. Find the probability that the sum of the 95 values is less than 7,400.
9. Find the sum that is two standard deviations above the mean of the sums.
10. Find the sum that is 1.5 standard deviations below the mean of the sums.

Use the following information to answer the next five exercises: The distribution of results from a cholesterol test has a mean of 180 and a standard deviation of 20. A sample size of 40 is drawn randomly.

11. Find the probability that the sum of the 40 values is greater than 7,500.
12. Find the probability that the sum of the 40 values is less than 7,000.
13. Find the sum that is one standard deviation above the mean of the sums.
14. Find the sum that is 1.5 standard deviations below the mean of the sums.
15. Find the percentage of sums between 1.5 standard deviations below the mean of the sums and one standard deviation above the mean of the sums.

Use the following information to answer the next six exercises: A researcher measures the amount of sugar in several cans of the same type of soda. The mean is 39.01 with a standard deviation of 0.5. The researcher randomly selects a sample of 100.

16. Find the probability that the sum of the 100 values is greater than 3,910.

17. Find the probability that the sum of the 100 values is less than 3,900.

18. Find the probability that the sum of the 100 values falls between the numbers you found in [link] (16) and [link] (17).

19. Find the sum with a *z*-score of –2.5.

20. Find the sum with a *z*-score of 0.5.

21. Find the probability that the sums will fall between the *z*-scores –2 and 1.

Use the following information to answer the next four exercises: An unknown distribution has a mean 12 and a standard deviation of one. A sample size of 25 is taken. Let X = the object of interest.

22. What is the mean of ΣX?

23. What is the standard deviation of ΣX?

24. What is $P(\Sigma x = 290)$?

25. What is $P(\Sigma x > 290)$?

26. True or False: Only the sums of normal distributions are also normal distributions.

27. In order for the sums of a distribution to approach a normal distribution, what must be true?

28. What three things must you know about a distribution to find the probability of sums?

29. An unknown distribution has a mean of 25 and a standard deviation of six. Let X = one object from this distribution. What is the sample size if the standard deviation of ΣX is 42?

30. An unknown distribution has a mean of 19 and a standard deviation of 20. Let X = the object of interest. What is the sample size if the mean of ΣX is 15,200?

Use the following information to answer the next three exercises: A market researcher analyzes how many electronics devices customers buy in a single purchase. The distribution has a mean of three with a standard deviation of 0.7. She samples 400 customers.

31. What is the *z*-score for $\Sigma x = 840$?

32. What is the *z*-score for $\Sigma x = 1{,}186$?

33. What is $P(\Sigma x < 1186)$?

Use the following information to answer the next three exercises: An unkwon distribution has a mean of 100, a standard deviation of 100, and a sample size of 100. Let X = one object of interest.

34. What is the mean of ΣX?

35. What is the standard deviation of ΣX?

36. What is $P(\Sigma x > 9000)$?

7.3 Using the Central Limit Theorem

Use the following information to answer the next 10 exercises: A manufacturer produces 25-pound lifting weights. The lowest actual weight is 24 pounds, and the highest is 26 pounds. Each weight is equally likely, so the distribution of weights is uniform. A sample of 100 weights is taken.

37. a. What is the distribution for the weights of one 25-pound lifting weight? What are the mean and standard deivation?

b. What is the distribution for the mean weight of 100 25-pound lifting weights?
c. Find the probability that the mean actual weight for the 100 weights is less than 24.9.

38. Draw the graph of Exercise 7.37.

39. Find the probability that the mean actual weight for the 100 weights is greater than 25.2.

40. Draw the graph of Exercise 7.39.

41. Find the 90th percentile for the mean weight for the 100 weights.

42. Draw the graph of Exercise 7.41.

43. a. What is the distribution for the sum of the weights of 100 25-pound lifting weights?
b. Find $P(\Sigma x < 2450)$.

44. Draw the graph of Exercise 7.43.

45. Find the 90th percentile for the total weight of the 100 weights.

46. Draw the graph of Exercise 7.45.

Use the following information to answer the next five exercises: The length of time a particular smartphone's battery lasts follows an exponential distribution with a mean of ten months. A sample of 64 of these smartphones is taken.

47. a. What is the standard deviation?
b. What is the parameter m?

48. What is the distribution for the length of time one battery lasts?

49. What is the distribution for the mean length of time 64 batteries last?

50. What is the distribution for the total length of time 64 batteries last?

51. Find the probability that the sample mean is between 7 and 11.

52. Find the 80th percentile for the total length of time 64 batteries last.

53. Find the interquartile range (IQR) for the mean amount of time 64 batteries last.

54. Find the middle 80 percent for the total amount of time 64 batteries last.

Use the following information to answer the next six exercises: A uniform distribution has a minimum of six and a maximum of ten. A sample of 50 is taken.

55. Find $P(\Sigma x > 420)$.

56. Find the 90th percentile for the sums.

57. Find the 15th percentile for the sums.

58. Find the first quartile for the sums.

59. Find the third quartile for the sums.

60. Find the 80th percentile for the sums.

Homework

7.1 The Central Limit Theorem for Sample Means (Averages)

61. Previously, De Anza's statistics students estimated that the amount of change daytime statistics students carry is exponentially distributed with a mean of $0.88. Suppose that we randomly pick 25 daytime statistics students.
a. In words, X = _____.
b. X ~ _____(_____, _____)
c. In words, $\overline{X}$ = _____.
d. $\overline{X}$ ~ _____ (_____, _____)
e. Find the probability that an individual had between $0.80 and $1.00. Graph the situation, and shade in the

area to be determined.

f. Find the probability that the average amount of change of the 25 students was between $0.80 and $1.00. Graph the situation, and shade in the area to be determined.

g. Explain why there is a difference in part (e) and part (f).

62. Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet. We randomly sample 49 fly balls.

a. If $\overline{X}$ = average distance in feet for 49 fly balls, then $\overline{X}$ ~ _____(_____, _____).

b. What is the probability that the 49 balls traveled an average of less than 240 feet? Sketch the graph. Scale the horizontal axis for $\overline{X}$. Shade the region corresponding to the probability. Find the probability.

c. Find the 80th percentile of the distribution of the average of 49 fly balls.

63. According to the Internal Revenue Service, the average length of time for an individual to complete (keep records for, learn, prepare, copy, assemble, and send) IRS Form 1040 is 10.53 hours (without any attached schedules). The distribution is unknown. Let us assume that the standard deviation is two hours. Suppose we randomly sample 36 taxpayers.

a. In words, X = _____.

b. In words, $\overline{X}$ = _____.

c. $\overline{X}$ ~ _____(_____, _____)

d. Would you be surprised if the 36 taxpayers finished their Form 1040s in an average of more than 12 hours? Explain why or why not in complete sentences.

e. Would you be surprised if one taxpayer finished his or her Form 1040 in more than 12 hours? In a complete sentence, explain why.

64. Suppose that a category of world-class runners are known to run a marathon (26 miles) in an average of 145 minutes with a standard deviation of 14 minutes. Consider 49 of the races. Let $\overline{X}$ be the average of the 49 races.

a. $\overline{X}$ ~ _____(_____, _____)

b. Find the probability that the runner will average between 142 and 146 minutes in these 49 marathons.

c. Find the 80th percentile for the average of these 49 marathons.

d. Find the median of the average running times.

65. The length of songs in a collector's online album collection is uniformly distributed from 2 to 3.5 minutes. Suppose we randomly pick five albums from the collection. There are a total of 43 songs on the five albums.

a. In words, X = _____.

b. X ~ _____

c. In words, $\overline{X}$ = _____.

d. $\overline{X}$ ~ _____(_____, _____)

e. Find the first quartile for the average song length.

f. The IQR for the average song length is _____–_____.

66. In 1940, the average size of a U.S. farm was 174 acres. Let's say that the standard deviation was 55 acres. Suppose we randomly survey 38 farmers from 1940.

a. In words, X = _____.

b. In words, $\overline{X}$ = _____.

c. $\overline{X}$ ~ _____(_____, _____)

d. The IQR for $\overline{X}$ is from _____ acres to _____ acres.

67. Determine which of the following are true and which are false. Then, in complete sentences, justify your answers.

a. When the sample size is large, the mean of $\overline{X}$ is approximately equal to the mean of X.

b. When the sample size is large, $\overline{X}$ is approximately normally distributed.

c. When the sample size is large, the standard deviation of $\overline{X}$ is approximately the same as the standard deviation of X.

68. The percentage of fat calories that a person in America consumes each day is normally distributed with a mean of about 36 and a standard deviation of about ten. Suppose that 16 individuals are randomly chosen. Let $\overline{X}$ = average percentage of fat calories.

a. $\overline{X}$ ~ _____(_____, _____)

b. For the group of 16, find the probability that the average percentage of fat calories consumed is more than

five. Graph the situation and shade in the area to be determined.
c. Find the first quartile for the average percentage of fat calories.

69. The distribution of income in some economically developing countries is considered wedge shaped (many very poor people, very few middle income people, and even fewer wealthy people). Suppose we pick a country with a wedge-shaped distribution. Let the average salary be $2,000 per year with a standard deviation of $8,000. We randomly survey 1,000 residents of that country.
a. In words, X = _____.
b. In words, $\overline{X}$ = _____.
c. $\overline{X}$ ~ _____(_____, _____)
d. How is it possible for the standard deviation to be greater than the average?
e. Why is it more likely that the average salary of the 1,000 residents will be from $2,000 to $2,100 than from $2,100 to $2,200?

70. Which of the following is NOT true about the distribution for averages?
a. The mean, median, and mode are equal.
b. The area under the curve is 1.
c. The curve never touches the x-axis.
d. The curve is skewed to the right.

71. The cost of unleaded gasoline in the Bay Area once followed an unknown distribution with a mean of $4.59 and a standard deviation of $0.10. Sixteen gas stations from the Bay Area are randomly chosen. We are interested in the average cost of gasoline for the 16 gas stations. The distribution to use for the average cost of gasoline for the 16 gas stations is:
a. $\overline{X} \sim N(4.59, 0.10)$
b. $\overline{X} \sim N\left(4.59, \frac{0.10}{\sqrt{16}}\right)$
c. $\overline{X} \sim N\left(4.59, \frac{16}{0.10}\right)$
d. $\overline{X} \sim N\left(4.59, \frac{\sqrt{16}}{0.10}\right)$

7.2 The Central Limit Theorem for Sums (Optional)

72. Which of the following is NOT true about the theoretical distribution of sums?
a. The mean, median, and mode are equal.
b. The area under the curve is one.
c. The curve never touches the x-axis.
d. The curve is skewed to the right.

73. Suppose that the duration of a particular type of criminal trial is known to have a mean of 21 days and a standard deviation of seven days. We randomly sample nine trials.
a. In words, ΣX = _____.
b. ΣX ~ _____(_____, _____)
c. Find the probability that the total length of the nine trials is at least 225 days.
d. Ninety percent of the total of nine of these types of trials will last at least how long?

74. Suppose that the weight of open boxes of cereal in a home with children is uniformly distributed from two to six pounds with a mean of four pounds and standard deviation of 1.1547. We randomly survey 64 homes with children.
a. In words, X = _____.
b. The distribution is _____.
c. In words, ΣX = _____.
d. ΣX ~ _____(_____, _____)
e. Find the probability that the total weight of the open boxes is less than 250 pounds.
f. Find the 35^{th} percentile for the total weight of open boxes of cereal.

75. Salaries for entry-level managers at a restaurant chain are normally distributed with a mean of $44,000 and a standard deviation of $6,500. We randomly survey 10 managers from these restaurants.

a. In words, X = _____.
b. X ~ _____(_____, _____)
c. In words, ΣX = _____.
d. ΣX ~ _____(_____, _____)
e. Find the probability that the managers earn a total of over $400,000.
f. Find the 90^{th} percentile for an individual manager's salary.
g. Find the 90^{th} percentile for the sum of ten managers' salary.
h. If we surveyed 70 managers instead of ten, graphically, how would that change the distribution in part (d)?
i. If each of the 70 managers received a $3,000 raise, graphically, how would that change the distribution in part (b)?

7.3 Using the Central Limit Theorem

76. The attention span of a two-year-old is exponentially distributed with a mean of about eight minutes. Suppose we randomly survey 60 two-year-olds.

a. In words, X = _____.
b. X ~ _____(_____, _____)
c. In words, $\overline{X}$ = _____.
d. $\overline{X}$ ~ _____(_____, _____)
e. Before doing any calculations, which do you think will be higher? Explain why.
    i. The probability that an individual attention span is less than 10 minutes.
    ii. The probability that the average attention span for the 60 children is less than 10 minutes.
f. Calculate the probabilities in part (e).
g. Explain why the distribution for $\overline{X}$ is not exponential.

77. The closing stock prices of 35 U.S. semiconductor manufacturers are given as follows:

Company	Closing Stock Prices
1	8.625
2	30.25
3	27.625
4	46.75
5	32.875
6	18.25
7	5
8	0.125
9	2.9375
10	6.875
11	28.25
12	24.25

Table 7.7

Company	Closing Stock Prices
13	21
14	1.5
15	30.25
16	71
17	43.5
18	49.25
19	2.5625
20	31
21	16.5
22	9.5
23	18.5
24	18
25	9
26	10.5
27	16.625
28	1.25
29	18
30	12.87
31	7
32	12.875
33	2.875
34	60.25
35	29.25

Table 7.7

a. In words, X = _____.
b. i. $\overline{x}$ = _____
   ii. s_X = _____
   iii. n = _____

c. Construct a histogram of the distribution of the averages. Start at $x = -0.0005$. Use bar widths of 10.
d. In words, describe the distribution of the stock prices.
e. Randomly average five stock prices together. (Use a random number generator.) Continue averaging five prices together until you have 10 averages. List those 10 averages.
f. Use the 10 averages from part (e) to calculate the following:
   i. $\overline{x}$ = _____
   ii. s_x = _____
g. Construct a histogram of the distribution of the averages. Start at $x = -0.0005$. Use bar widths of 10.
h. Does this histogram look like the graph in Part (c)?
i. In one or two complete sentences, explain why the graphs either look the same or look different.
j. Based on the theory of the central limit theorem, $\overline{X}$ ~ _____(_____, ____).

Use the following information to answer the next three exercises: Richard's Furniture Company delivers furniture from 10 a.m. to 2 p.m. continuously and uniformly. We are interested in how long (in hours) past the 10 a.m. start time that individuals wait for their delivery.

78. X ~ _____(_____, _____)
a. $U(0, 4)$
b. $U(10, 2)$
c. $Exp(2)$
d. $N(2, 1)$

79. The average wait time is:
a. one hour
b. two hours
c. two and a half hours
d. four hours

80. Suppose that it is now past noon on a delivery day. The probability that a person must wait at least one and a half more hours is
a. $\frac{1}{4}$
b. $\frac{1}{2}$
c. $\frac{3}{4}$
d. $\frac{3}{8}$

Use the following information to answer the next two exercises: The time to wait for a particular rural bus is distributed uniformly from zero to 75 minutes. One hundred riders are randomly sampled to learn how long they waited.

81. The 90^{th} percentile sample average wait time (in minutes) for a sample of 100 riders is:
a. 315.0
b. 40.3
c. 38.5
d. 65.2

82. Would you be surprised, based on numerical calculations, if the sample average wait time (in minutes) for 100 riders was less than 30 minutes?
a. yes
b. no
c. There is not enough information.

Use the following to answer the next two exercises: The cost of unleaded gasoline in the Bay Area once followed an unknown distribution with a mean of \$4.59 and a standard deviation of \$0.10. Sixteen gas stations from the Bay Area are randomly chosen. We are interested in the average cost of gasoline for the 16 gas stations.

83. What's the approximate probability that the average price for 16 gas stations is more than \$4.69?
a. almost zero

b. 0.1587
c. 0.0943
d. unknown

84. Find the probability that the average price for 30 gas stations is less than $4.55.
a. 0.6554
b. 0.3446
c. 0.0142
d. 0.9858
e. 0

85. Suppose in a local kindergarten through 12th grade (K–12) school district, 53 percent of the population favor a charter school for grades K through five. A simple random sample of 300 is surveyed. Calculate the following using the normal approximation to the binomial distribtion.
a. Find the probability that less than 100 favor a charter school for grades K through 5.
b. Find the probability that 170 or more favor a charter school for grades K through 5.
c. Find the probability that no more than 140 favor a charter school for grades K through 5.
d. Find the probability that there are fewer than 130 that favor a charter school for grades K through 5.
e. Find the probability that exactly 150 favor a charter school for grades K through 5.

If you have access to an appropriate calculator or computer software, try calculating these probabilities using the technology.

86. Four friends, Janice, Barbara, Kathy, and Roberta, decided to carpool together to get to school. Each day the driver would be chosen by randomly selecting one of the four names. They carpool to school for 96 days. Use the normal approximation to the binomial to calculate the following probabilities. Round the standard deviation to four decimal places.
a. Find the probability that Janice is the driver at most 20 days.
b. Find the probability that Roberta is the driver more than 16 days.
c. Find the probability that Barbara drives exactly 24 of those 96 days.

87. $X \sim N(60, 9)$. Suppose that you form random samples of 25 from this distribution. Let $\overline{X}$ be the random variable of averages. Let ΣX be the random variable of sums. For parts (c) through (f), sketch the graph, shade the region, label and scale the horizontal axis for $\overline{X}$, and find the probability.
a. Sketch the distributions of X and $\overline{X}$ on the same graph.
b. $\overline{X} \sim$ _____(_____, _____)
c. $P(\overline{x} < 60) =$ _____
d. Find the 30th percentile for the mean.
e. $P(56 < \overline{x} < 62) =$ _____
f. $P(18 < \overline{x} < 58) =$ _____
g. $\Sigma x \sim$ _____(_____, _____)
h. Find the minimum value for the upper quartile for the sum.
i. $P(1400 < \Sigma x < 1550) =$ _____

88. Suppose that the length of research papers is uniformly distributed from 10 to 25 pages. We survey a class in which 55 research papers were turned in to a professor. The 55 research papers are considered a random collection of all papers. We are interested in the average length of the research papers.
a. In words, $X =$ _____.
b. $X \sim$ _____(_____, _____)
c. $\mu_X =$ _____
d. $\sigma_X =$ _____
e. In words, $\overline{X} =$ _____.
f. $\overline{X} \sim$ _____(_____, _____)
g. In words, $\Sigma X =$ _____.
h. $\Sigma X \sim$ _____(_____, _____)
i. Without doing any calculations, do you think that it's likely the professor will need to read a total of more than 1,050 pages? Why?
j. Calculate the probability that the professor will need to read a total of more than 1,050 pages.
k. Why is it so unlikely that the average length of the papers will be less than 12 pages?

89. Salaries for managers in a restaurant chain are normally distributed with a mean of $44,000 and a standard deviation of $6,500. We randomly survey 10 managers from that district.
   a. Find the 90th percentile for an individual manager's salary.
   b. Find the 90th percentile for the average manager's salary.

90. The average length of a maternity stay in a U.S. hospital is said to be 2.4 days with a standard deviation of 0.9 days. We randomly survey 80 women who recently bore children in a U.S. hospital.
   a. In words, X = _____.
   b. In words, $\overline{X}$ = _____.
   c. $\overline{X}$ ~ _____(_____, _____)
   d. In words, ΣX = _____.
   e. ΣX ~ _____(_____, _____)
   f. Is it likely that an individual stayed more than five days in the hospital? Why or why not?
   g. Is it likely that the average stay for the 80 women was more than five days? Why or why not?
   h. Which is more likely:
      i. An individual stayed more than five days.
      ii. The average stay of 80 women was more than five days.
   i. If we were to sum up the women's stays, is it likely that collectively, they spent more than a year in the hospital? Why or why not?

For each problem, wherever possible, provide graphs and use a calculator.

91. NeverReady batteries has engineered a newer, longer-lasting AAA battery. The company claims this battery has an average life span of 17 hours with a standard deviation of 0.8 hours. Your statistics class questions this claim. As a class, you randomly select 30 batteries and find that the sample mean life span is 16.7 hours. If the process is working properly, what is the probability of getting a random sample of 30 batteries in which the sample mean life span is 16.7 hours or less? Is the company's claim reasonable?

92. Men have an average weight of 172 pounds with a standard deviation of 29 pounds.
   a. Find the probability that 20 randomly selected men will have a sum weight greater than 3,600 pounds.
   b. If 20 men have a sum weight greater than 3,500 pounds, then their total weight exceeds the safety limits for water taxis. Based on (a), is this a safety concern? Explain.

93. Large bags of a brand of multicolored candies have a claimed net weight of 396.9 g. The standard deviation for the weight of the individual candies is 0.017 g. The following table is from a stats experiment conducted by a statistics class.

Red (g)	Orange (g)	Yellow (g)	Brown (g)	Blue (g)	Green (g)
0.751	0.735	0.883	0.696	0.881	0.925
0.841	0.895	0.769	0.876	0.863	0.914
0.856	0.865	0.859	0.855	0.775	0.881
0.799	0.864	0.784	0.806	0.854	0.865
0.966	0.852	0.824	0.840	0.810	0.865
0.859	0.866	0.858	0.868	0.858	1.015
0.857	0.859	0.848	0.859	0.818	0.876

Table 7.8

Red (g)	Orange (g)	Yellow (g)	Brown (g)	Blue (g)	Green (g)
0.942	0.838	0.851	0.982	0.868	0.809
0.873	0.863			0.803	0.865
0.809	0.888			0.932	0.848
0.890	0.925			0.842	0.940
0.878	0.793			0.832	0.833
0.905	0.977			0.807	0.845
	0.850			0.841	0.852
	0.830			0.932	0.778
	0.856			0.833	0.814
	0.842			0.881	0.791
	0.778			0.818	0.810
	0.786			0.864	0.881
	0.853			0.825	
	0.864			0.855	
	0.873			0.942	
	0.880			0.825	
	0.882			0.869	
	0.931			0.912	
				0.887	

Table 7.8

The bag contained 465 candies and the listed weights in the table came from randomly selected candies. Count the weights.

a. Find the mean sample weight and the standard deviation of the sample weights of candies in the table.
b. Find the sum of the sample weights in the table and the standard deviation of the sum of the weights.
c. If 465 candies are randomly selected, find the probability that their weights sum to at least 396.9 g.
d. Is the candy company's labeling accurate?

94. The Screw Right Company claims their $\frac{3}{4}$inch screws are within ±0.23 of the claimed mean diameter of 0.750 inches with a standard deviation of 0.115 inches. The following data were recorded.

0.757	0.723	0.754	0.737	0.757	0.741	0.722	0.741	0.743	0.742
0.740	0.758	0.724	0.739	0.736	0.735	0.760	0.750	0.759	0.754
0.744	0.758	0.765	0.756	0.738	0.742	0.758	0.757	0.724	0.757
0.744	0.738	0.763	0.756	0.760	0.768	0.761	0.742	0.734	0.754
0.758	0.735	0.740	0.743	0.737	0.737	0.725	0.761	0.758	0.756

Table 7.9

The screws were randomly selected from the local home repair store.

a. Find the mean diameter and standard deviation for the sample.
b. Find the probability that 50 randomly selected screws will be within the stated tolerance levels. Is the company's diameter claim plausible?

95. Your company has a contract to perform preventive maintenance on thousands of air conditioners in a large city. Based on service records from previous years, the time that a technician spends servicing a unit averages one hour with a standard deviation of one hour. In the coming week, your company will service a simple random sample of 70 units in the city. You plan to budget an average of 1.1 hours per technician to complete the work. Will this be enough time?

96. A typical adult has an average IQ score of 105 with a standard deviation of 20. If 20 randomly selected adults are given an IQ test, what is the probability that the sample mean scores will be between 85 and 125 points?

97. Certain coins have an average weight of 5.201 g with a standard deviation of 0.065 g. If a vending machine is designed to accept coins whose weights range from 5.111 g to 5.291 g, what is the expected number of rejected coins when 280 randomly selected coins are inserted into the machine?

References

7.1 The Central Limit Theorem for Sample Means (Averages)

Baran, D. (2010). 20 percent of Americans have never used email. *WebGuild*. Retrieved from http://www.webguild.org/20080519/20-percent-of-americans-have-never-used-email

The Flurry Blog. (2013). Retrieved from http://blog.flurry.com

U.S. Department of Agriculture. (n.d.). Retrieved from https://www.usda.gov/

7.2 The Central Limit Theorem for Sums (Optional)

Farago, P. (2012, Oct. 29). The truth about cats and dogs: Smartphone vs tablet usage differences. *Flurry Analytics Blog*. Retrieved from http://flurrymobile.tumblr.com/post/113379683050/the-truth-about-cats-and-dogs-smartphone-vs

7.3 Using the Central Limit Theorem

The Wall Street Journal. (n.d.). Retrieved from https://www.wsj.com/

Centers for Disease Control and Prevention. (2017, April 16). National health and nutrition examination survey. *National Center for Health Statistics*. Retrieved from http://www.cdc.gov/nchs/nhanes.htm

Solutions

1. mean = 4 hours, standard deviation = 1.2 hours, sample size = 16

3. a. Check student's solution.
b. 3.5, 4.25, 0.2441

5. The fact that the two distributions are different accounts for the different probabilities.

7. 0.3345

9. 7833.92

11. 0.0089

13. 7326.49

15. 77.45%

17. 0.4207

19. 3,888.5

21. 0.8186

23. 5

25. 0.9772

27. The sample size, n, gets larger.

29. 49

31. 26.00

33. 0.1587

35. 1000

37.
a. $U(24, 26)$, 25, 0.5774
b. $N(25, 0.0577)$
c. 0.0416

39. 0.0003

41. 25.07

43.
a. $N(2500, 5.7735)$
b. 0

45. 2507.40

47.
a. 10
b. $\frac{1}{10}$

49. $N\left(10, \frac{10}{8}\right)$

51. 0.7799

53. 1.69

55. 0.0072

57. 391.54

59. 405.51

61. a. X = amount of change students carry
b. $X \sim Exp(1/0.88)$ or approximately $X \sim Exp(1.1364)$
c. $\overline{X}$ = average amount of change carried by a sample of 25 students.
d. $\overline{X} \sim N(0.88, 0.176)$
e. 0.0819
f. 0.4276
g. The probability in part (e) represents the probability of an individual value. In part (f), the probability describes the mean of a sample of 25. Part (f) relies on the central limit theorem, so the distributions are different. Part (e) is exponential and part (f) is normal.

63. a. length of time for an individual to complete IRS form 1040, in hours
b. mean length of time for a sample of 36 taxpayers to complete IRS form 1040, in hours
c. $N\left(10.53, \frac{1}{3}\right)$
d. Yes, I would be surprised, because the probability is almost 0.
e. No, I would not be totally surprised because the probability is 0.2312.

65. a. the length of a song, in minutes, in the collection
b. $U(2, 3.5)$
c. the average length, in minutes, of the songs from a sample of five albums from the collection
d. $N(2.75, 0.0220)$
e. 2.74 minutes
f. 0.03 minutes

67. a. True. The mean of a sampling distribution of the means is approximately the mean of the data distribution.
b. True. According to the central limit theorem, the larger the sample, the closer the sampling distribution of the means becomes normal.
c. The standard deviation of the sampling distribution of the means will decrease, making it approximately the same as the standard deviation of X as the sample size increases.

69. a. X = the yearly income of someone in a Third World country
b. the average salary from samples of 1,000 residents of a Third World country
c. $\overline{X} \sim N\left(2{,}000, \frac{8{,}000}{\sqrt{1{,}000}}\right)$
d. Very wide differences in data values can have averages smaller than standard deviations.
e. The distribution of the sample mean will have higher probabilities closer to the population mean.
$P(2{,}000 < \overline{X} < 2{,}100) = 0.1537$
$P(2{,}100 < \overline{X} < 2{,}200) = 0.1317$

71. b

73. a. the total length of time for nine criminal trials
b. $N(189, 21)$
c. 0.0432
d. 162.09; 90 percent of the total nine trials of this type will last 162 days or more.

75. a. X = the salary of one elementary school teacher in the district
b. $X \sim N(44000, 6500)$
c. ΣX ~ sum of the salaries of 10 elementary school teachers in the sample
d. $\Sigma X \sim N(44{,}000, 20{,}554.80)$
e. 0.9742
f. \$52,330.09
g. 466,342.04
h. Sampling 70 teachers instead of 10 would cause the distribution to be more spread out. It would be a more symmetrical normal curve.

i. If every teacher received a $3,000 raise, the distribution of X would shift to the right by $3,000. In other words, it would have a mean of $47,000.

77. a. X = the closing stock prices for U.S. semiconductor manufacturers
b. i. $20.71, ii. $17.31, iii. 35
c.
d. exponential distribution, $X \sim Exp\left(\frac{1}{20.71}\right)$
e. Answers will vary.
f. i. $20.71, ii. $11.14
g. Answers will vary.
h. Answers will vary.
i. Answers will vary.
j. $N\left(20.71, \frac{17.31}{\sqrt{5}}\right)$

79. b

81. b

83. a

85. a. 0
b. 0.1123
c. 0.0162
d. 0.0003
e. 0.0268

87. a. Check student's solution.
b. $\overline{X} \sim N\left(60, \frac{9}{\sqrt{25}}\right)$
c. 0.5000
d. 59.06
e. 0.8536
f. 0.1333
g. $N(1500, 45)$
h. 1530.35
i. 0.6877

89. a. $52,330
b. $46,634

91.
- We have $\mu = 17$, $\sigma = 0.8$, $\overline{x} = 16.7$, and $n = 30$. To calculate the probability, we use `normalcdf`(lower, upper, μ, $\frac{\sigma}{\sqrt{n}}$) = `normalcdf`$\left(E-99, 16.7, 17, \frac{0.8}{\sqrt{30}}\right)$ = 0.0200.
- If the process is working properly, then the probability that a sample of 30 batteries would have at most 16.7 life span hours is only 2%. Therefore, the class was justified to question the claim.

93. a. For the sample, we have $n = 100$, $\overline{x} = 0.862$, and $s = 0.05$.
b. $\Sigma\overline{x} = 85.65$, $\Sigma s = 5.18$
c. `normalcdf`$(396.9, E99, (465)(0.8565), (0.05)(\sqrt{465})) \approx 1$
d. Because the probability of a sample of size of 465 having at least a mean sum of 396.9 is appproximately 1, we can conclude that the company is correctly labeling their candy packages.

95. Use $\texttt{normalcdf}\left(E-99,1.1,1,\frac{1}{\sqrt{70}}\right)$ = 0.7986. This means that there is an 80 percent chance that the service time will be less than 1.1 hours. It may be wise to schedule more time because there is an associated 20 percent chance that the maintenance time will be greater than 1.1 hours.

97. Because we have $\texttt{normalcdf}\left(5.111,5.291,5.201,\frac{0.065}{\sqrt{280}}\right) \approx 1$, we can conclude that practically all the coins are within the limits; therefore, there should be no rejected coins out of a well-selected sample size of 280.

8 CONFIDENCE INTERVALS

Figure 8.1 Have you ever wondered what the average number of chocolate candies in a bag at the grocery store is? You can use confidence intervals to answer this question. (credit: comedy_nose/flickr)

Chapter Objectives

By the end of this chapter, the student should be able to do the following:

- Calculate and interpret confidence intervals for estimating a population mean and a population proportion
- Interpret the Student's *t* probability distribution as the sample size changes
- Discriminate between problems applying the normal and the Student's *t*-distributions
- Calculate the sample size required to estimate a population mean and a population proportion, given a desired confidence level and margin of error

Introduction

Suppose you were trying to determine the mean rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. You would have obtained a point estimate of the true mean. If you are trying to determine the percentage of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempt. In this case, you would have obtained a point estimate for the true proportion.

We use sample data to make generalizations about an unknown population. This part of statistics is called **inferential statistics**. *The sample data help us to make an estimate of a population* ***parameter***. We realize that the point estimate is most likely not the exact value of the population parameter, but close to it. After calculating point estimates, we construct interval estimates, called *confidence intervals*.

In this chapter, you will learn to construct and interpret confidence intervals. You will also learn a new distribution, the Student's-*t*, and how it is used with those intervals. Throughout the chapter, it is important to keep in mind that the confidence interval is a random variable. It is the population parameter that is fixed.

If you worked in the marketing department of an entertainment company, you might be interested in the mean number of songs a consumer downloads a month from an internet music store. If so, you could conduct a survey and calculate the sample mean, $\overline{x}$, and the sample standard deviation, s. You would use $\overline{x}$ to estimate the population mean and s to estimate the population standard deviation. The sample mean, $\overline{x}$, is the **point estimate** for the population mean, μ. The sample standard deviation, s, is the point estimate for the population standard deviation, σ.

Each instance of $\overline{x}$ and s is called a *statistic*.

A **confidence interval** is another type of estimate but, instead of being just one number, it is an interval of numbers. The interval of numbers is a range of values calculated from a given set of sample data. The confidence interval is likely to include an unknown population parameter.

Suppose, for the internet music example, we do not know the population mean, μ, but we do know that the population standard deviation is $\sigma = 1$ and our sample size is 100. Then, by the central limit theorem, the standard deviation for the sample mean is

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1.$$

The **Empirical Rule**, which applies to bell-shaped distributions, says that in approximately 95 percent of the samples, the sample mean, $\overline{x}$, will be within two standard deviations of the population mean, μ. For our internet music example, two standard deviations would be calculated as (2)(0.1) = 0.2. The sample mean, $\overline{x}$, is likely to be within 0.2 units of μ.

In this example, we do not know the true population mean μ (because we do not have information from all the internet music users!), but we can compute the sample mean $\overline{x}$ based on our sample of 100 individuals. Because the sample mean is likely to be within 0.2 units of the true population mean 95 percent of the times that we take a sample of 100 users, we can say with 95 percent confidence that μ is within 0.2 units of $\overline{x}$. In other words, μ is somewhere between $\overline{x} - 0.2$ and $\overline{x} + 0.2$.

Suppose that from the sample of 100 internet music customers, we compute a sample mean download of $\overline{x} = 2$ songs per month. Since we know that the population standard deviation is $\sigma = 1$, according to the central limit theorem, the standard deviation for the sample means is $\sigma = \frac{1}{\sqrt{100}} = 0.1$.

We know that there is a 95 percent chance that the true population mean value μ is between two standard deviations from the sample mean. That is, with 95 percent confidence we can say that μ is between $\overline{x} + 2 \times \frac{\sigma}{\sqrt{n}}$ and $\overline{x} - 2 \times \frac{\sigma}{\sqrt{n}}$.

Replacing the symbols for their values in this example, we say that we are **95 percent confident** that the true average number of songs downloaded from an internet music store per month is between

$$\overline{x} - 2 \times \frac{\sigma}{\sqrt{n}} = 2 - 2 \times \frac{\sigma}{\sqrt{100}} = 2 - 0.2 = 1.8, \text{ and}$$

$$\overline{x} + 2 \times \frac{\sigma}{\sqrt{n}} = 2 + 2 \times \frac{\sigma}{\sqrt{100}} = 2 + 0.2 = 2.2. \quad \textbf{8.1}$$

The 95 percent confidence interval for μ is (1.8, 2.2).

The 95 percent confidence interval implies two possibilities. Either the interval (1.8, 2.2) contains the true mean, μ, or our sample produced an $\overline{x}$ that is not within 0.2 units of the true mean μ. The second possibility happens for only 5 percent of all the samples (95–100 percent).

Remember that a confidence interval is created for an unknown population parameter like the population mean, μ. Confidence intervals for some parameters have the form

(point estimate – margin of error, point estimate + margin of error).

The margin of error depends on the confidence level or percentage of confidence and the standard error of the mean.

When you read newspapers and journals, you might notice that some reports use the phrase *margin of error*. Other reports will not use that phrase, but include a confidence interval as the point estimate plus or minus the margin of error. Those are two ways of expressing the same concept.

Note

Although the text covers only symmetrical confidence intervals, there are non-symmetrical confidence intervals (for example, a confidence interval for the standard deviation).

COLLABORATIVE EXERCISE

Have your instructor record the number of meals each student in your class eats out in a week. Assume that the standard deviation is known to be three meals. Construct an approximate 95 percent confidence interval for the true mean number of meals students eat out each week.

1. Calculate the sample mean.
2. Let $\sigma = 3$ and n = the number of students surveyed.
3. Construct the interval. $\left(\bar{x} - 2\left(\frac{\sigma}{\sqrt{n}}\right)\right), \left(\bar{x} + 2\left(\frac{\sigma}{\sqrt{n}}\right)\right)$

We say we are approximately 95 percent confident that the true mean number of meals that students eat out in a week is between _____ and _____.

8.1 A Single Population Mean Using the Normal Distribution

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x} = 10$ and we have constructed the 90 percent confidence interval (5, 15), where the margin of error = 5.

Calculating the Confidence Interval

To construct a confidence interval for a single unknown population mean, μ, *where the population standard deviation is known*, we need $\bar{x}$ as an estimate for μ, and we need the margin of error. The margin of error for the population mean is called the **error bound for a population mean** (*EBM*). The sample mean, $\bar{x}$, is the **point estimate** of the unknown population mean, μ.

The confidence interval (*CI*) estimate will have the form:

(point estimate – error bound, point estimate + error bound) or, in symbols, $(\bar{x} - EBM, \bar{x} + EBM)$.

The margin of error (*EBM*) depends on the **confidence level** (***CL***). The confidence level is often considered the probability that the calculated confidence interval estimate will contain the true population parameter. However, it is more accurate to state that the confidence level is the percentage of confidence intervals that contain the true population parameter when repeated samples are taken. Most often, the person constructing the confidence interval will choose a confidence level of 90 percent or higher, because that person wants to be reasonably certain of his or her conclusions.

Another probability, which is called alpha (α) is related to the confidence level, *CL*. Alpha is the probability that the confidence interval does not contain the unknown population parameter. Mathematically, alpha can be computed as $\alpha = 1 - CL$.

EXAMPLE 8.1

Suppose we have collected data from a sample. We know the sample mean, but we do not know the mean for the entire population.
The sample mean is seven, and the error bound for the mean is 2.5.

$$\bar{x} \text{ and } EBM = 2.5.$$

The confidence interval is (7 – 2.5, 7 + 2.5), and calculating the values gives (4.5, 9.5).

If the confidence level is 95 percent, then we say, "We estimate with 95 percent confidence that the true value of the population mean is between 4.5 and 9.5."

TRY IT 8.1

Suppose we have data from a sample. The sample mean is 15, and the error bound for the mean is 3.2.

What is the confidence interval estimate for the population mean?

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\overline{x} = 10$, and we have constructed the 90 percent confidence interval (5, 15) where $EBM = 5$.

To get a 90 percent confidence interval, we must include the central 90 percent of the probability of the normal distribution. If we include the central 90 percent, we leave out a total of $\alpha = 10$ percent in both tails, or 5 percent in each tail, of the normal distribution.

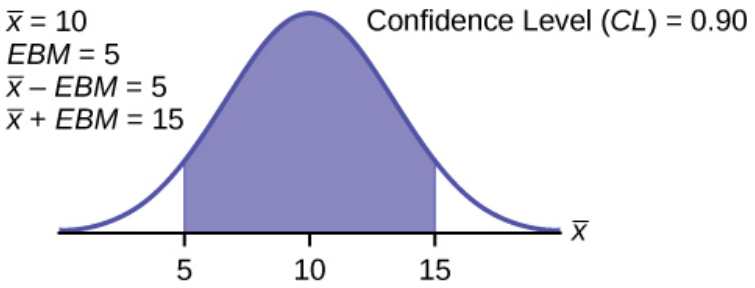


Figure 8.2

The critical value *1.645* is the ***z**-score* in a standard normal probability distribution that puts an area of *0.90* in the center, an area of *0.05* in the far left tail, and an area of *0.05* in the far right tail. To capture the central 90 percent, we must go out *1.645 standard deviations* on either side of the calculated sample mean. The critical value will change depending on the confidence level of the interval.

It is important that the *standard deviation* used be appropriate for the parameter we are estimating, so in this section, we need to use the standard deviation that applies to sample means, which is $\frac{\sigma}{\sqrt{n}}$. The fraction $\frac{\sigma}{\sqrt{n}}$ is commonly called the *standard error of the mean* in order to distinguish clearly the standard deviation for a mean from the population standard deviation, σ.

In summary, as a result of the central limit theorem, the following statements apply:

- $\overline{X}$ is normally distributed, that is, $\overline{X} \sim N\left(\mu_X, \frac{\sigma}{\sqrt{n}}\right)$.
- **When the population standard deviation σ is known, we use a normal distribution to calculate the error bound.**

Calculating the Confidence Interval

To construct a confidence interval estimate for an unknown population mean, we need data from a random sample. The steps to construct and interpret the confidence interval are as follows:

- Calculate the sample mean, $\overline{x}$, from the sample data. Remember, in this section, we already know the population standard deviation, σ.
- Find the z-score that corresponds to the confidence level.
- Calculate the error bound EBM.
- Construct the confidence interval.
- If we denote the critical z-score by $z_{\frac{\alpha}{2}}$, and the sample size by n, then the formula for the confidence interval with confidence level $Cl = 1 - \alpha$, is given by $(\overline{x} - z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}})$.
- Write a sentence that interprets the estimate in the context of the situation in the problem. (Explain what the confidence interval means, in the words of the problem.)

We will first examine each step in more detail and then illustrate the process with some examples.

Finding the *z*-Score for the Stated Confidence Level

When we know the population standard deviation, σ, we use a standard normal distribution to calculate the error bound *EBM* and construct the confidence interval. We need to find the value of z that puts an area equal to the confidence level (in decimal form) in the middle of the standard normal distribution $Z \sim N(0, 1)$.

The confidence level, *CL*, is the area in the middle of the standard normal distribution. $CL = 1 - \alpha$, so α is the area that is split equally between the two tails. Each of the tails contains an area equal to $\frac{\alpha}{2}$.

The z-score that has an area to the right of $\frac{\alpha}{2}$ is denoted by $z_{\frac{\alpha}{2}}$.

For example, when $CL = 0.95$, $\alpha = 0.05$, and $\frac{\alpha}{2} = 0.025$, we write $z_{\frac{\alpha}{2}} = z_{0.025}$.

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is 1 – 0.025 = 0.975.

$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$, using a calculator, computer, or standard normal probability table.

Normal table (see appendices) shows that the probability for 0 to 1.96 is 0.47500, and so the probability to the right tail of the critical value 1.96 is 0.5 – 0.475 = 0.025

USING THE TI-83, 83+, 84, 84+ CALCULATOR

`invNorm`(0.975, 0, 1) = 1.96. In this command, the value 0.975 is the total area to the left of the critical value that we are looking to calculate. The parameters 0 and 1 are the mean value and the standard deviation of the standard normal distribution Z.

Note

Remember to use the area to the LEFT of $z_{\frac{\alpha}{2}}$. In this chapter, the last two inputs in the invNorm command are 0, 1, because you are using a standard normal distribution Z with mean 0 and standard deviation 1.

Calculating the Margin of Error *EBM*

The error bound formula for an unknown population mean, μ, when the population standard deviation, σ, is known is

Margin of error = $\left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$.

Constructing the Confidence Interval

The confidence interval estimate has the format sample mean plus or minus the margin of error.

The graph gives a picture of the entire situation

$CL + \frac{\alpha}{2} + \frac{\alpha}{2} = CL + \alpha = 1$.

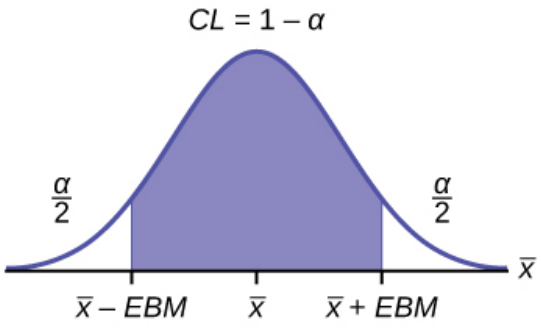


Figure 8.3

Writing the Interpretation

The interpretation should clearly state the confidence level (*CL*), explain which population parameter is being estimated (here, a **population mean**), and state the confidence interval (both endpoints): "We estimate with ___percent confidence that the true population mean (include the context of the problem) is between ___ and ___ (include appropriate units)."

EXAMPLE 8.2

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of three points. A random sample of 36 scores is taken and gives a sample mean (sample mean score) of 68. Find a confidence interval estimate for the population mean exam score (the mean score on all exams).

Problem

Find a 90 percent confidence interval for the true (population) mean of statistics exam scores.

Solution

- You can use technology to calculate the confidence interval directly.
- The first solution is shown step-by-step (Solution A).
- The second solution uses the TI-83, 83+, and 84+ calculators (Solution B).

Solution A

To find the confidence interval, you need the sample mean, $\overline{x}$, and the *EBM*.

$$\overline{x} = 68 \tag{8.2}$$

$$EBM = (z_{\frac{\alpha}{2}})\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$\sigma = 3;\ n = 36;$$

The confidence level is 90 percent (*CL* = 0.90).

$$CL = 0.90, \text{ so } \alpha = 1 - CL = 1 - 0.90 = 0.10.$$

$$\frac{\alpha}{2} = 0.05, z_{\frac{\alpha}{2}} = z_{0.05}$$

The area to the right of $z_{0.05}$ is 0.05 and the area to the left of $z_{0.05}$ is 1 – 0.05 = 0.95.

$$z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$$

using invNorm(0.95, 0, 1) on the TI-83,83+, and 84+ calculators. This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.

EBM = (1.645)$\left(\frac{3}{\sqrt{36}}\right)$ = 0.8225

$\overline{x}$ – *EBM* = 68 – 0.8225 = 67.1775

$\overline{x}$ + *EBM* = 68 + 0.8225 = 68.8225

The 90 percent confidence interval is **(67.1775, 68.8225).**

Solution

Solution B

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Press STAT and arrow over to TESTS.
Arrow down to `7:ZInterval`.
Press ENTER.
Arrow to `Stats` and press ENTER.
Arrow down and enter 3 for σ, 68 for $\overline{x}$, 36 for n, and .90 for `C-level`.
Arrow down to `Calculate` and press ENTER.

The confidence interval is (to three decimal places)(67.178, 68.822).

Interpretation

We estimate with 90 percent confidence that the true population mean exam score for all statistics students is between 67.18 and 68.82.

Explanation of 90 percent Confidence Level

Ninety percent of all confidence intervals constructed in this way contain the true mean statistics exam score. For example, if we constructed 100 of these confidence intervals, we would expect 90 of them to contain the true population mean exam score.

TRY IT 8.2

Suppose average pizza delivery times are normally distributed with an unknown population mean and a population standard deviation of 6 minutes. A random sample of 28 pizza delivery restaurants is taken and has a sample mean delivery time of 36 min.

Find a 90 percent confidence interval estimate for the population mean delivery time.

EXAMPLE 8.3

The specific absorption rate (SAR) for a cell phone measures the amount of radio frequency (RF) energy absorbed by the user's body when using the handset. Every cell phone emits RF energy. Different phone models have different SAR measures. For certification from the Federal Communications Commission for sale in the United States, the SAR level for a cell phone must be no more than 1.6 watts per kilogram. Table 8.1 shows the highest SAR level for a random selection of cell phone models of a random cell phone company.

Phone Model #	SAR	Phone Model #	SAR	Phone Model #	SAR
800	1.11	1800	1.36	2800	0.74
900	1.48	1900	1.34	2900	0.5
1000	1.43	2000	1.18	3000	0.4
1100	1.3	2100	1.3	3100	0.867
1200	1.09	2200	1.26	3200	0.68
1300	0.455	2300	1.29	3300	0.51
1400	1.41	2400	0.36	3400	1.13
1500	0.82	2500	0.52	3500	0.3
1600	0.78	2600	1.6	3600	1.48
1700	1.25	2700	1.39	3700	1.38

Table 8.1

Problem

Find a 98 percent confidence interval for the true (population) mean of the SARs for cell phones. Assume that the population standard deviation is $\sigma = 0.337$.

Solution

Solution A

To find the confidence interval, start by finding the point estimate: the sample mean,

$$\overline{x} = 1.024.$$

This is calculated by adding the specific absorption rate for the 30 cell phones in the sample, and dividing the result by 30.

Next, find the *EBM*. Because you are creating a 98 percent confidence interval, $CL = 0.98$.

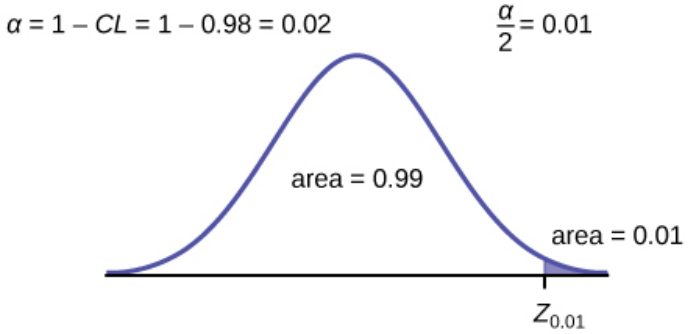


Figure 8.4

You need to find $z_{0.01}$, having the property that the area under the normal density curve to the right of $z_{0.01}$ is 0.01 and the area to the left is 0.99. Use your calculator, a computer, or a probability table for the standard normal distribution to find $z_{0.01} = 2.326$.

$$EBM = (z_{0.01})\frac{\sigma}{\sqrt{n}} = (2.326)\frac{0.337}{\sqrt{30}} = 0.1431$$

To find the 98 percent confidence interval, find $\overline{x} \pm EBM$.

$$\overline{x} - EBM = 1.024 - 0.1431 = 0.8809$$
$$\overline{x} + EBM = 1.024 + 0.1431 = 1.1671$$

We estimate with 98 percent confidence that the true SAR mean for the population of cell phones in the United States is between 0.8809 and 1.1671 watts per kilogram.

Solution

Solution B

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Press STAT and arrow over to TESTS.
Arrow down to 7:ZInterval.
Press ENTER.
Arrow to Stats and press ENTER.
Arrow down and enter the following values:
σ: 0.337
$\overline{x}$: 1.024
n: 30
C-level: 0.98

Arrow down to Calculate and press ENTER.

The confidence interval is (to three decimal places) (0.881, 1.167).

TRY IT 8.3

Table 8.2 shows a different random sampling of 20 cell phone models. Use these data to calculate a 93 percent confidence interval for the true mean SAR for cell phones certified for use in the United States. As previously, assume that the population standard deviation is $\sigma = 0.337$.

Phone Model	SAR	Phone Model	SAR
450	1.48	1450	1.53
550	0.8	1550	0.68
650	1.15	1650	1.4
750	1.36	1750	1.24
850	0.77	1850	0.57
950	0.462	1950	0.2
1050	1.36	2050	0.51
1150	1.39	2150	0.3
1250	1.3	2250	0.73
1350	0.7	2350	0.869

Table 8.2

Notice the difference in the confidence intervals calculated in Example 8.3 and the following Try It exercise. These intervals are different for several reasons: they are calculated from different samples, the samples are different sizes, and the intervals are calculated for different levels of confidence. Even though the intervals are different, they do not yield conflicting information. The effects of these kinds of changes are the subject of the next section in this chapter.

Changing the Confidence Level or Sample Size

EXAMPLE 8.4

Problem

Suppose we change the original problem in Example 8.2 by using a 95 percent confidence level. Find a 95 percent confidence interval for the true (population) mean statistics exam score.

Solution

To find the confidence interval, you need the sample mean, $\overline{x}$, and the *EBM*.

$$\overline{x} = 68$$

$$EBM=(z_{\frac{\alpha}{2}})(\frac{\sigma}{\sqrt{n}})$$

$$\sigma= 3;\ n= 36$$

The confidence level is 95 percent (*CL* = 0.95).

$$CL = 0.95, \text{ so } \alpha = 1 - CL = 1 - 0.95 = 0.05.$$

$$\frac{\alpha}{2} = 0.025 \qquad z_{\frac{\alpha}{2}} = z_{0.025}$$

The area to the right of $z_{0.025}$ is 0.025, and the area to the left of $z_{0.025}$ is 1 – 0.025 = 0.975.

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96,$$

when using invnorm(0.975,0,1) on the TI-83, 83+, or 84+ calculators. (This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.)

$$EBM = (1.96)\left(\frac{3}{\sqrt{36}}\right) = 0.98$$

$$\overline{x} - EBM = 68 - 0.98 = 67.02$$

$$\overline{x} + EBM = 68 + 0.98 = 68.98$$

Notice that the *EBM* is larger for a 95 percent confidence level in the original problem.

Interpretation

We estimate with 95 percent confidence that the true population mean for all statistics exam scores is between 67.02 and 68.98.

Explanation of 95 percent Confidence Level

95 percent of all confidence intervals constructed in this way contain the true value of the population mean statistics exam score.

Comparing the Results

The 90 percent confidence interval is (67.18, 68.82). The 95 percent confidence interval is (67.02, 68.98). The 95 percent confidence interval is wider. If you look at the graphs, because the area 0.95 is larger than the area 0.90, it makes sense that the 95 percent confidence interval is wider. For more certainty that the confidence interval actually does contain the true value of the population mean for all statistics exam scores, the confidence interval necessarily needs to be wider.

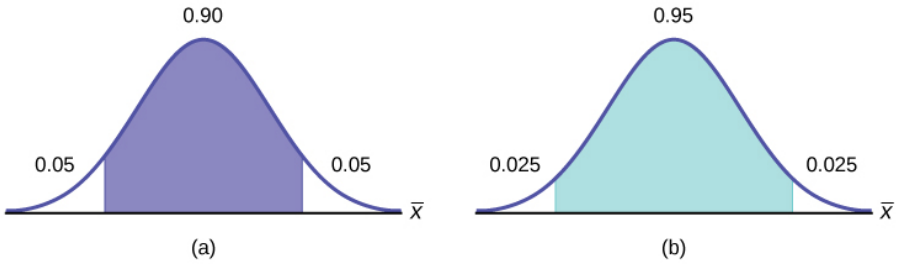


Figure 8.5

Summary: Effect of Changing the Confidence Level

- Increasing the confidence level increases the error bound, making the confidence interval wider.
- Decreasing the confidence level decreases the error bound, making the confidence interval narrower.

TRY IT 8.4

Refer back to the pizza-delivery Try It exercise. The population standard deviation is six minutes and the sample mean deliver time is 36 minutes. Use a sample size of 20. Find a 95 percent confidence interval estimate for the true mean pizza-delivery time.

EXAMPLE 8.5

Suppose we change the original problem in Example 8.2 to see what happens to the error bound if the sample size is changed.

Problem

Leave everything the same except the sample size. Use the original 90 percent confidence level. What happens to the error bound and the confidence interval if we increase the sample size and use $n = 100$ instead of $n = 36$? What happens if we decrease the sample size to $n = 25$ instead of $n = 36$?

- $\overline{x} = 68$
- $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$
- $\sigma = 3$, the confidence level is 90 percent ($CL = 0.90$), $z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$.

Solution

Solution A

If we *increase* the sample size n to 100, we *decrease* the margin of error.

When $n = 100$, $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right) = (1.645)\left(\frac{3}{\sqrt{100}}\right) = 0.4935$.

Solution

Solution B

If we *decrease* the sample size n to 25, we *increase* the error bound.

When $n = 25$, $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right) = (1.645)\left(\frac{3}{\sqrt{25}}\right) = 0.987$.

Summary: Effect of Changing the Sample Size

- Increasing the sample size causes the error bound to decrease, making the confidence interval narrower.
- Decreasing the sample size causes the error bound to increase, making the confidence interval wider.

TRY IT 8.5

Refer back to the pizza-delivery Try It exercise. The mean delivery time is 36 minutes and the population standard deviation is six minutes. Assume the sample size is changed to 50 restaurants with the same sample mean. Find a 90 percent confidence interval estimate for the population mean delivery time.

Working Backward to Find the Error Bound or Sample Mean

When we calculate a confidence interval, we find the sample mean, calculate the error bound, and use them to calculate the confidence interval. However, sometimes when we read statistical studies, the study may state the confidence interval only. If we know the confidence interval, we can work backward to find both the error bound and the sample mean.

Finding the Error Bound

- From the upper value for the interval, subtract the sample mean,
- Or, from the upper value for the interval, subtract the lower value. Then divide the difference by 2.

Finding the Sample Mean

- Subtract the error bound from the upper value of the confidence interval,
- Or, average the upper and lower endpoints of the confidence interval.

Notice that there are two methods to perform each calculation. You can choose the method that is easier to use with the information you know.

EXAMPLE 8.6

Suppose we know that a confidence interval is (67.18, 68.82) and we want to find the error bound. We may know that the sample mean is 68, or perhaps our source only gives the confidence interval and does not tell us the value of the sample mean.

Calculate the error bound:

- If we know that the sample mean is 68, $EBM = 68.82 - 68 = 0.82$.
- If we do not know the sample mean, $EBM = \frac{(68.82-67.18)}{2} = 0.82$. The margin of error is the quantity that we add and subtract from the sample mean to obtain the confidence interval. Therefore, the margin of error is half of the length of the interval.

Calculate the sample mean:

- If we know the error bound, $\overline{x} = 68.82 - 0.82 = 68$.
- If we do not know the error bound, $\overline{x} = \frac{(67.18+68.82)}{2} = 68$.

TRY IT 8.6

Suppose we know that a confidence interval is (42.12, 47.88). Find the error bound and the sample mean.

Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size. In this situation, we are given the desired margin of error, *EBM*, and we need to compute the sample size *n*.

The formula for sample size is $n = \frac{z^2\sigma^2}{EBM^2}$, found by solving the error bound formula for *n*. Always round up the value of *n* to the closest integer.

In this formula, *z* is the critical value $z_{\frac{\alpha}{2}}$, corresponding to the desired confidence level. A researcher planning a study who wants a specified confidence level and error bound can use this formula to calculate the size of the sample needed for the study.

EXAMPLE 8.7

The population standard deviation for the age of Foothill College students is 15 years. If we want to be 95 percent confident that the sample mean age is within two years of the true population mean age of Foothill College students, how many randomly selected Foothill College students must be surveyed?

From the problem, we know that $\sigma = 15$ and $EBM = 2$.
$z = z_{0.025} = 1.96$, because the confidence level is 95 percent.

$n = \frac{z^2\sigma^2}{EBM^2} = \frac{(1.96)^2(15)^2}{2^2} = 216.09$ using the sample size equation.
Use $n = 217$. Always round the answer up to the next higher integer to ensure that the sample size is large enough.

Therefore, 217 Foothill College students should be surveyed in order to be 95 percent confident that we are within two years of the true population mean age of Foothill College students.

TRY IT 8.7

The population standard deviation for the height of high school basketball players is three inches. If we want to be 95 percent confident that the sample mean height is within one inch of the true population mean height, how many randomly selected students must be surveyed?

8.2 A Single Population Mean Using the Student's t-Distribution

In practice, we rarely know the **population standard deviation**. In the past, when the sample size was large, this unknown number did not present a problem to statisticians. They used the sample standard deviation s as an estimate for σ and proceeded as before to calculate a **confidence interval** with close-enough results. However, statisticians ran into problems when the sample size was small. A small sample size caused inaccuracies in the confidence interval.

William S. Gosset (1876–1937) of the Guinness brewery in Dublin, Ireland, ran into this problem. His experiments with hops and barley produced very few samples. Just replacing σ with s did not produce accurate results when he tried to calculate a confidence interval. He realized that he could not use a normal distribution for the calculation; he found that the actual distribution depends on the sample size. This problem led him to *discover* what is called the **Student's *t*-distribution**. The name comes from the fact that Gosset wrote under the pen name *Student*.

Up until the mid-1970s, some statisticians used the **normal distribution** approximation for large sample sizes and used the Student's t-distribution only for sample sizes of at most 30. With graphing calculators and computers, the practice now is to use the Student's t-distribution whenever s is used as an estimate for σ.

If you draw a simple random sample of size n from a population that has an approximately normal distribution with mean μ and unknown population standard deviation σ and calculate the t-score $t = \frac{\overline{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)}$, then the t-scores follow a *Student's t-distribution with n – 1 degrees of freedom*. The t-score has the same interpretation as the ***z*-score**: It measures how far $\overline{x}$ is from its mean μ. For each sample size n, there is a different Student's t-distribution.

The **degrees of freedom (*df*)**, n -– 1, are the sample size minus 1.

Properties of the Student's *t*-distribution

- The graph for the Student's t-distribution is similar to the standard normal curve.
- The mean for the Student's t-distribution is zero, and the distribution is symmetric about zero.
- The Student's t-distribution has more probability in its tails than the standard normal distribution. Figure 8.6 shows the graphs of the student t-distribution for 1, 2 and 5 degrees of freedom: (v), compare to the standard normal distribution (in black).

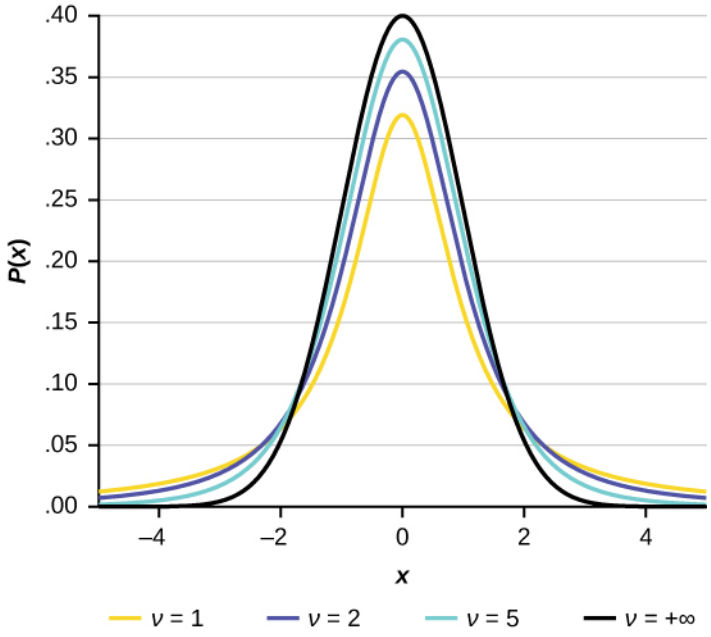


Figure 8.6

- The exact shape of the Student's t-distribution depends on the degrees of freedom. As the degrees of freedom increase, the graph of the Student's t-distribution becomes more like the graph of the standard normal distribution.
- The underlying population of individual observations is assumed to be normally distributed with unknown population mean μ and unknown population standard deviation σ. The size of the underlying population is generally not relevant unless it is very small. If it is bell-shaped (normal), then the assumption is met and does not need discussion. Random sampling is assumed, but that is a completely separate assumption from normality.

Calculators and computers can easily calculate any Student's *t*-probabilities. The TI-83, 83+, and 84+ have a tcdf function to find the probability for given values of *t*. The grammar for the tcdf command is tcdf(lower bound, upper bound, degrees of freedom). However, for confidence intervals, we need to use inverse probability to find the value of *t* when we know the probability.

For the TI-84+, you can use the invT command on the DISTRibution menu. The invT command works similarly to the invnorm. The invT command requires two inputs: invT(area to the left, degrees of freedom). The output is the *t*-score that corresponds to the area we specified.

The TI-83 and 83+ do not have the invT command. (The TI-89 has an inverse T command.)

A probability table for the Student's *t*-distribution can also be used. The table gives critical *t*-values that correspond to the confidence level (column) and degrees of freedom (row). (The TI-86 does not have an invT program or command, so if you are using that calculator, you need to use a probability table for the Student's *t*-distribution.) When using a *t*-table, note that some tables are formatted to show the confidence level in the column headings, while the column headings in some tables may show only corresponding area in one or both tails.

A Student's *t*-table (see Appendix H Tables) gives *t*-scores given the degrees of freedom and the right-tailed probability. The table is very limited. *Calculators and computers can easily calculate any Student's t-probabilities.*

If the population standard deviation is not known, the error bound for a population mean is

- $EBM = \left(t_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right)$,
- $t_{\frac{\alpha}{2}}$ is the *t*-score with area to the right equal to $\frac{\alpha}{2}$,
- use $df = n - 1$ degrees of freedom, and
- s = sample standard deviation.

The format for the confidence interval is

$$(\overline{x} - EBM, \overline{x} + EBM). \quad \textbf{8.3}$$

USING THE TI-83, 83+, 84, 84+ CALCULATOR

To calculate the confidence interval directly, do the following:
Press STAT.
Arrow over to TESTS.
Arrow down to `8:TInterval` and press ENTER (or just press 8).

EXAMPLE 8.8

Problem

Suppose you do a study of acupuncture to determine how effective it is in relieving pain. You measure sensory rates for 15 subjects with the results given. Use the sample data to construct a 95 percent confidence interval for the mean sensory rate for the population (assumed normal) from which you took the data.
The solution is shown step-by-step and by using the TI-83, 83+, or 84+ calculators.

8.6; 9.4; 7.9; 6.8; 8.3; 7.3; 9.2; 9.6; 8.7; 11.4; 10.3; 5.4; 8.1; 5.5; 6.9

Solution

- The first solution is step-by-step (Solution A).
- The second solution uses the TI-83+ and TI-84 calculators (Solution B).

To find the confidence interval, you need the sample mean, $\overline{x}$, and the *EBM*.

$$\overline{x} = \frac{8.6 + 9.4 + 7.9 + 6.8 + 8.3 + 7.3 + 9.2 + 9.6 + 8.7 + 11.4 + 10.3 + 5.4 + 8.1 + 5.5 + 6.9}{15} = 8.2267;$$

$$s = \sqrt{\frac{(8.6-\overline{x})^2 + (9.4 - \overline{x})^2 + \cdots + (5.5 - \overline{x})^2 + (6.9 - \overline{x})^2}{14}} = 1.6722;$$

$n = 15$

$df = 15 - 1 = 14$ CL, so $\alpha = 1 - CL = 1 - 0.95 = 0.05$

$\frac{\alpha}{2} = 0.025$; $t_{\frac{\alpha}{2}} = t_{0.025}$

The area to the right of $t_{0.025}$ is 0.025, and the area to the left of $t_{0.025}$ is 1 – 0.025 = 0.975.

$t_{\frac{\alpha}{2}} = t_{0.025} = 2.14$ using invT(.975,14) on the TI-84+ calculator.

$$EBM = \left(t_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right) \quad \textbf{8.4}$$

$$EBM = (2.14)\left(\frac{1.6722}{\sqrt{15}}\right) = 0.924$$

$$\overline{x} - EBM = 8.2267 - 0.9240 = 7.3$$

$$\overline{x} + EBM = 8.2267 + 0.9240 = 9.15$$

The 95 percent confidence interval is (7.30, 9.15).

We estimate with 95 percent confidence that the true population mean sensory rate is between 7.30 and 9.15.

Solution

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Press STAT and arrow over to TESTS.
Arrow down to `8:TInterval` and press ENTER (or you can just press 8).
Arrow to `Data` and press ENTER.
Arrow down to `List` and enter the list name where you put the data.
There should be a 1 after `Freq`.
Arrow down to `C-level` and enter 0.95.
Arrow down to `Calculate` and press ENTER.
The 95 percent confidence interval is (7.3006, 9.1527).

Note

When calculating the error bound, you can also use a probability table for the Student's *t*-distribution to find the value of *t*. The table gives *t*-scores that correspond to the confidence level (column) and degrees of freedom (row); the *t*-score is found where the row and column intersect in the table.

TRY IT 8.8

You do a study of hypnotherapy to determine how effective it is in increasing the number of hours of sleep subjects get each night. You measure hours of sleep for 12 subjects with the following results. Construct a 95 percent confidence interval for the mean number of hours slept for the population (assumed normal) from which you took

the data.

8.2, 9.1, 7.7, 8.6, 6.9, 11.2, 10.1, 9.9, 8.9, 9.2, 7.5, 10.5

EXAMPLE 8.9

Problem

A group of researchers is working to understand the scope of industrial pollution in the human body. Industrial chemicals may enter the body through pollution or as ingredients in consumer products. In October 2008, the scientists tested cord-blood samples for 20 newborn infants in the United States. The cord blood of the *in utero/newborn* group was tested for 430 industrial compounds, pollutants, and other chemicals, including chemicals linked to brain and nervous-system toxicity, immune-system toxicity, reproductive toxicity, and fertility problems. There are health concerns about the effects of some chemicals on the brain and nervous system. Table 8.3 shows how many of the targeted chemicals were found in each infant's cord blood.

79	145	147	160	116	100	159	151	156	126
137	83	156	94	121	144	123	114	139	99

Table 8.3

Use this sample data to construct a 90 percent confidence interval for the mean number of targeted industrial chemicals to be found in an infant's blood.

Solution

Solution A

From the sample data, you can calculate $\overline{x} = \frac{79 + 145 + \cdots + 139 + 99}{20} = 127.45$ There are 20 infants in the sample, so $n = 20$, and $df = 20 - 1 = 19$.

$$s = \sqrt{\frac{(79-\overline{x})^2 + (145-\overline{x})^2 + \cdots + (139-\overline{x})^2 + (99-\overline{x})^2}{19}} = 25.965.$$

You are asked to calculate a 90 percent confidence interval: $CL = 0.90$, so $\alpha = 1 - CL = 1 - 0.90 = 0.10$. $\frac{\alpha}{2} = 0.05, t_{\frac{\alpha}{2}} = t_{0.05}$

By definition, the area to the right of $t_{0.05}$ is 0.05, and so the area to the left of $t_{0.05}$ is $1 - 0.05 = 0.95$.

Use a table, calculator, or computer to find that $t_{0.05} = 1.729$.

$$EBM = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) = 1.729\left(\frac{25.965}{\sqrt{20}}\right) \approx 10.038$$

$$\overline{x} - EBM = 127.45 - 10.038 = 117.412$$

$$\overline{x} + EBM = 127.45 + 10.038 = 137.488$$

We estimate with 90 percent confidence that the mean number of all targeted industrial chemicals found in cord blood in the United States is between 117.412 and 137.488.

Solution

Solution B

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Enter the data as a list.
Press STAT and arrow over to TESTS.

Arrow down to `8:TInterval` and press ENTER (or you can just press 8). Arrow to Data and press ENTER.
Arrow down to `List` and enter the list name where you put the data.
Arrow down to `Freq` and enter 1.
Arrow down to `C-level` and enter 0.90.
Arrow down to `Calculate` and press ENTER.
The 90 percent confidence interval is (117.41, 137.49).

TRY IT 8.9

A random sample of statistics students was asked to estimate the total number of hours they spend watching television in an average week. The responses are recorded in Table 8.4. Use the following sample data to construct a 98 percent confidence interval for the mean number of hours statistics students will spend watching television in one week.

0	3	1	20	9
5	10	1	10	4
14	2	4	4	5

Table 8.4

8.3 A Population Proportion

During an election year, we see articles in the newspaper that state **confidence intervals** in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has 40 percent of the vote within 3 percentage points (if the sample is large enough). Often, election polls are calculated with 95 percent confidence, so the pollsters would be 95 percent confident that the true proportion of voters who favored the candidate would be between 0.37 and 0.43 (0.40 – 0.03, 0.40 + 0.03).

Investors in the stock market are interested in the true proportion of stocks that go up and down each week. Businesses that sell personal computers are interested in the proportion of households in the United States that own personal computers. Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households in the United States that own personal computers.

The procedure to find the confidence interval, the sample size, the **error bound for a population (*EBP*)**, and the **confidence level** for a proportion is similar to that for the population mean, but the formulas are different.

How do you know you are dealing with a proportion problem? First, the data that you are collecting is categorical, consisting of two categories: Success or Failure, Yes or No. Examples of situations where you are the following trying to estimate the true population proportion are the following: What proportion of the population smoke? What proportion of the population will vote for candidate A? What proportion of the population has a college-level education?

The distribution of the sample proportions (based on samples of size n) is denoted by P' (read "P prime").

The central limit theorem for proportions asserts that the sample proportion distribution P' follows a normal distribution with mean value p, and standard deviation $\sqrt{\frac{p \bullet q}{n}}$, where p is the population proportion and $q = 1 - p$.

The confidence interval has the form ($p' - EBP$, $p' + EBP$). EBP is error bound for the proportion.

$$p' = \frac{x}{n}$$

p' = the *estimated proportion* of successes (p' is a *point estimate* for p, the true proportion.)

x = the *number* of successes

n = the size of the sample

The error bound for a proportion is

$EBP = \left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{p'q'}{n}}\right)$, where $q' = 1 - p'$.

This formula is similar to the error bound formula for a mean, except that the "appropriate standard deviation" is different. For a mean, when the population standard deviation is known, the appropriate standard deviation that we use is $\frac{\sigma}{\sqrt{n}}$. For a proportion, the appropriate standard deviation is $\sqrt{\frac{pq}{n}}$.

However, in the error bound formula, we use $\sqrt{\frac{p'q'}{n}}$ as the standard deviation, instead of $\sqrt{\frac{pq}{n}}$.

In the error bound formula, the *sample proportions p′ and q′, are estimates of the unknown population proportions p and q*. The estimated proportions p' and q' are used because p and q are not known. The sample proportions p' and q' are calculated from the data: p' is the estimated proportion of successes, and q' is the estimated proportion of failures.

The confidence interval can be used only if the number of successes np' and the number of failures nq' are both greater than five.

That is, in order to use the formula for confidence intervals for proportions, you need to verify that both $np' \geq 5$ and $nq' \geq 5$.

EXAMPLE 8.10

Problem

Suppose that a market research firm is hired to estimate the percentage of adults living in a large city who have cell phones. Five hundred randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes, they own cell phones. Using a 95 percent confidence level, compute a confidence interval estimate for the true proportion of adult residents of this city who have cell phones.

Solution

- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+, or 84 calculators (Solution B).

Let X = the number of people in the sample who have cell phones. X is binomial. $X \sim B\left(500, \frac{421}{500}\right)$.

To calculate the confidence interval, you must find p', q', and *EBP*.

$n = 500$

x = the number of successes = 421

$$p' = \frac{x}{n} = \frac{421}{500} = 0.842$$

$p' = 0.842$ is the sample proportion; this is the point estimate of the population proportion.

$$q' = 1 - p' = 1 - 0.842 = 0.158$$

Because $CL = 0.95$, then $\alpha = 1 - CL = 1 - 0.95 = 0.05$ $\left(\frac{\alpha}{2}\right) = 0.025$.

Then, $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$.

Use the TI-83, 83+, or 84+ calculator command invNorm(0.975,0,1) to find $z_{0.025}$. Remember that the area to the right of $z_{0.025}$is 0.025, and the area to the left of $z_{0.025}$is 0.975. This can also be found using appropriate commands on other calculators, using a computer, or using a standard normal probability table.

$$EBP = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} = (1.96)\sqrt{\frac{(0.842)(0.158)}{500}} = 0.032$$

$$p' - EBP = 0.842 - 0.032 = 0.81$$

$$p' + EBP = 0.842 + 0.032 = 0.874$$

The confidence interval for the true binomial population proportion is (p' – EBP, p' + EBP) = (0.810, 0.874).

Interpretation

We estimate with 95 percent confidence that between 81 percent and 87.4 percent of all adult residents of this city have cell phones.

Explanation of 95 percent Confidence Level

Ninety-five percent of the confidence intervals constructed in this way would contain the true value for the population proportion of all adult residents of this city who have cell phones.

Solution

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Press STAT and arrow over to TESTS.
Arrow down to `A:1-PropZint`. Press ENTER.
Arrow down to x and enter 421.
Arrow down to n and enter 500.
Arrow down to `C-Level` and enter .95.
Arrow down to `Calculate` and press ENTER.
The confidence interval is (0.81003, 0.87397).

TRY IT 8.10

Suppose 250 randomly selected people are surveyed to determine whether they own tablets. Of the 250 surveyed, 98 reported owning tablets. Using a 95 percent confidence level, compute a confidence interval estimate for the true proportion of people who own tablets.

EXAMPLE 8.11

Problem

For a class project, a political science student at a large university wants to estimate the percentage of students who are registered voters. He surveys 500 students and finds that 300 are registered voters. Compute a 90 percent confidence interval for the true percentage of students who are registered voters, and interpret the confidence interval.

Solution

- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+, or 84 calculators (Solution B).

Solution A

$$x = 300 \text{ and } n = 500$$

$$p' = \frac{x}{n} = \frac{300}{500} = 0.600$$

$$q' = 1 - p' = 1 - 0.600 = 0.400$$

Because $CL = 0.90$, then $\alpha = 1 - CL = 1 - 0.90 = 0.10\left(\frac{\alpha}{2}\right) = 0.05$.

$$z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$$

Use the TI-83, 83+, or 84+ calculator command invNorm(0.95,0,1) to find $z_{0.05}$. Remember that the area to the right of $z_{0.05}$ is 0.05, and the area to the left of $z_{0.05}$ is 0.95. This can also be found using appropriate commands on other calculators, using a computer, or using a standard normal probability table.

$$EBP = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} = (1.645)\sqrt{\frac{(0.60)(0.40)}{500}} = 0.036$$

$$p' - EBP = 0.60 - 0.036 = 0.564$$

$$p' + EBP = 0.60 + 0.036 = 0.636$$

The confidence interval for the true binomial population proportion is (p' – *EBP* , p' + *EBP*) = (0.564, 0.636).

Interpretation

- We estimate with 90 percent confidence that the true percentage of all students who are registered voters is between 56.4 percent and 63.6 percent.
- Alternate wording: We estimate with 90 percent confidence that between 56.4 percent and 63.6 percent of all students are registered voters.

Explanation of 90 percent Confidence Level

Ninety percent of all confidence intervals constructed in this way contain the true value for the population percentage of students who are registered voters.

Solution

Solution B

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Press STAT and arrow over to TESTS.
Arrow down to `A:1-PropZint`. Press ENTER.
Arrow down to *x* and enter 300.
Arrow down to *n* and enter 500.
Arrow down to `C-Level` and enter 0.90.
Arrow down to `Calculate` and press ENTER.
The confidence interval is (0.564, 0.636).

TRY IT 8.11

A student polls her school to determine whether students in the school district are for or against the new legislation regarding school uniforms. She surveys 600 students and finds that 480 are against the new legislation.

a. Compute a 90 percent confidence interval for the true percentage of students who are against the new legislation, and interpret the confidence interval.

b. In a sample of 300 students, 68 percent said they own an iPod and a smartphone. Compute a 97 percent confidence interval for the true percentage of students who own an iPod and a smartphone.

Plus-Four Confidence Interval for *p*

There is a certain amount of error introduced into the process of calculating a confidence interval for a proportion. Because we do not know the true proportion for the population, we are forced to use point estimates to calculate the appropriate standard deviation of the sampling distribution. Studies have shown that the resulting estimation of the standard deviation can be flawed.

Fortunately, there is a simple adjustment that allows us to produce more accurate confidence intervals: We simply pretend that we have four additional observations. Two of these observations are successes, and two are failures. The new sample size, then, is $n + 4$, and the new count of successes is $x + 2$.

Computer studies have demonstrated the effectiveness of the **plus-four confidence interval for *p* method**. It should be

used when the confidence level desired is at least 90 percent and the sample size is at least ten.

EXAMPLE 8.12

Problem

A random sample of 25 statistics students was asked: "Have you used a product in the past week?" Six students reported using the product within the past week. Use the plus-four method to find a 95 percent confidence interval for the true proportion of statistics students who use the product weekly.

Solution

Solution A

Six students out of 25 reported using a product within the past week, so $x = 6$ and $n = 25$. Because we are using the plus-four method, we will use $x = 6 + 2 = 8$, and $n = 25 + 4 = 29$.

$$p' = \frac{x}{n} = \frac{8}{29} \approx 0.276$$

$$q' = 1 - p' = 1 - 0.276 = 0.724$$

Because $CL = 0.95$, we know $\alpha = 1 - 0.95 = 0.05$, and $\frac{\alpha}{2} = 0.025$.

$$z_{0.025} = 1.96$$

$$EPB = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} = (1.96)\sqrt{\frac{0.276(0.724)}{29}} \approx 0.163$$

$$p' - EPB = 0.276 - 0.163 = 0.113$$

$$p' + EPB = 0.276 + 0.163 = 0.439$$

We are 95 percent confident that the true proportion of all statistics students who use the product is between 0.113 and 0.439.

Solution

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Press STAT and arrow over to TESTS.
Arrow down to `A:1-PropZint`. Press ENTER.

Arrow down to x and enter 8.
Arrow down to n and enter 29.
Arrow down to `C-Level` and enter 0.95.
Arrow down to `Calculate` and press ENTER.
The confidence interval is (0.113, 0.439).

Reminder

Remember that the plus-four method assumes an additional four trials: two successes and two failures. You do not need to change the process for calculating the confidence interval; simply update the values of x and n to reflect these additional trials.

TRY IT 8.12

Out of a random sample of 65 freshmen at State University, 31 students have declared their majors. Use the plus-four method to find a 96 percent confidence interval for the true proportion of freshmen at State University who have declared their majors.

EXAMPLE 8.13

Problem

A group of researchers recently conducted a study analyzing the privacy management habits of teen internet users. In a group of 50 teens, 13 reported having more than 500 friends on a social media site. Use the plus four method to find a 90 percent confidence interval for the true proportion of teens who would report having more than 500 online friends.

Solution

Using plus-four, we have x = 13 + 2 = 15, and n = 50 + 4 = 54.

$$p' = \frac{15}{54} \approx 0.278 \quad \textbf{8.5}$$

$$q' = 1 - p' = 1 - 0.278 = 0.722$$

Because CL = 0.90, we know α = 1 – 0.90 = 0.10, and $\frac{\alpha}{2}$ = 0.05.

$$z_{0.05} = 1.645$$

$$EPB = (z_{\frac{\alpha}{2}}) \left(\sqrt{\frac{p'q'}{n}} \right) = (1.645) \left(\sqrt{\frac{(0.278)(0.722)}{54}} \right) \approx 0.100$$

$$p' - EPB = 0.278 - 0.100 = 0.178$$

$$p' + EPB = 0.278 + 0.160 = 0.378$$

We are 90 percent confident that between 17.8 percent and 37.8 percent of all teens would report having more than 500 friends on a social media site.

Solution

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Press STAT and arrow over to TESTS.
Arrow down to `A:1-PropZint`. Press ENTER.
Arrow down to x and enter 15.
Arrow down to n and enter 54.
Arrow down to `C-Level` and enter 0.90.
Arrow down to `Calculate` and press ENTER.
The confidence interval is (0.178, 0.378).

TRY IT 8.13

The research group referenced in Example 8.13 talked to teens in smaller focus groups but also interviewed additional teens over the phone. When the study was complete, 588 teens had answered the question about their social media site friends, with 159 saying that they have more than 500 friends. Use the plus-four method to find a 90 percent confidence interval for the true proportion of teens who would report having more than 500 online friends based on this larger sample. Compare the results to those in Example 8.13.

Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size.

The margin of error formula for a population proportion is

- $EBP = z_{\frac{\alpha}{2}} \times \sqrt{\frac{p' \bullet q'}{n}}$, where p' is the sample proportion, $q' = 1 - p'$, and n is the sample size.
- Solving for n gives you an equation for the sample size.
- $n = \frac{\left(z_{\frac{\alpha}{2}}\right)^2 (p'q')}{EBP^2}$. This formula tells us that we can compute the sample size n required for a confidence level of $Cl = 1 - \alpha$ by taking the square of the critical value $z_{\frac{\alpha}{2}}$, multiplying by the point estimate p', and by $q' = 1 - p'$ and finally dividing the result by the square of the margin of error. *Always remember to round up the value of n.*

EXAMPLE 8.14

Problem

Suppose a mobile phone company wants to determine the current percentage of customers ages 50+ who use text messaging on their cell phones. How many customers ages 50+ should the company survey in order to be 90 percent confident that the estimated (sample) proportion is within 3 percentage points of the true population proportion of customers ages 50+ who use text messaging on their cell phones? Assume that $p' = 0.5$.

Solution

From the problem, we know that $EBP = 0.03$ (3 percent=0.03), and $z_{\frac{\alpha}{2}}$ $z_{0.05} = 1.645$ because the confidence level is 90 percent.

To calculate the sample size n, use the formula and make the substitutions.

$$n = \frac{z^2 p' q'}{EBP^2} \; gives \; n = \frac{1.645^2 (0.5)(0.5)}{0.03^2} = 751.7$$

Round the answer to the next higher value. The sample size should be 752 cell phone customers ages 50+ in order to be 90 percent confident that the estimated (sample) proportion is within 3 percentage points of the true population proportion of all customers ages 50+ who use text messaging on their cell phones.

TRY IT 8.14

An internet marketing company wants to determine the current percentage of customers who click on ads on their smartphones. How many customers should the company survey in order to be 90 percent confident that the estimated proportion is within 5 percentage points of the true population proportion of customers who click on ads on their smartphones? Assume that the sample proportion p' is 0.50.

8.4 Confidence Interval (Home Costs)

Stats Lab

Confidence Interval (Home Costs)

Student Learning Outcomes

- The student will calculate the 90 percent confidence interval for the mean cost of a home in the area in which this school is located.
- The student will interpret confidence intervals.
- The student will determine the effects of changing conditions on the confidence interval.

Collect the Data

Check the Real Estate section in your local newspaper. Record the sale prices for 35 randomly selected homes recently listed in the county.

Note

Many newspapers list them only one day per week. Also, we will assume that homes come up for sale randomly.

1. Complete the following table:

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

Table 8.5

Describe the Data

1. Compute the following:
   a. $\overline{x}$ = _____
   b. s_x = _____
   c. n = _____
2. In words, define the random variable $\overline{X}$.
3. State the estimated distribution to use. Use both words and symbols.

Find the Confidence interval

1. Calculate the confidence interval and the error bound.
   a. Confidence interval: _____
   b. Error Bound: _____
2. How much area is in both tails (combined)? α = _____
3. How much area is in each tail? $\frac{\alpha}{2}$ = _____
4. Fill in the blanks on the graph with the area in each section. Then, fill in the number line with the upper and lower limits of the confidence interval and the sample mean.

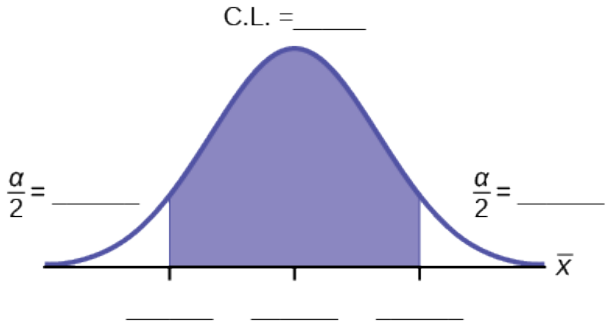


Figure 8.7

5. Some students think that a 90 percent confidence interval contains 90 percent of the data. Use the list of data on the first page and count how many of the data values lie within the confidence interval. What percentage is this? Is

this percentage close to 90 percent? Explain why this percentage should or should not be close to 90 percent.

Describe the Confidence Interval

1. In two to three complete sentences, explain what a confidence interval means (in general), as if you were talking to someone who has not taken statistics.
2. In one to two complete sentences, explain what this confidence interval means for this particular study.

Use the Data to Construct Confidence Intervals

1. Using the given information, construct a confidence interval for each confidence level given.

Confidence Level	*EBM*/Error Bound	Confidence Interval
50%		
80%		
95%		
99%		

Table 8.6

2. What happens to the *EBM* as the confidence level increases? Does the width of the confidence interval increase or decrease? Explain why this happens.

8.5 Confidence Interval (Place of Birth)

Stats Lab

Confidence Interval (Place of Birth)

Student Learning Outcomes

- The student will calculate the 90 percent confidence interval of the proportion of students in this school who were born in this state.
- The student will interpret confidence intervals.
- The student will determine the effects of changing conditions on the confidence interval.

Collect the Data

1. Survey the students in your class, asking them whether they were born in this state. Let X = the number who were born in this state.
   a. n = _____
   b. x = _____
2. In words, define the random variable P'.
3. State the estimated distribution to use.

Find the Confidence interval and Error bound

1. Calculate the confidence interval and the error bound.
   a. Confidence interval: ____
   b. Error Bound: ____
2. How much area is in both tails (combined)? α = ____
3. How much area is in each tail? $\frac{\alpha}{2}$ = ____
4. Fill in the blanks on the graph with the area in each section. Then, fill in the number line with the upper and lower limits of the confidence interval and the sample proportion.

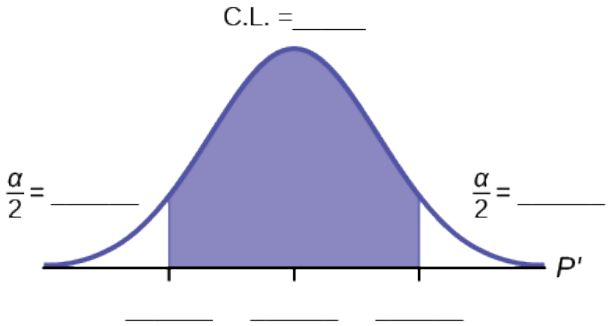


Figure 8.8

Describe the Confidence Interval

1. In two to three complete sentences, explain what a confidence interval means (in general), as though you were talking to someone who has not taken statistics.
2. In one to two complete sentences, explain what this confidence interval means for this particular study.
3. Construct a confidence interval for each confidence level given.

Confidence Level	*EBP*/Error Bound	Confidence Interval
50%		
80%		
95%		
99%		

Table 8.7

4. What happens to the *EBP* as the confidence level increases? Does the width of the confidence interval increase or decrease? Explain why this happens.

8.6 Confidence Interval (Women's Heights)

Stats Lab

Confidence Interval (Women's Heights)

Student Learning Outcomes

- The student will calculate a 90 percent confidence interval using the given data.
- The student will determine the relationship between the confidence level and the percentage of constructed intervals that contain the population mean.

Given:

59.4	71.6	69.3	65.0	62.9	66.5	61.7	55.2
67.5	67.2	63.8	62.9	63.0	63.9	68.7	65.5

Table 8.8 Heights of 100 Women (in Inches)

61.9	69.6	58.7	63.4	61.8	60.6	69.8	60.0
64.9	66.1	66.8	60.6	65.6	63.8	61.3	59.2
64.1	59.3	64.9	62.4	63.5	60.9	63.3	66.3
61.5	64.3	62.9	60.6	63.8	58.8	64.9	65.7
62.5	70.9	62.9	63.1	62.2	58.7	64.7	66.0
60.5	64.7	65.4	60.2	65.0	64.1	61.1	65.3
64.6	59.2	61.4	62.0	63.5	61.4	65.5	62.3
65.5	64.7	58.8	66.1	64.9	66.9	57.9	69.8
58.5	63.4	69.2	65.9	62.2	60.0	58.1	62.5
62.4	59.1	66.4	61.2	60.4	58.7	66.7	67.5
63.2	56.6	67.7	62.5				

Table 8.8 Heights of 100 Women (in Inches)

1. Table 8.8 lists the heights of 100 women. Use a random number generator to select 10 data values randomly.
2. Calculate the sample mean and the sample standard deviation. Assume that the population standard deviation is known to be 3.3 in. With these values, construct a 90 percent confidence interval for your sample of 10 values. Write the confidence interval you obtained in the first space of Table 8.9.
3. Now write your confidence interval on the board. As others in the class write their confidence intervals on the board, copy them into Table 8.9.

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

Table 8.9 90 percent Confidence Intervals

Discussion Questions

1. The actual population mean for the 100 heights given in Table 8.8 is $\mu = 63.4$. Using the class listing of confidence intervals, count how many of them contain the population mean μ; i.e., for how many intervals does the value of μ lie between the endpoints of the confidence interval?

2. Divide this number by the total number of confidence intervals generated by the class to determine the percentage of confidence intervals that contain the mean μ. Write that percentage here: _____.
3. Is the percentage of confidence intervals that contain the population mean μ close to 90 percent?
4. Suppose we had generated 100 confidence intervals. What do you think would happen to the percentage of confidence intervals that contained the population mean?
5. When we construct a 90 percent confidence interval, we say that we are *90 percent confident that the true population mean lies within the confidence interval.* Using complete sentences, explain what we mean by this phrase.
6. Some students think that a 90 percent confidence interval contains 90 percent of the data. Use the list of data given (the heights of women) and count how many of the data values lie within the confidence interval that you generated based on that data. How many of the 100 data values lie within your confidence interval? What percentage is this? Is this percentage close to 90 percent?
7. Explain why it does not make sense to count data values that lie in a confidence interval. Think about the random variable that is being used in the problem.
8. Suppose you obtained the heights of 10 women and calculated a confidence interval from this information. Without knowing the population mean μ, would you have any way of knowing *for certain* whether your interval actually contained the value of μ? Explain.

Key Terms

binomial distribution a discrete random variable (RV) that arises from Bernoulli trials; there are a fixed number, *n*, of independent trials
Independent means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances, the binomial RV *X* is defined as the number of successes in *n* trials. The notation is $X \sim B(\mathbf{n}, \mathbf{p})$. The mean is $\mu = np$, and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly *x* successes in *n* trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$.

confidence interval (*CI*) an interval estimate for an unknown population parameter.
This depends on the following:
- the desired confidence level,
- information that is known about the distribution (for example, known standard deviation), and
- the sample and its size.

confidence level (*CL*) the percentage expression for the probability that the confidence interval contains the true population parameter; for example, if the *CL* = 90 percent, then in 90 out of 100 samples, the interval estimate will enclose the true population parameter

degrees of freedom (*df*) the number of objects in a sample that are free to vary

error bound for a population mean (*EBM*) the margin of error; depends on the confidence level, sample size, and known or estimated population standard deviation

error bound for a population proportion (*EBP*) the margin of error; depends on the confidence level, the sample size, and the estimated (from the sample) proportion of successes

inferential statistics also called statistical inference or inductive statistics; this facet of statistics deals with estimating a population parameter based on a sample statistic
For example, if four out of the 100 calculators sampled are defective, we might infer that 4 percent of the production is defective.

normal distribution a bell-shaped continuous random variable *X*, with center at the mean value (μ) and distance from the center to the inflection points of the bell curve given by the standard deviation (σ).
We write $X \sim N(\mu, \sigma)$. If the mean value is 0 and the standard deviation is 1, the random variable is called the standard normal distribution, and it is denoted with the letter *Z*

parameter a numerical characteristic of a population

plus-four confidence interval plus-four confidence interval when you add two imaginary successes and two imaginary failures (four overall) to your sample

point estimate a single number computed from a sample and used to estimate a population parameter

standard deviation a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: *s* for sample standard deviation and σ for population standard deviation

Student's *t*-distribution investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student
the major characteristics of the random variable (RV) are as follows:
- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as *n* get larger.
- There is a family of *t*-distributions: Each representative of the family is completely defined by the number of degrees of freedom, which is one less than the number of data.

Chapter Review

8.1 A Single Population Mean Using the Normal Distribution

In this module, we learned how to calculate the confidence interval for a single population mean where the population standard deviation is known. When estimating a population mean, the margin of error is called the error bound for a population mean (*EBM*). A confidence interval has the general form

(lower bound, upper bound) = (point estimate – *EBM*, point estimate + *EBM*).

The calculation of *EBM* depends on the size of the sample and the level of confidence desired. The confidence level is the percentage of all possible samples that can be expected to include the true population parameter. As the confidence level increases, the corresponding *EBM* increases as well. As the sample size increases, the *EBM* decreases. By the central limit theorem,

$$EBM = z\frac{\sigma}{\sqrt{n}.}$$

Given a confidence interval, you can work backward to find the error bound (*EBM*) or the sample mean. To find the error bound, find the difference of the upper bound of the interval and the mean. If you do not know the sample mean, you can find the error bound by calculating half of the difference of the upper and lower bounds. To find the sample mean given a confidence interval, find the difference of the upper bound and the error bound. If the error bound is unknown, then average the upper and lower bounds of the confidence interval to find the sample mean.

Sometimes researchers know in advance that they want to estimate a population mean within a specific margin of error for a given level of confidence. In that case, solve the *EBM* formula for *n* to discover the size of the sample that is needed to achieve this goal:

$$n = \frac{z^2\sigma^2}{EBM^2}$$

8.2 A Single Population Mean Using the Student's t-Distribution

In many cases, the researcher does not know the population standard deviation, σ, of the measure being studied. In these cases, it is common to use the sample standard deviation, *s*, as an estimate of σ. The normal distribution creates accurate confidence intervals when σ is known, but it is not as accurate when *s* is used as an estimate. In this case, the Student's *t*-distribution is much better. Define a *t*-score using the following formula:

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

The *t*-score follows the Student's *t*-distribution with $n - 1$ degrees of freedom. The confidence interval under this distribution is calculated with $EBM = \left(t_{\frac{\alpha}{2}}\right)\frac{s}{\sqrt{n,}}$ where $t_{\frac{\alpha}{2}}$ is the *t*-score with area to the right equal to $\frac{\alpha}{2}$, *s* is the sample standard deviation, and *n* is the sample size. Use a table, calculator, or computer to find $t_{\frac{\alpha}{2}}$ for a given α.

8.3 A Population Proportion

Some statistical measures, like many survey questions, measure qualitative rather than quantitative data. In this case, the population parameter being estimated is a proportion. It is possible to create a confidence interval for the true population proportion by following procedures similar to those used in creating confidence intervals for population means. The formulas are slightly different, but they follow the same reasoning.

Let p' represent the sample proportion, x/n, where *x* represents the number of successes, and *n* represents the sample size. Let $q' = 1 - p'$. Then the confidence interval for a population proportion is given by the following formula:

$$\text{(lower bound, upper bound)} = (p' - EBP, p' + EBP) = (p' - z\sqrt{\frac{p'q'}{n}}, p' + z\sqrt{\frac{p'q'}{n}}).$$

The **plus–four method** for calculating confidence intervals is an attempt to balance the error introduced by using estimates of the population proportion when calculating the standard deviation of the sampling distribution. Simply imagine four additional trials in the study; two are successes and two are failures. Calculate $p' = \frac{x+2}{n+4}$, and proceed to find the confidence interval. When sample sizes are small, this method has been demonstrated to provide more accurate confidence intervals than the standard formula used for larger samples.

Formula Review

8.1 A Single Population Mean Using the Normal Distribution

$\overline{X} \sim N\left(\mu_X, \frac{\sigma}{\sqrt{n}}\right)$ The distribution of sample means is normally distributed with mean equal to the population mean and standard deviation given by the population standard deviation divided by the square root of the sample size.

The general form for a confidence interval for a single population mean, known standard deviation, normal distribution is given by

(lower bound, upper bound) = (point estimate – *EBM*, point estimate + *EBM*)

$$= (\overline{x} - EBM, \overline{x} + EBM)$$

$$= \left(\overline{x} - z\frac{\sigma}{\sqrt{n}}, \overline{x} + z\frac{\sigma}{\sqrt{n}}\right).$$

$EBM = z\frac{\sigma}{\sqrt{n}}$ = the error bound for the mean, or the margin of error for a single population mean; this formula is used when the population standard deviation is known.

CL = confidence level, or the proportion of confidence intervals created that is expected to contain the true

population parameter

$\alpha = 1 - CL$ = the proportion of confidence intervals that will not contain the population parameter

$z_{\frac{\alpha}{2}}$ = the *z*-score with the property that the area to the right of the *z*-score is $\frac{\alpha}{2}$; this is the *z*-score, used in the calculation of *EBM*, where α = 1 – *CL*.

$n = \frac{z^2\sigma^2}{EBM^2}$ = the formula used to determine the sample size (*n*) needed to achieve a desired margin of error at a given level of confidence

General form of a confidence interval

(lower value, upper value) = (point estimate error bound, point estimate + error bound)

To find the error bound when you know the confidence interval,

error bound = upper value point estimate **or** error bound = $\frac{\text{upper value} - \text{lower value}}{2}$.

Single population mean, known standard deviation, normal distribution

Use the normal distribution for means; population standard deviation is known: $EBM = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

The confidence interval has the format ($\overline{x}$ – *EBM*, $\overline{x}$ + *EBM*).

8.2 A Single Population Mean Using the Student's t-Distribution

s = the standard deviation of sample values

$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$ is the formula for the *t*-score, which measures how far away a measure is from the population mean in the Student's *t*-distribution.

df = *n* – 1; the degrees of freedom for a Student's *t*-distribution, where *n* represents the size of the sample

$T \sim t_{df}$ the random variable, *T*, has a Student's *t*-distribution with *df* degrees of freedom

$EBM = t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$ = the error bound for the population mean when the population standard deviation is unknown

$t_{\frac{\alpha}{2}}$ is the *t*-score in the Student's *t*-distribution with area to the right equal to $\frac{\alpha}{2}$.

The general form for a confidence interval for a single mean, population standard deviation unknown, Student's *t* is given by
(lower bound, upper bound) = (point estimate – *EBM*, point estimate + *EBM*)

$$= \left(\overline{x} - \frac{ts}{\sqrt{n}}, \overline{x} + \frac{ts}{\sqrt{n}}\right).$$

8.3 A Population Proportion

$p' = x/n$, where *x* represents the number of successes and *n* represents the sample size. The variable p' is the sample proportion and serves as the point estimate for the true population proportion.

$$q' = 1 - p'$$

$p' \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$ The variable p' has a **binomial distribution** that can be approximated with the normal distribution shown here,

$$EBP = \text{the error bound for a proportion} = z_{\frac{\alpha}{2}} \sqrt{\frac{p'q'}{n}}.$$

Confidence interval for a proportion:

$$\text{(lower bound, upper bound)} = (p' - EBP, p' + EBP) = \left(p' - z\sqrt{\frac{p}{}}\right.$$

$n = \frac{z_{\frac{\alpha}{2}}{}^2 p'q'}{EBP^2}$ provides the number of participants needed to estimate the population proportion with confidence 1 – *α* and margin of error *EBP*.

Use the normal distribution for a single population proportion $p' = \frac{x}{n}$.

$$EBP = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}}\ p' + q' = 1$$

The confidence interval has the format (p' – *EBP*, p' + *EBP*).

$\overline{x}$ is a point estimate for μ.

p' is a point estimate for ρ.

s is a point estimate for σ.

Practice

8.1 A Single Population Mean Using the Normal Distribution

Use the following information to answer the next five exercises: The standard deviation of the weights of elephants is known to be approximately 15 lb. We wish to construct a 95 percent confidence interval for the mean weight of newborn elephant calves. Fifty newborn elephants are weighed. The sample mean is 244 lb. The sample standard deviation is 11 lb.

1. Identify the following:
   a. $\overline{x}$ = _____

b. σ = _____
c. n = _____

2. In words, define the random variables X and $\overline{X}$.
3. Which distribution should you use for this problem?
4. Construct a 95 percent confidence interval for the population mean weight of newborn elephants. State the confidence interval, sketch the graph, and calculate the error bound.
5. What will happen to the confidence interval obtained, if 500 newborn elephants are weighed instead of 50? Why?

Use the following information to answer the next seven exercises: The U.S. Census Bureau conducts a study to determine the time needed to complete the short form. The bureau surveys 200 people. The sample mean is 8.2 minutes. There is a known standard deviation of 2.2 minutes. The population distribution is assumed to be normal.

6. Identify the following:
   a. $\overline{x}$ = _____
   b. σ = _____
   c. n = _____
7. In words, define the random variables X and $\overline{X}$.
8. Which distribution should you use for this problem?
9. Construct a 90 percent confidence interval for the population mean time to complete the forms. State the confidence interval, sketch the graph, and calculate the error bound.
10. If the Census wants to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?
11. If the Census did another survey, kept the error bound the same, and surveyed only 50 people instead of 200, what would happen to the level of confidence? Why?
12. Suppose the Census needed to be 98 percent confident of the population mean length of time. Would the Census have to survey more people? Why or why not?

Use the following information to answer the next 10 exercises: A sample of 20 heads of lettuce was selected. Assume that the population distribution of head weight is normal. The weight of each head of lettuce was then recorded. The mean weight was 2.2 lb, with a standard deviation of 0.1 lb. The population standard deviation is known to be 0.2 lb.

13. Identify the following:
   a. $\overline{x}$ = _____
   b. σ = _____
   c. n = _____
14. In words, define the random variable X.
15. In words, define the random variable $\overline{X}$.
16. Which distribution should you use for this problem?
17. Construct a 90 percent confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the error bound.
18. Construct a 95 percent confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the error bound.
19. In complete sentences, explain why the confidence interval in Exercise 8.17 is larger than in Exercise 8.18.
20. In complete sentences, give an interpretation of what the interval in Exercise 8.18 means.
21. What would happen if 40 heads of lettuce were sampled instead of 20 and the error bound remained the same?
22. What would happen if 40 heads of lettuce were sampled instead of 20 and the confidence level remained the same?

Use the following information to answer the next 14 exercises: The mean age for all Foothill College students for a recent fall term was 33.2. The population standard deviation has been pretty consistent at 15. Suppose that 25 winter students were randomly selected. The mean age for the sample was 30.4. We are interested in the true mean age for winter Foothill College students. Let X = the age of a winter Foothill College student.

23. $\overline{x}$ = _____

24. n = _____

25. _____ = 15

26. In words, define the random variable $\overline{X}$.

27. What is $\overline{x}$ estimating?

28. Is σ_x known?

29. As a result of your answer to Exercise 8.26, state the exact distribution to use when calculating the confidence interval.

Construct a 95 percent confidence interval for the true mean age of winter Foothill College students by working out and then answering the next eight exercises.

30. How much area is in both tails (combined)? α =_____

31. How much area is in each tail? $\frac{\alpha}{2}$ =_____

32. Identify the following specifications:
   a. lower limit
   b. upper limit
   c. error bound

33. The 95 percent confidence interval is _____.

34. Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample mean.

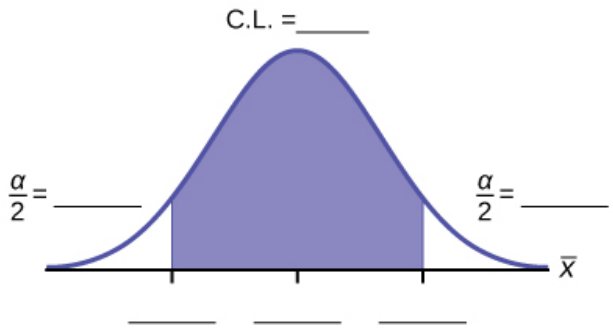


Figure 8.9

35. In one complete sentence, explain what the interval means.

36. Using the same mean, standard deviation, and level of confidence, suppose that n were 69 instead of 25. Would the error bound become larger or smaller? How do you know?

37. Using the same mean, standard deviation, and sample size, how would the error bound change if the confidence level were reduced to 90 percent? Why?

8.2 A Single Population Mean Using the Student's t-Distribution

Use the following information to answer the next five exercises: A hospital is trying to cut down on emergency room wait times. It is interested in the amount of time patients must wait before being called back to be examined. An investigation committee randomly surveyed 70 patients. The sample mean was 1.5 hr, with a sample standard deviation of 0.5 hr.

38. Identify the following:

a. $\overline{x}$ =_____
b. s_x =_____
c. n =_____
d. $n - 1$ =_____

39. Define the random variables X and $\overline{X}$ in words.

40. Which distribution should you use for this problem?

41. Construct a 95 percent confidence interval for the population mean time spent waiting. State the confidence interval, sketch the graph, and calculate the error bound.

42. Explain in complete sentences what the confidence interval means.

Use the following information to answer the next six exercises: One hundred eight Americans were surveyed to determine the number of hours they spend watching television each month. It was revealed that they watch an average of 151 hours each month, with a standard deviation of 32 hours. Assume that the underlying population distribution is normal.

43. Identify the following:
a. $\overline{x}$ =_____
b. s_x =_____
c. n =_____
d. $n - 1$ =_____

44. Define the random variable X in words.

45. Define the random variable $\overline{X}$ in words.

46. Which distribution should you use for this problem?

47. Construct a 99 percent confidence interval for the population mean hours spent watching television per month. State the confidence interval, sketch the graph, and calculate the error bound.

48. Why would the error bound change if the confidence level were lowered to 95 percent?

Use the following information to answer the next 13 exercises: The data in Table 8.10 are the result of a random survey of 39 national flags (with replacement between picks) from various countries. We are interested in finding a confidence interval for the true mean number of colors on a national flag. Let X = the number of colors on a national flag.

X	Freq.
1	1
2	7
3	18
4	7
5	6

Table 8.10

49. Calculate the following:
a. $\overline{x}$ =_____
b. s_x =_____
c. n =_____

50. Define the random variable $\overline{X}$ in words.

51. What is $\overline{x}$ estimating?

52. Is σ_x known?

53. As a result of your answer to Exercise 8.52, state the exact distribution to use when calculating the confidence interval.

Construct a 95 percent confidence interval for the true mean number of colors on national flags.

54. How much area is in both tails (combined)?

55. How much area is in each tail?

56. Calculate the following:
a. lower limit
b. upper limit
c. error bound

57. The 95 percent confidence interval is____.

58. Fill in the blanks on the graph with the areas, the upper and lower limits of the confidence interval, and the sample mean.

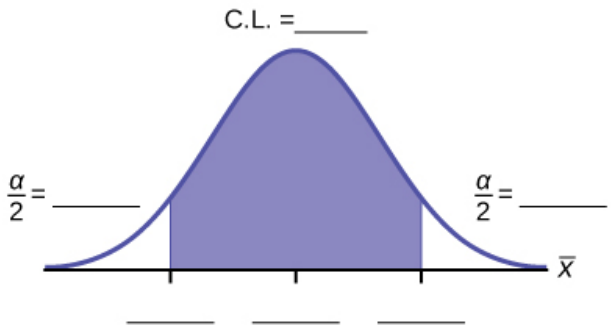


Figure 8.10

59. In one complete sentence, explain what the interval means.

60. Using the same $\overline{x}$, s_x, and level of confidence, suppose that *n* were 69 instead of 39. Would the error bound become larger or smaller? How do you know?

61. Using the same $\overline{x}$, s_x, and *n* = 39, how would the error bound change if the confidence level were reduced to 90 percent? Why?

8.3 A Population Proportion

Use the following information to answer the next two exercises: Marketing companies are interested in knowing the population percentage of women who make the majority of household purchasing decisions.

62. When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 90 percent confident that the population proportion is estimated to within 0.05?

63. If it were later determined that it was important to be more than 90 percent confident and a new survey were commissioned, how would it affect the minimum number you need to survey? Why?

Use the following information to answer the next five exercises: Suppose a marketing company conducted a survey. It randomly surveyed 200 households and found that in 120 of them, the women made the majority of the purchasing decisions. We are interested in the population proportion of households where women make the majority of the purchasing decisions.

64. Identify the following:
a. *x* = _____

b. n = _____
c. p' = _____

65. Define the random variables X and P' in words.

66. Which distribution should you use for this problem?

67. Construct a 95 percent confidence interval for the population proportion of households where the women make the majority of the purchasing decisions. State the confidence interval, sketch the graph, and calculate the error bound.

68. List two difficulties the company might have in obtaining random results if this survey were done by email.

Use the following information to answer the next five exercises: Of 1,050 randomly selected adults, 360 identified themselves as manual laborers, 280 identified themselves as non-manual wage earners, 250 identified themselves as mid-level managers, and 160 identified themselves as executives. In the survey, 82 percent of manual laborers preferred trucks, 62 percent of non-manual wage earners preferred trucks, 54 percent of mid-level managers preferred trucks, and 26 percent of executives preferred trucks.

69. We are interested in finding the 95 percent confidence interval for the percentage of executives who prefer trucks. Define random variables X and P' in words.

70. Which distribution should you use for this problem?

71. Construct a 95 percent confidence interval. State the confidence interval, sketch the graph, and calculate the error bound.

72. Suppose we want to lower the sampling error. What is one way to accomplish that?

73. The sampling error given in the survey is ±2 percent. Explain what the ±2 percent means.

Use the following information to answer the next five exercises: A poll of 1,200 voters asked what the most significant issue was in the upcoming election. Sixty-five percent answered "the economy." We are interested in the population proportion of voters who believe the economy is the most important.

74. Define the random variable X in words.

75. Define the random variable P' in words.

76. Which distribution should you use for this problem?

77. Construct a 90 percent confidence interval, and state the confidence interval and the error bound.

78. What would happen to the confidence interval if the level of confidence were 95 percent?

Use the following information to answer the next 16 exercises: The Ice Chalet offers dozens of different beginning ice-skating classes. All of the class names are put into a bucket. The 5 p.m., Monday night, ages 8 to 12, beginning ice-skating class is picked. In that class are 64 girls and 16 boys. Suppose that we are interested in the true proportion of girls, ages 8 to 12, in all beginning ice-skating classes at the Ice Chalet. Assume that the children in the selected class are a random sample of the population.

79. What is being counted?

80. In words, define the random variable X.

81. Calculate the following:
a. x = _____
b. n = _____
c. p' = _____

82. State the estimated distribution of X. X~ _____

83. Define a new random variable P'. What is p' estimating?

84. In words, define the random variable P'.

85. State the estimated distribution of P'. Construct a 92 percent confidence interval for the true proportion of girls in the ages 8 to 12 beginning ice-skating classes at the Ice Chalet.

86. How much area is in both tails (combined)?

87. How much area is in each tail?

88. Calculate the following:
- a. lower limit
- b. upper limit
- c. error bound

89. The 92 percent confidence interval is _____.

90. Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample proportion.

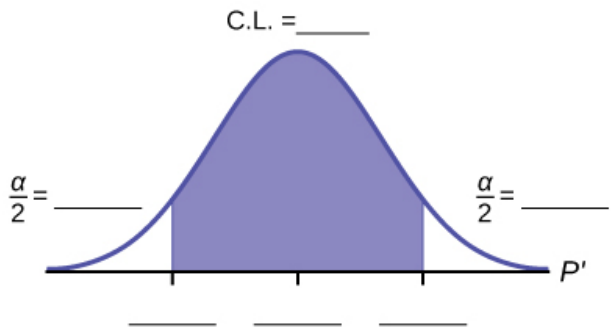


Figure 8.11

91. In one complete sentence, explain what the interval means.

92. Using the same p' and level of confidence, suppose that n were increased to 100. Would the error bound become larger or smaller? How do you know?

93. Using the same p' and $n = 80$, how would the error bound change if the confidence level were increased to 98 percent? Why?

94. If you decreased the allowable error bound, why would the minimum sample size increase (keeping the same level of confidence)?

Homework

8.1 A Single Population Mean Using the Normal Distribution

95. Among various ethnic groups, the standard deviation of heights is known to be approximately three inches. We wish to construct a 95 percent confidence interval for the mean height of male Swedes. 48 male Swedes are surveyed. The sample mean is 71 inches. The sample standard deviation is 2.8 in.
- a. i. $\overline{x}$ =_____
  ii. σ =_____
  iii. n =_____
- b. In words, define the random variables X and $\overline{X}$.
- c. Which distribution should you use for this problem? Explain your choice.
- d. Construct a 95 percent confidence interval for the population mean height of male Swedes.
  i. State the confidence interval.
  ii. Sketch the graph.
  iii. Calculate the error bound.
- e. What will happen to the level of confidence obtained if 1,000 male Swedes are surveyed instead of 48? Why?

96. Announcements for 84 upcoming engineering conferences were randomly picked from a stack of *IEEE Spectrum* magazines. The mean length of the conferences was 3.94 days, with a standard deviation of 1.28 days. Assume

the underlying population is normal.

a. In words, define the random variables X and $\overline{X}$.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95 percent confidence interval for the population mean length of engineering conferences.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

97. Suppose that an accounting firm does a study to determine the time needed to complete one person's tax forms. It randomly surveys 100 people. The sample mean is 23.6 hours. There is a known standard deviation of 7.0 hours. The population distribution is assumed to be normal.

a. i. $\overline{x}$ =_____
   ii. σ =_____
   iii. n =_____
b. In words, define the random variables X and $\overline{X}$.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 90 percent confidence interval for the population mean time to complete the tax forms.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
e. If the firm wished to increase its level of confidence and keep the error bound the same by taking another survey, which changes should it make?
f. If the firm did another survey, kept the error bound the same, and only surveyed 49 people, what would happen to the level of confidence? Why?
g. Suppose that the firm decided that it needed to be at least 96 percent confident of the population mean length of time to within one hour. How would the number of people the firm surveys change? Why?

98. A sample of 16 small bags of the same brand of candies was selected. Assume that the population distribution of bag weights is normal. The weight of each bag was then recorded. The mean weight was two ounces with a standard deviation of 0.12 ounces. The population standard deviation is known to be 0.1 ounce.

a. i. $\overline{x}$ =_____
   ii. σ =_____
   iii. s_X =_____
b. In words, define the random variable X.
c. In words, define the random variable $\overline{X}$.
d. Which distribution should you use for this problem? Explain your choice.
e. Construct a 90 percent confidence interval for the population mean weight of the candies.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
f. Construct a 98 percent confidence interval for the population mean weight of the candies.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
g. In complete sentences, explain why the confidence interval in Part f is larger than the confidence interval in Part e.
h. In complete sentences, give an interpretation of what the interval in Part f means.

99. A camp director is interested in the mean number of letters each child sends during his or her camp session. The population standard deviation is known to be 2.5. A survey of 20 campers is taken. The mean from the sample is 7.9, with a sample standard deviation of 2.8.

a. i. $\overline{x}$ =_____
   ii. σ =_____
   iii. n =_____
b. Define the random variables X and $\overline{X}$ in words.

c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 90 percent confidence interval for the population mean number of letters campers send home.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
e. What will happen to the error bound and confidence interval if 500 campers are surveyed? Why?

100. What is meant by the term *90 percent confident* when constructing a confidence interval for a mean?
a. If we took repeated samples, approximately 90 percent of the samples would produce the same confidence interval.
b. If we took repeated samples, approximately 90 percent of the confidence intervals calculated from those samples would contain the sample mean.
c. If we took repeated samples, approximately 90 percent of the confidence intervals calculated from those samples would contain the true value of the population mean.
d. If we took repeated samples, the sample mean would equal the population mean in approximately 90 percent of the samples.

101. The Federal Election Commission collects information about campaign contributions and disbursements for candidates and political committees during each election cycle. During the 2012 campaign season, there were 1,619 candidates for the House of Representatives across the United States who received contributions from individuals. Table 8.11 shows the total receipts from individuals for a random selection of 40 House candidates rounded to the nearest $100. The standard deviation for this data to the nearest hundred is σ = $909,200.

$3,600	$1,243,900	$10,900	$385,200	$581,500
$7,400	$2,900	$400	$3,714,500	$632,500
$391,000	$467,400	$56,800	$5,800	$405,200
$733,200	$8,000	$468,700	$75,200	$41,000
$13,300	$9,500	$953,800	$1,113,500	$1,109,300
$353,900	$986,100	$88,600	$378,200	$13,200
$3,800	$745,100	$5,800	$3,072,100	$1,626,700
$512,900	$2,309,200	$6,600	$202,400	$15,800

Table 8.11

a. Find the point estimate for the population mean.
b. Using 95 percent confidence, calculate the error bound.
c. Create a 95 percent confidence interval for the mean total individual contributions.
d. Interpret the confidence interval in the context of the problem.

102. The American Community Survey (ACS), part of the U.S. Census Bureau, conducts a yearly census similar to the one taken every 10 years, but with a smaller percentage of participants. The most recent survey estimates with 90 percent confidence that the mean household income in the United States falls between $69,720 and $69,922. Find the point estimate for mean U.S. household income and the error bound for mean U.S. household income.

103. The average height of young adult males has a normal distribution with standard deviation of 2.5 in. You want to estimate the mean height of students at your college or university to within 1 in. with 93 percent confidence. How many male students must you measure?

8.2 A Single Population Mean Using the Student's t-Distribution

104. In six packages of multicolored fruit snacks, there were five red snack pieces. The total number of snack pieces

in the six bags was 68. We wish to calculate a 96 percent confidence interval for the population proportion of red snack pieces.

a. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Calculate p'.
d. Construct a 96 percent confidence interval for the population proportion of red snack pieces per bag.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
e. Do you think that six packages of fruit snacks yield enough data to give accurate results? Why or why not?

105. A random survey of enrollment at 35 community colleges across the United States yielded the following figures: 6,414, 1,550, 2,109, 9,350, 21,828, 4,300, 5,944, 5,722, 2,825, 2,044, 5,481, 5,200, 5,853, 2,750, 10,012, 6,357, 27,000, 9,414, 7,681, 3,200, 17,500, 9,200, 7,380, 18,314, 6,557, 13,713, 17,768, 7,493, 2,771, 2,861, 1,263, 7,285, 28,165, 5,080, 11,622. Assume the underlying population is normal.

a. i. $\overline{x}$ = _____
   ii. s_X = _____
   iii. n = _____
   iv. $n - 1$ = _____
b. Define the random variables X and $\overline{X}$ in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95 percent confidence interval for the population mean enrollment at community colleges in the United States.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
e. What will happen to the error bound and confidence interval if 500 community colleges are surveyed? Why?

106. Suppose that a committee is studying whether there is wasted time in our judicial system. It is interested in the mean amount of time individuals waste at the courthouse waiting to be called for jury duty. The committee randomly surveyed 81 people who recently served as jurors. The sample mean wait time was 8 hr, with a sample standard deviation of 4 hr.

a. i. $\overline{x}$ = _____
   ii. s_X = _____
   iii. n = _____
   iv. $n - 1$ = _____
b. Define the random variables X and $\overline{X}$ in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95 percent confidence interval for the population mean time wasted.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
e. Explain in a complete sentence what the confidence interval means.

107. A pharmaceutical company makes a drug used during surgery. It is assumed that the distribution for the length of time the drug lasts is approximately normal. Researchers in a hospital used the drug on a random sample of nine patients. The effective period of the antibiotic drug for each patient (in hours) was as follows: 2.7, 2.8, 3.0, 2.3, 2.3, 2.2, 2.8, 2.1, and 2.4.

a. i. $\overline{x}$ = _____
   ii. s_X = _____
   iii. n = _____
   iv. $n - 1$ = _____
b. Define the random variable X in words.
c. Define the random variable $\overline{X}$ in words.
d. Which distribution should you use for this problem? Explain your choice.

e. Construct a 95 percent confidence interval for the population mean length of time.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

f. What does it mean to be *95 percent confident* in this problem?

108. Suppose that 14 children who were learning to ride two-wheel bikes were surveyed to determine how long they had to use training wheels. It was revealed that they used them an average of six months, with a sample standard deviation of three months. Assume that the underlying population distribution is normal.

a. i. $\overline{x}$ = _____
   ii. s_x = _____
   iii. n = _____
   iv. $n - 1$ = _____

b. Define the random variable X in words.
c. Define the random variable $\overline{X}$ in words.
d. Which distribution should you use for this problem? Explain your choice.
e. Construct a 99 percent confidence interval for the population mean length of time using training wheels.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

f. Why would the error bound change if the confidence level were lowered to 90 percent?

109. The Federal Election Commission (FEC) collects information about campaign contributions and disbursements for candidates and political committees during each election cycle. A political action committee (PAC) is a committee formed to raise money for candidates and campaigns. A Leadership PAC is a PAC formed by a federal politician (senator or representative) to raise money to help other candidates' campaigns.
The FEC has reported financial information for 556 Leadership PACs that operated during the 2011–2012 election cycle. The following table shows the total receipts during this cycle for a random selection of 20 Leadership PACs.

$46,500.00	$0	$40,966.50	$105,887.20	$5,175.00
$29,050.00	$19,500.00	$181,557.20	$31,500.00	$149,970.80
$2,555,363.20	$12,025.00	$409,000.00	$60,521.70	$18,000.00
$61,810.20	$76,530.80	$119,459.20	$0	$63,520.00
$6,500.00	$502,578.00	$705,061.10	$708,258.90	$135,810.00
$2,000.00	$2,000.00	$0	$1,287,933.80	$219,148.30

Table 8.12

$\overline{x} = \$251,854.23$
$s = \$521,130.41$
Use the sample data to construct a 96 percent confidence interval for the mean amount of money raised by all Leadership PACs during the 2011–2012 election cycle. Use the Student's *t*-distribution.

110. A major business magazine published data on the best small firms in 2012. These were firms that have been publicly traded for at least a year, have a stock price of at least $5 per share, and have reported annual revenue between $5 million and $1 billion. Table 8.13 shows the ages of the corporate CEOs for a random sample of these firms.

48	58	51	61	56
59	74	63	53	50
59	60	60	57	46
55	63	57	47	55
57	43	61	62	49
67	67	55	55	49

Table 8.13

Use the sample data to construct a 90 percent confidence interval for the mean age of CEOs for these top small firms. Use the Student's *t*-distribution.

111. Unoccupied seats on flights cause airlines to lose revenue. Suppose a large airline wants to estimate its mean number of unoccupied seats per flight over the past year. To accomplish this, the records of 225 flights are randomly selected, and the number of unoccupied seats is noted for each of the sampled flights. The sample mean is 11.6 seats, and the sample standard deviation is 4.1 seats.

a. i. $\overline{x}$ = _____
   ii. s_x = _____
   iii. n = _____
   iv. $n - 1$ = _____

b. Define the random variables X and $\overline{X}$ in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 92 percent confidence interval for the population mean number of unoccupied seats per flight.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

112. In a recent sample of 84 used car sales costs, the sample mean was $6,425, with a standard deviation of $3,156. Assume the underlying distribution is approximately normal.

a. Which distribution should you use for this problem? Explain your choice.
b. Define the random variable $\overline{X}$ in words.
c. Construct a 95 percent confidence interval for the population mean cost of a used car.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
d. Explain what a *95 percent confidence interval* means for this study.

113. Six different national brands of chocolate chip cookies were randomly selected at the supermarket. The grams of fat per serving are as follows: 8, 8, 10, 7, 9, 9. Assume the underlying distribution is approximately normal.

a. Construct a 90 percent confidence interval for the population mean grams of fat per serving of chocolate chip cookies sold in supermarkets.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
b. If you wanted a smaller error bound while keeping the same level of confidence, what should have been changed in the study before it was done?
c. Go to the store and record the grams of fat per serving of six brands of chocolate chip cookies.
d. Calculate the mean.
e. Is the mean within the interval you calculated in Part a? Did you expect it to be? Why or why not?

114. A survey of the mean number of cents off given by coupons was conducted by randomly surveying one coupon per page from the coupons section of a local newspaper. The following data were collected: 20¢, 75¢, 50¢, 65¢, 30¢, 55¢, 40¢, 40¢, 30¢, 55¢, $1.50, 40¢, 65¢, 40¢. Assume the underlying distribution is approximately normal.

a. i. $\overline{x}$ = _____
   ii. s_x = _____
   iii. n = _____
   iv. $n - 1$ = _____

b. Define the random variables X and $\overline{X}$ in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95 percent confidence interval for the population mean worth of coupons.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
e. If many random samples were collected with 14 samples as the size, which percentage of the confidence intervals constructed should contain the population mean worth of coupons? Explain why.

Use the following information to answer the next two exercises: A quality control specialist for a restaurant chain takes a random sample of size 12 to check the amount of soda served in the 16-oz serving size. The sample mean is 13.30, with a sample standard deviation of 1.55. Assume the underlying population is normally distributed.

115. Find the 95 percent confidence interval for the true population mean for the amount of soda served.

a. (12.42, 14.18)
b. (12.32, 14.29)
c. (12.50, 14.10)
d. Impossible to determine

116. Which of the following is the error bound?

a. 0.87
b. 1.98
c. 0.99
d. 1.74

8.3 A Population Proportion

117. Insurance companies are interested in knowing the population percentage of drivers who always buckle up before riding in a car.

a. When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 95 percent confident that the population proportion is estimated to within 0.03?
b. If it were later determined that it was important to be more than 95 percent confident and a new survey was commissioned, how would that affect the minimum number you would need to survey? Why?

118. Suppose that the insurance companies did conduct a survey. They randomly surveyed 400 drivers and found that 320 claimed they always buckle up. We are interested in the population proportion of drivers who claim they always buckle up.

a. i. x = _____
   ii. n = _____
   iii. p' = _____

b. Define the random variables X and P' in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95 percent confidence interval for the population proportion who claim they always buckle up.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
e. If this survey were done by telephone, list three difficulties the companies might have in obtaining random results.

119. According to a recent survey of 1,200 people, 61 percent believe that the president is doing an acceptable job. We are interested in the population proportion of people who believe the president is doing an acceptable job.
   a. Define the random variables X and P' in words.
   b. Which distribution should you use for this problem? Explain your choice.
   c. Construct a 90 percent confidence interval for the population proportion of people who believe the president is doing an acceptable job.
      i. State the confidence interval.
      ii. Sketch the graph.
      iii. Calculate the error bound.

120. An article regarding dating and marriage recently appeared in a major newspaper. Of the 1,709 randomly selected adults, 315 identified themselves as ethnicity A, 323 identified themselves as ethnicity B, 254 identified themselves as ethnicity C, and 779 identified themselves as ethnicity D. In this survey, 86 percent of ethnicity B said that they would welcome a person of ethnicity A into their families. Among ethnicity C, 77 percent would welcome a person of ethnicity D into their families, 71 percent would welcome a person of ethnicity A, and 66 percent would welcome a person of ethnicity B.
   a. We are interested in finding the 95 percent confidence interval for the percent of all ethnicity B adults who would welcome a person of ethnicity D into their families. Define the random variables X and P' in words.
   b. Which distribution should you use for this problem? Explain your choice.
   c. Construct a 95 percent confidence interval.
      i. State the confidence interval.
      ii. Sketch the graph.
      iii. Calculate the error bound.

121. Refer to the information in Exercise 8.120.
   a. Construct three 95 percent confidence intervals:
      i. percentage of all ethnicity C who would welcome a person of ethnicity D into their families
      ii. percentage of all ethnicity C who would welcome a person of ethnicity A into their families
      iii. percentage of all ethnicity C who would welcome a person of ethnicity B into their families
   b. Even though the three point estimates are different, do any of the confidence intervals overlap? Which?
   c. For any intervals that do overlap, in words, what does this imply about the significance of the differences in the true proportions?
   d. For any intervals that do not overlap, in words, what does this imply about the significance of the differences in the true proportions?

122. Stanford University conducted a study of whether running is healthy for men and women over age 50. During the first eight years of the study, 1.5 percent of the 451 members of the 50-Plus Fitness Association died. We are interested in the proportion of people over 50 who ran and died in the same eight year period.
   a. Define the random variables X and P' in words.
   b. Which distribution should you use for this problem? Explain your choice.
   c. Construct a 97 percent confidence interval for the population proportion of people over 50 who ran and died in the same 8-year period.
      i. State the confidence interval.
      ii. Sketch the graph.
      iii. Calculate the error bound.
   d. Explain what a *97 percent confidence interval* means for this study.

123. A telephone poll of 1,000 adult Americans was reported in an issue of a national magazine. One of the questions asked, “What is the main problem facing the country?” Twenty percent responded "crime". We are interested in the population proportion of adult Americans who believe that crime is the main problem.
   a. Define the random variables X and P' in words.
   b. Which distribution should you use for this problem? Explain your choice.
   c. Construct a 95 percent confidence interval for the population proportion of adult Americans who believe that crime is the main problem.
      i. State the confidence interval.
      ii. Sketch the graph.
      iii. Calculate the error bound.

d. Suppose we want to lower the sampling error. What is one way to accomplish that?
e. The sampling error given by the group of researchers who conducted the poll is ±3 percent. In one to three complete sentences, explain what the ±3 percent represents.

124. Refer to Exercise 8.123. Another question in the poll asked, "[How much are] you worried about the quality of education in our schools?" Sixty-three percent responded "a lot". We are interested in the population proportion of adult Americans who are worried a lot about the quality of education in our schools.
a. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95 percent confidence interval for the population proportion of adult Americans who are worried a lot about the quality of education in our schools.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
d. The sampling error given by the group of researchers who conducted the poll is ±3 percent. In one to three complete sentences, explain what the ±3 percent represents.

Use the following information to answer the next three exercises: According to a Field Poll, 79 percent of California adults (actual results are 400 out of 506 surveyed) believe that *education and our schools* is one of the top issues facing California. We wish to construct a 90 percent confidence interval for the true proportion of California adults who believe that education and the schools is one of the top issues facing California.

125. A point estimate for the true population proportion is _____.
a. 0.90
b. 1.27
c. 0.79
d. 400

126. A 90 percent confidence interval for the population proportion is _____.
a. (0.761, 0.820)
b. (0.125, 0.188)
c. (0.755, 0.826)
d. (0.130, 0.183)

127. The error bound is approximately _____.
a. 1.581
b. 0.791
c. 0.059
d. 0.030

Use the following information to answer the next two exercises: Five hundred eleven (511) homes in a certain southern California community are randomly surveyed to determine whether they meet minimal earthquake preparedness recommendations. One hundred seventy-three (173) of the homes surveyed meet the minimum recommendations for earthquake preparedness, and 338 do not.

128. Find the confidence interval at the 90 percent confidence level for the true population proportion of southern California community homes meeting at least the minimum recommendations for earthquake preparedness.
a. (0.2975, 0.3796)
b. (0.6270, 0.6959)
c. (0.3041, 0.3730)
d. (0.6204, 0.7025)

129. The point estimate for the population proportion of homes that do not meet the minimum recommendations for earthquake preparedness is _____.
a. 0.6614
b. 0.3386

c. 173
d. 338

130. On May 23, 2013, a polling group reported that of the 1,005 people surveyed, 76 percent of U.S. workers believe that they will continue working past retirement age. The confidence level for this study was reported at 95 percent with a ±3 percent margin of error.
a. Determine the estimated proportion from the sample.
b. Determine the sample size.
c. Identify *CL* and α.
d. Calculate the error bound based on the information provided.
e. Compare the error bound in Part d to the margin of error reported by the polling group. Explain any differences between the values.
f. Create a confidence interval for the results of this study.
g. A reporter is covering the release of this study for a local news station. How should she explain the confidence interval to her audience?

131. A national survey of 1,000 adults was conducted on May 13, 2013, by a group of researchers. It concluded with 95 percent confidence that 49 percent to 55 percent of Americans believe that big-time college sports programs corrupt the process of higher education.
a. Find the point estimate and the error bound for this confidence interval.
b. Can we (with 95 percent confidence) conclude that more than half of all American adults believe this?
c. Use the point estimate from Part a and $n = 1{,}000$ to calculate a 75 percent confidence interval for the proportion of American adults who believe that major college sports programs corrupt higher education.
d. Can we (with 75 percent confidence) conclude that at least half of all American adults believe this?

132. A polling group recently conducted a survey asking adults across the United States about music preferences. When asked, 80 of the 571 participants download music weekly.
a. Create a 99 percent confidence interval for the true proportion of American adults who download music weekly.
b. This survey was conducted through automated telephone interviews on May 6 and 7, 2013. The error bound of the survey compensates for sampling error, or natural variability among samples. List some factors that could affect the survey's outcome that are not covered by the margin of error.
c. Without performing any calculations, describe how the confidence interval would change if the confidence level decreased from 99 percent to 90 percent.

133. You plan to conduct a survey on your college campus to learn about the political awareness of students. You want to estimate the true proportion of college students on your campus who voted in the 2012 presidential election with 95 percent confidence and a margin of error no greater than 5 percent. How many students must you interview?

134. In a recent poll, 9 of 48 respondents rated the likelihood of a certain event occurring in their community as *likely* or *very likely*. Use the plus-four method to create a 97 percent confidence interval for the proportion of American adults who believe that the event is likely or very likely. Explain what this confidence interval means in the context of the problem. A local poll in a New England town found that nine of 48 households think winter-proofing their cars is *very important*. Use the plus-four method to create a 97 percent confidence interval for the proportion of town residents who think winter-proofing their cars is *very important*. Explain what this confidence interval means in the context of this scenario.

References

8.1 A Single Population Mean Using the Normal Distribution

Centers for Disease Control and Prevention. (n.d.). *National health and nutrition examination survey*. Retrieved from http://www.cdc.gov/nchs/nhanes.htm

Foothill De Anza Community College District. (n.d.). *Headcount enrollment trends by student demographics ten-year fall trends to most recently completed fall*. Retrieved from http://research.fhda.edu/factbook/FH_Demo_Trends/FoothillDemographicTrends.htm

Kuczmarski, R. J., et al. (2002). *2000 CDC growth charts for the United States: Methods and development*. Retrieved from http://www.cdc.gov/growthcharts/2000growthchart-us.pdf

La, L., and; German, K. (n.d.). Cell phones with the highest radiation levels. *CNET*. Retrieved from http://reviews.cnet.com/cell-phone-radiation-levels/

U.S. Census Bureau. (n.d.). *American FactFinder*. Retrieved from http://factfinder2.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t

U.S. Census Bureau. (2011). *Mean income in the past 12 months (in 2011 inflation-adjusted dollars): 2011 American Community Survey 1-year estimates*. Retrieved from http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_11_1YR_S1902&prodType=table

U.S. Federal Election Commission. (n.d.). *Disclosure data catalog: Candidate summary report 2012*. Retrieved from http://www.fec.gov/data/CandidateSummary.do?format=html&election_yr=2012

U.S. Federal Election Commission. (n.d.). *Metadata description of candidate summary file*. Retrieved from http://www.fec.gov/finance/disclosure/metadata/metadataforcandidatesummary.shtml

8.2 A Single Population Mean Using the Student's t-Distribution

Bloomberg Businessweek. (n.d.). Retrieved from http://www.businessweek.com/

Environmental Working Group. (n.d.). *Human toxome project: Mapping the pollution in people*. Retrieved from http://www.ewg.org/sites/humantoxome/participants/participant-group.php?group=in+utero%2Fnewborn

Federal Election Commission. (n.d.). *Disclosure data catalog: 2012 leadership PACs and sponsors*. Retrieved from http://www.fec.gov/data/index.jsp

Federal Election Commission. (n.d.). *Metadata description of leadership PAC list*. Retrieved from http://www.fec.gov/finance/disclosure/metadata/metadataLeadershipPacList.shtml

Forbes. (2013). America's best small companies. Retrieved from http://www.forbes.com/best-small-companies/list/

Forbes. (n.d.). Retrieved from http://www.forbes.com/

Microsoft Bookshelf. (n.d.).

8.3 A Population Proportion

Jensen, T. (2013). *Democrats, Republicans divided on opinion of music icons*. Retrieved from http://www.publicpolicypolling.com/Day2MusicPoll.pdf

Madden, M., et al. (2013). *Teens, social media, and privacy*. Retrieved from http://www.pewinternet.org/Reports/2013/Teens-Social-Media-And-Privacy.aspx

Princeton Survey Research Associates International. (2012). *2012 teens and privacy management survey*. Retrieved from http://www.pewinternet.org/~/media//Files/Questionnaire/2013/Methods%20and%20Questions_Teens%20and%20Social%20Media.pdf

Rasmussen Reports. (2013). *52% say big-time college athletics corrupt education process*. Retrieved from http://www.rasmussenreports.com/public_content/lifestyle/sports/may_2013/52_say_big_time_college_athletics_corrupt_education_process

Saad, L. (2013). *Three in four U.S. workers plan to work past retirement age*. Retrieved from http://www.gallup.com/poll/162758/three-fourworkers-plan-work-past-retirement-age.aspx

The Field Poll. (n.d.). Retrieved from http://field.com/fieldpollonline/subscribers/

Zogby Analytics. (2013). *New SUNYIT/Zogby analytics poll: Few Americans worry about emergency situations occurring in their community; Only one in three have an emergency plan; 70% support infrastructure "investment" for national security*. Retrieved from http://www.zogbyanalytics.com/news/299-americans-neither-worried-norprepared-in-case-of-a-disaster-sunyit-zogby-analytics-poll

Solutions

1. a. 244
b. 15
c. 50

3. $N\left(244, \frac{15}{\sqrt{50}}\right)$

5. As the sample size increases, there will be less variability in the mean, so the interval size decreases.

7. X is the time in minutes it takes to complete the U.S. Census short form. $\overline{X}$ is the mean time it took a sample of 200 people to complete the U.S. Census short form.

9. CI: (7.9441, 8.4559)

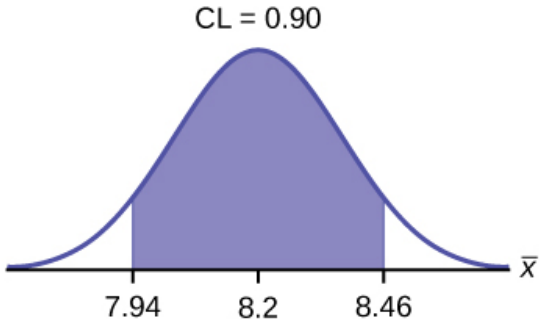


Figure 8.12

EBM = 0.26

11. The level of confidence would decrease, because decreasing n makes the confidence interval wider, so at the same error bound, the confidence level decreases.

13. a. $\overline{x} = 2.2$
b. $\sigma = 0.2$
c. $n = 20$

15. $\overline{X}$ is the mean weight of a sample of 20 heads of lettuce.

17. *EBM* = 0.07
CI: (2.1264, 2.2736)

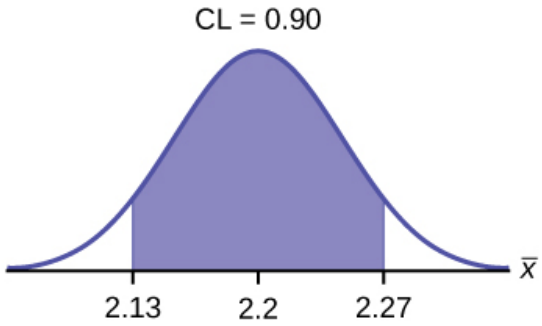


Figure 8.13

19. The interval is greater, because the level of confidence increased. If the only change made in the analysis is a change in confidence level, then all we are doing is changing how much area is being calculated for the normal distribution. Therefore, a larger confidence level results in larger areas and larger intervals.

21. The confidence level would increase.

23. 30.4

25. σ

27. μ

29. normal

31. 0.025

33. (24.52,36.28)

35. We are 95 percent confident that the true mean age for winter Foothill College students is between 24.52 and 36.28.

37. The error bound for the mean would decrease, because as the CL decreases, you need less area under the normal curve (which translates into a smaller interval) to capture the true population mean.

39. X is the number of hours a patient waits in the emergency room before being called back to be examined. $\overline{X}$ is the mean wait time of 70 patients in the emergency room.

41. CI: (1.3808, 1.6192)

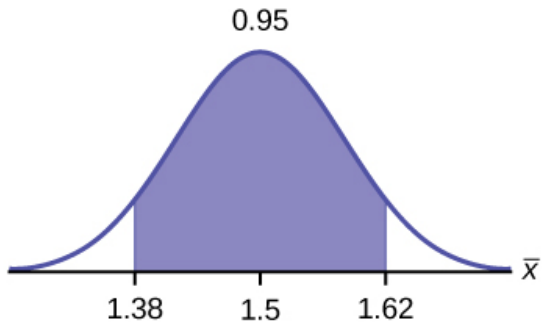


Figure 8.14

EBM = 0.12

43. a. $\overline{x}$ = 151
b. s_x = 32
c. n = 108
d. n – 1 = 107

45. $\overline{X}$ is the mean number of hours spent watching television per month from a sample of 108 Americans.

47. CI: (142.92, 159.08)

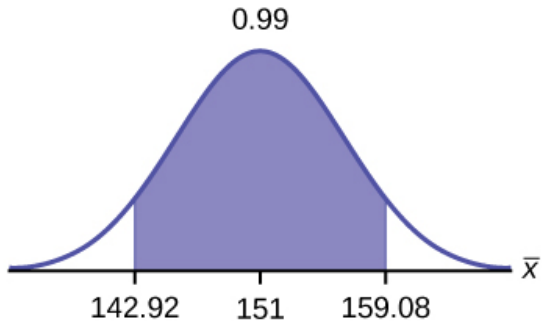


Figure 8.15

EBM = 8.08

49. a. 3.26

b. 1.02
c. 39

51. μ

53. t_{38}

55. 0.025

57. (2.93, 3.59)

59. We are 95 percent confident that the true mean number of colors for national flags is between 2.93 colors and 3.59 colors.

60. The error bound would become EBM = 0.245. This error bound decreases, because as sample sizes increase, variability decreases, and we need less interval length to capture the true mean.

63. It would decrease, because the *z*-score would decrease, which would reduce the numerator and lower the number.

65. *X* is the number of *successes* where the woman makes the majority of the purchasing decisions for the household. *P′* is the percentage of households sampled where the woman makes the majority of the purchasing decisions for the household.

67. CI: (0.5321, 0.6679)

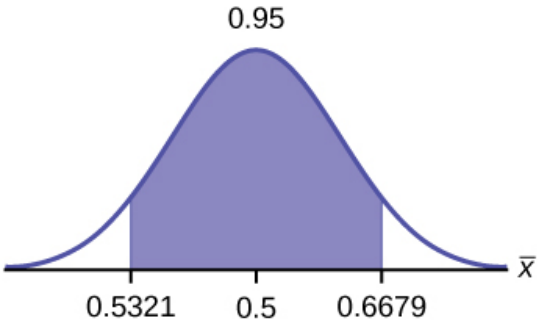


Figure 8.16

EBM: 0.0679

69. *X* is the number of *successes* where an executive prefers a truck. *P′* is the percentage of executives sampled who prefer a truck.

71. CI: (0.19432, 0.33068)

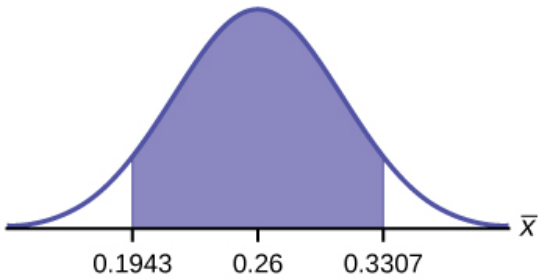


Figure 8.17

EBM: 0.0707

73. The sampling error means that the true mean can be 2 percent above or below the sample mean.

75. P' is the proportion of voters sampled who said the economy is the most important issue in the upcoming election.

77. CI: (0.62735, 0.67265);
EBM: 0.02265

79. the number of girls, ages 8 to 12, in the 5 p.m. Monday night beginning ice-skating class

81. a. $x = 64$
b. $n = 80$
c. $p' = 0.8$

83. p

85. $P' \sim N\left(0.8, \sqrt{\frac{(0.8)(0.2)}{80}}\right)$
CI = (0.72171, 0.87829).

87. 0.04

89. (0.72; 0.88)

91. With 92 percent confidence, we estimate the proportion of girls, ages 8 to 12, in a beginning ice-skating class at the Ice Chalet to be between 72 percent and 88 percent.

93. The error bound would increase. Assuming all other variables are kept constant, as the confidence level increases, the area under the curve corresponding to the confidence level becomes larger, which creates a wider interval and thus a larger error.

95. a. i. 71
ii. 3
iii. 48

b. X is the height of a Swedish male, and is the mean height from a sample of 48 Swedish males.
c. Normal. We know the standard deviation for the population, and the sample size is greater than 30.
d. i. CI: (70.151, 71.49)

ii.

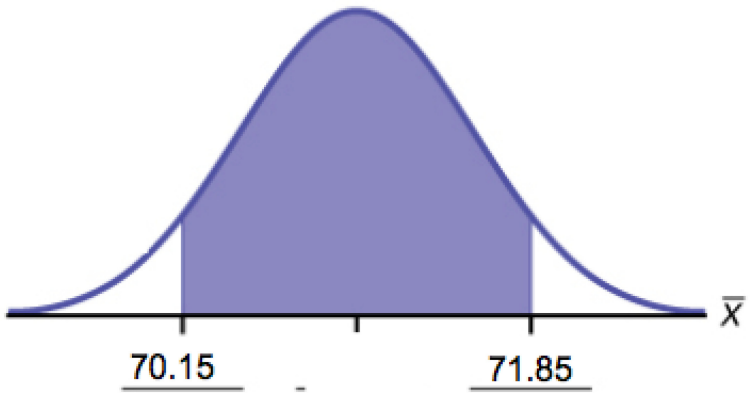


Figure 8.18

iii. EBM = 0.849

e. The confidence interval will decrease in size, because the sample size increased. Recall, when all factors remain unchanged, an increase in sample size decreases variability. Thus, we do not need as large an interval to capture the true population mean.

97. a. i. $\overline{x} = 23.6$

ii. $\sigma = 7$
iii. n = 100

b. X is the time needed to complete an individual tax form. $\overline{X}$ is the mean time to complete tax forms from a sample of 100 customers.

c. $N\left(23.6, \frac{7}{\sqrt{100}}\right)$ because we know sigma.

d. i. (22.228, 24.972)

ii.

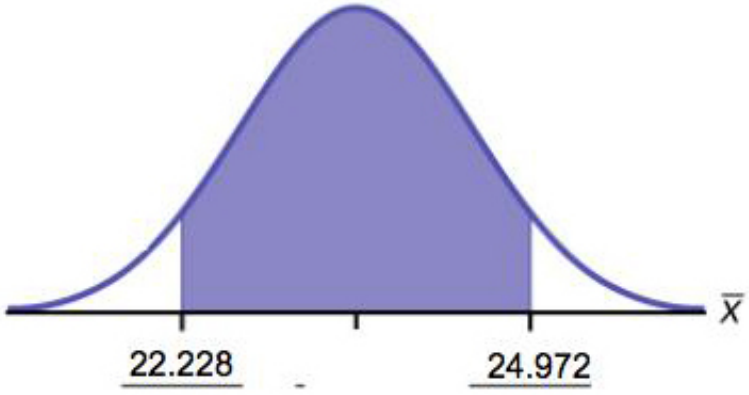


Figure 8.19

iii. *EBM* = 1.372

e. It will need to change the sample size. The firm needs to determine what the confidence level should be and then apply the error bound formula to determine the necessary sample size.

f. The confidence level would increase as a result of a larger interval. Smaller sample sizes result in more variability. To capture the true population mean, we need to have a larger interval.

g. According to the error bound formula, the firm needs to survey 206 people. Because we increase the confidence level, we need to increase either our error bound or the sample size.

99. a. i. 7.9
ii. 2.5
iii. 20

b. X is the number of letters a single camper will send home. $\overline{X}$ is the mean number of letters sent home from a sample of 20 campers.

c. $N\,7.9\left(\frac{2.5}{\sqrt{20}}\right)$

d. i. CI: (6.98, 8.82)

ii.

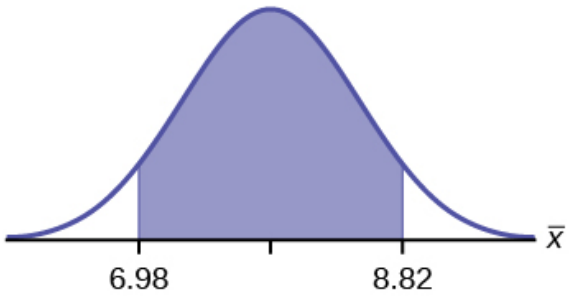


Figure 8.20

iii. *EBM*: 0.92

e. The error bound and confidence interval will decrease.

101. a. $\overline{x}$ = \$568,873

b. $CL = 0.95$, $\alpha = 1 - 0.95 = 0.05$, $z_{\frac{\alpha}{2}} = 1.96$

$EBM = z_{0.025}\frac{\sigma}{\sqrt{n}} = 1.96\ \frac{909200}{\sqrt{40}} = \$281{,}764$

c. $\overline{x} - EBM = 568{,}873 - 281{,}764 = 287{,}109$

$\overline{x} + EBM = 568{,}873 + 281{,}764 = 850{,}637$

Alternate solution:

USING THE TI-83, 83+, 84, 84+ CALCULATOR

1. Press STAT and arrow over to TESTS.
2. Arrow down to `7:ZInterval`.
3. Press ENTER.
4. Arrow to Stats and press ENTER.
5. Arrow down and enter the following values:
   - σ: 909,200
   - $\overline{x}$: 568,873
   - n: 40
   - CL: 0.95
6. Arrow down to Calculate and press ENTER.
7. The confidence interval is (\$287,114, \$850,632).
8. Notice the small difference between the two solutions—these differences are simply due to rounding error in the hand calculations.

d. We estimate with 95 percent confidence that the mean amount of contributions received from all individuals by House candidates is between \$287,109 and \$850,637.

103. Use the formula for EBM, solved for n:

$n = \frac{z^2\sigma^2}{EBM^2}$

From the statement of the problem, you know that $\sigma = 2.5$, and you need $EBM = 1$.

$z = z_{0.035} = 1.812$.

(This is the value of z for which the area under the density curve to the *right* of z is 0.035.)

$n = \frac{z^2\sigma^2}{EBM^2} = \frac{1.812^2 2.5^2}{1^2} \approx 20.52.$

You need to measure at least 21 male students to achieve your goal.

105. a. i. 8,629

ii. 6,944

iii. 35

iv. 34

b. t_{34}

c. i. CI: (6244, 11,014)

ii.

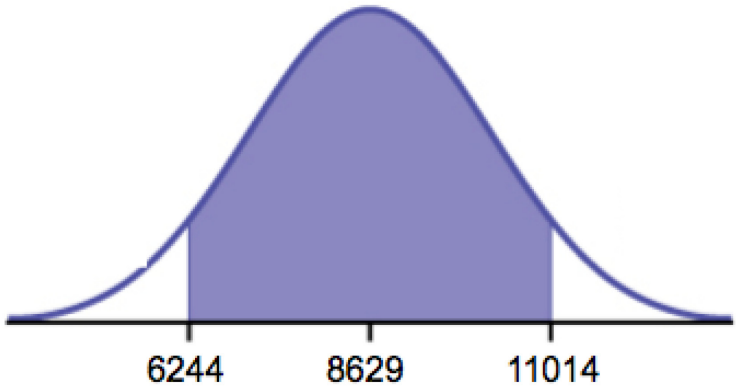


Figure 8.21

iii. EB = 2385

d. It will become smaller.

107. a. i. $\overline{x}$ = 2.51
ii. s_x = 0.318
iii. n = 9
iv. n - 1 = 8

b. The effective length of time for a tranquilizer
c. The mean effective length of time of tranquilizers from a sample of nine patients
d. We need to use a Student's t-distribution, because we do not know the population standard deviation.
e. i. CI: (2.27, 2.76)
ii. Check student's solution.
iii. *EBM*: 0.25

f. If we were to sample many groups of nine patients, 95 percent of the samples would contain the true population mean length of time.

109. $\overline{x} = \$251,854.23$;
$s = \$521,130.41$.
Note that we are not given the population standard deviation, only the standard deviation of the sample.
There are 30 measures in the sample, so n = 30, and df = 30 - 1 = 29.
CL = 0.96, so α = 1 - CL = 1 - 0.96 = 0.04.
$\frac{\alpha}{2} = 0.02 t_{\frac{\alpha}{2}} = t_{0.02}$ = 2.150.

$$EBM = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) = 2.150\left(\frac{521,130.41}{\sqrt{30}}\right) \sim \$204,561.66.$$

$\overline{x}$ - *EBM* = \$251,854.23 - \$204,561.66 = \$47,292.57.
$\overline{x}$ + *EBM* = \$251,854.23 + \$204,561.66 = \$456,415.89.
We estimate with 96 percent confidence that the mean amount of money raised by all Leadership PACs during the 2011–2012 election cycle lies between \$47,292.57 and \$456,415.89.
Alternate Solution

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Enter the data as a list.
Press

STAT

and arrow over to

TESTS

.
Arrow down to

8:TInterval

.
Press

ENTER

.
Arrow to Data and press

ENTER

.
Arrow down and enter the name of the list where the data are stored.
Enter

Freq

: 1
Enter

C-Level

: 0.96
Arrow down to

Calculate

and press

Enter

.
The 96 percent confidence interval is ($47,262, $456,447).

The difference between solutions arises from rounding differences.

111. a. i. $\overline{x}$ =
ii. s_x =
iii. n =
iv. n - 1 =

b. X is the number of unoccupied seats on a single flight. $\overline{X}$ is the mean number of unoccupied seats from a sample of 225 flights.
c. We will use a Student's t-distribution, because we do not know the population standard deviation.
d. i. CI: (11.12 , 12.08)
ii. Check student's solution.
iii. *EBM*: 0.48

113. a. i. CI: (7.64, 9.36)

ii.

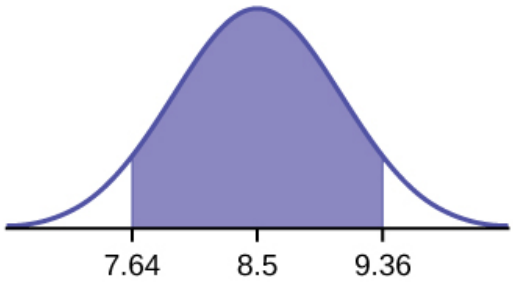


Figure 8.22

iii. *EBM*: 0.86

b. The sample should have been increased.
c. Answers will vary.
d. Answers will vary.
e. Answers will vary.

115. b

117. a. 1,068
b. The sample size would need to be increased, because the critical value increases as the confidence level increases.

119. a. X = the number of people who believe that the president is doing an acceptable job;

P' = the proportion of people in a sample who believe that the president is doing an acceptable job.

b. $N\left(0.61, \sqrt{\frac{(0.61)(0.39)}{1200}}\right)$

c. i. CI: (0.59, 0.63)
ii. Check student's solution.
iii. *EBM*: 0.02

121. a. i. (0.72, 0.82)
ii. (0.65, 0.76)
iii. (0.60, 0.72)

b. Yes, the intervals (0.72, 0.82) and (0.65, 0.76) overlap, and the intervals (0.65, 0.76) and (0.60, 0.72) overlap.
c. We can say that there does not appear to be a significant difference between the proportion of Asian adults who say that their families would welcome a white person into their families and the proportion of Asian adults who say that their families would welcome a Latino person into their families.
d. We can say that there is a significant difference between the proportion of Asian adults who say that their families would welcome a white person into their families and the proportion of Asian adults who say that their families would welcome a black person into their families.

123. a. X = the number of adult Americans who believe that crime is the main problem; P' = the proportion of adult Americans who believe that crime is the main problem.
b. Because we are estimating a proportion, that $P' = 0.2$ and $n = 1{,}000$, the distribution we should use is $N\left(0.2, \sqrt{\frac{(0.2)(0.8)}{1000}}\right)$.

c. i. CI: (0.18, 0.22)
ii. Check student's solution.
iii. *EBM*: 0.02

d. One way to lower the sampling error is to increase the sample size.
e. The stated *± 3 percent* represents the maximum error bound. This means that those doing the study are

reporting a maximum error of 3 percent. Thus, they estimate the percentage of adult Americans who the percentage of adult Americans who that crime is the main problem to be between 18 percent and 22 percent.

125. c

127. d

129. a

131. a. $p' = \frac{(0.55 + 0.49)}{2} = 0.52$; $EBP = 0.55 - 0.52 = 0.03$

b. No, the confidence interval includes values less than or equal to 0.50. It is possible that less than half of the population believe this.

c. $CL = 0.75$, so $\alpha = 1 - 0.75 = 0.25$ and $\frac{\alpha}{2} = 0.125$. $z_{\frac{\alpha}{2}} = 1.150$. (The area to the right of this z is 0.125, so the area to the left is 1 – 0.125 = 0.875.)

$$EBP = (1.150)\sqrt{\frac{0.52(0.48)}{1{,}000}} \approx 0.018$$

$(p' - EBP, p' + EBP) = (0.52 - 0.018, 0.52 + 0.018) = (0.502, 0.538)$

Alternate Solution

USING THE TI-83, 83+, 84, 84+ CALCULATOR

STAT TESTS A: 1-PropZinterval with $x = (0.52)(1{,}000)$, $n = 1{,}000$, CL = 0.75.

Answer is (0.502, 0.538).

d. Yes, this interval does not fall below 0.50, so we can conclude that at least half of all American adults believe that major sports programs corrupt education – but we do so with only 75 percent confidence.

133. $CL = 0.95$; $\alpha = 1 - 0.95 = 0.05$; $\frac{\alpha}{2} = 0.025$; $z_{\frac{\alpha}{2}} = 1.96$. Use $p' = q' = 0.5$.

$$n = \frac{z_{\frac{\alpha}{2}}^2 p' q'}{EBP^2} = \frac{1.96^2(0.5)(0.5)}{0.05^2} = 384.16.$$

You need to interview at least 385 students to estimate the proportion to within 5 percent at 95 percent confidence.

9

HYPOTHESIS TESTING WITH ONE SAMPLE

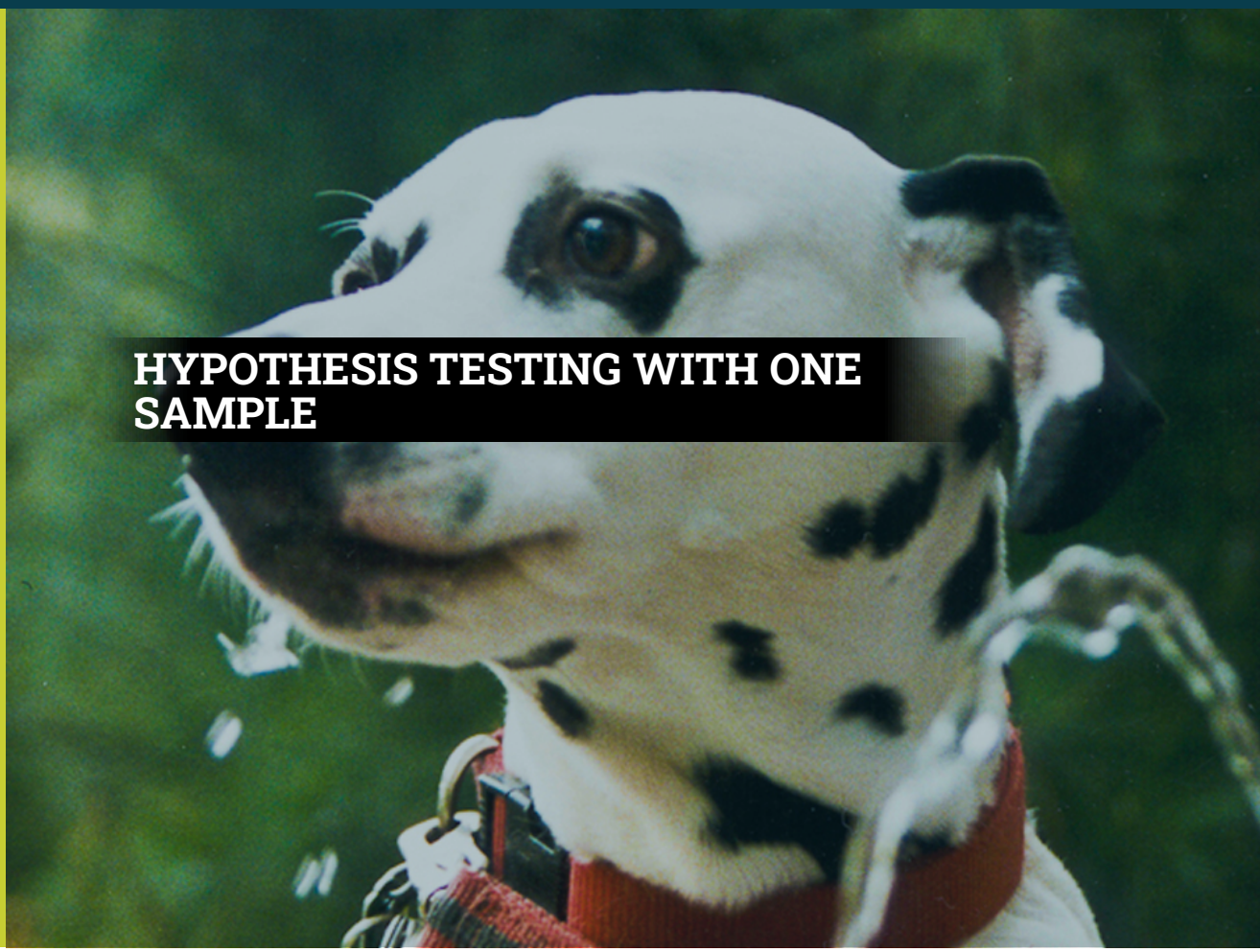


Figure 9.1 You can use a hypothesis test to decide if a dog breeder's claim that every Dalmatian has 35 spots is statistically sound. (credit: Robert Neff)

Chapter Objectives

By the end of this chapter, the student should be able to do the following:

- Differentiate between Type I and Type II errors
- Describe hypothesis testing in general and in practice
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation known
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation unknown
- Conduct and interpret hypothesis tests for a single population proportion

Introduction

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. **Confidence intervals** are one way to estimate a population parameter. Another way to make a statistical inference is to make a decision about a parameter. For instance, a car dealer advertises that its new small truck gets 35 miles per gallon, on average. A tutoring service claims that its method of tutoring helps 90 percent of its students get an A or a B. A company says that women managers in their company earn an average of $60,000 per year.

A statistician will make a decision about these claims. This process is called **hypothesis testing.** A hypothesis test involves collecting data from a sample and evaluating the data. Then, the statistician makes a decision as to whether or not there is sufficient evidence, based upon analyses of the data, to reject the null hypothesis.

In this chapter, you will conduct hypothesis tests on single means and single proportions. You will also learn about the

errors associated with these tests.

Hypothesis testing consists of two contradictory hypotheses or statements, a decision based on the data, and a conclusion. To perform a hypothesis test, a statistician will do the following:

1. Set up two contradictory hypotheses.
2. Collect sample data. In homework problems, the data or summary statistics will be given to you.
3. Determine the correct distribution to perform the hypothesis test.
4. Analyze sample data by performing the calculations that ultimately will allow you to reject or decline to reject the null hypothesis.
5. Make a decision and write a meaningful conclusion.

Note

To do the hypothesis test homework problems for this chapter and later chapters, make copies of the appropriate special solution sheets. See Appendix E.

9.1 Null and Alternative Hypotheses

The actual test begins by considering two **hypotheses**. They are called the **null hypothesis** and the **alternative hypothesis**. These hypotheses contain opposing viewpoints.

H_0, the —**null hypothesis:** a statement of no difference between sample means or proportions or no difference between a sample mean or proportion and a population mean or proportion. In other words, the difference equals 0.

H_a—, the **alternative hypothesis:** a claim about the population that is contradictory to H_0 and what we conclude when we reject H_0.

Since the null and alternative hypotheses are contradictory, you must examine evidence to decide if you have enough evidence to reject the null hypothesis or not. The evidence is in the form of sample data.

After you have determined which hypothesis the sample supports, you make a **decision.** There are two options for a decision. They are *reject* H_0 if the sample information favors the alternative hypothesis or *do not reject* H_0 or *decline to reject* H_0 if the sample information is insufficient to reject the null hypothesis.

Mathematical Symbols Used in H_0 and H_a:

H_0	H_a
equal (=)	not equal (≠) **or** greater than (>) **or** less than (<)
greater than or equal to (≥)	less than (<)
less than or equal to (≤)	more than (>)

Table 9.1

Note

H_0 always has a symbol with an equal in it. H_a never has a symbol with an equal in it. The choice of symbol depends on the wording of the hypothesis test. However, be aware that many researchers use = in the null hypothesis, even with > or < as the symbol in the alternative hypothesis. This practice is acceptable because we only make the decision to reject or not reject the null hypothesis.

EXAMPLE 9.1

H_0: No more than 30 percent of the registered voters in Santa Clara County voted in the primary election. $p \leq 30$

H_a: More than 30 percent of the registered voters in Santa Clara County voted in the primary election. $p > 30$

TRY IT 9.1

A medical trial is conducted to test whether or not a new medicine reduces cholesterol by 25 percent. State the null and alternative hypotheses.

EXAMPLE 9.2

We want to test whether the mean GPA of students in American colleges is different from 2.0 (out of 4.0). The null and alternative hypotheses are the following:
H_0: $\mu = 2.0$
H_a: $\mu \neq 2.0$

TRY IT 9.2

We want to test whether the mean height of eighth graders is 66 inches. State the null and alternative hypotheses. Fill in the correct symbol (=, ≠, ≥, <, ≤, >) for the null and alternative hypotheses.

a. H_0: μ __ 66
b. H_a: μ __ 66

EXAMPLE 9.3

We want to test if college students take fewer than five years to graduate from college, on the average. The null and alternative hypotheses are the following:
H_0: $\mu \geq 5$
H_a: $\mu < 5$

TRY IT 9.3

We want to test if it takes fewer than 45 minutes to teach a lesson plan. State the null and alternative hypotheses. Fill in the correct symbol (=, ≠, ≥, <, ≤, >) for the null and alternative hypotheses.

a. H_0: μ __ 45
b. H_a: μ __ 45

EXAMPLE 9.4

An article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams and a third of the students pass. The same article stated that 6.6 percent of U.S. students take advanced placement exams and 4.4 percent pass. Test if the percentage of U.S. students who take advanced placement exams is more than 6.6 percent. State the null and alternative hypotheses.
H_0: $p \leq 0.066$
H_a: $p > 0.066$

TRY IT 9.4

On a state driver's test, about 40 percent pass the test on the first try. We want to test if more than 40 percent pass on the first try. Fill in the correct symbol (=, ≠, ≥, <, ≤, >) for the null and alternative hypotheses.

a. H_0: p __ 0.40
b. H_a: p __ 0.40

COLLABORATIVE EXERCISE

Bring to class a newspaper, some news magazines, and some internet articles. In groups, find articles from which your group can write null and alternative hypotheses. Discuss your hypotheses with the rest of the class.

9.2 Outcomes and the Type I and Type II Errors

When you perform a hypothesis test, there are four possible outcomes depending on the actual truth, or falseness, of the null hypothesis H_0 and the decision to reject or not. The outcomes are summarized in the following table:

ACTION	H_0 IS ACTUALLY	...
	True	False
Do not reject H_0	Correct outcome	Type II error
Reject H_0	Type I error	Correct outcome

Table 9.2

The four possible outcomes in the table are as follows:

1. The decision is *not to reject* H_0 when ***H_0* is true** (correct decision).
2. The decision is to *reject* H_0 when, in fact, ***H_0 is true*** (incorrect decision known as a **Type I error**).
3. The decision is *not to reject* H_0 when, in fact, ***H_0* is false** (incorrect decision known as a **Type II error**).
4. The decision is to *reject* H_0 when ***H_0* is false** (correct decision whose probability is called the **Power of the Test**).

Each of the errors occurs with a particular probability. The Greek letters α and β represent the probabilities.

α = probability of a Type I error = ***P*(Type I error)** = probability of rejecting the null hypothesis when the null hypothesis is true.

β = probability of a Type II error = ***P*(Type II error)** = probability of not rejecting the null hypothesis when the null hypothesis is false.

α and β should be as small as possible because they are probabilities of errors. They are rarely zero.

The Power of the Test is 1 – β. Ideally, we want a high power that is as close to one as possible. Increasing the sample size can increase the Power of the Test.

The following are examples of Type I and Type II errors.

EXAMPLE 9.5

Suppose the null hypothesis, H_0, is: Frank's rock climbing equipment is safe.

Type I error: Frank does not go rock climbing because he considers that the equipment is not safe, when in fact, the equipment is really safe. Frank is making the mistake of rejecting the null hypothesis, when the equipment is actually safe!

Type II error: Frank goes climbing, thinking that his equipment is safe, but this is a mistake, and he painfully realizes that his equipment is not as safe as it should have been. Frank assumed that the null hypothesis was true, when it was not.

α = *probability* that Frank thinks his rock climbing equipment may not be safe when, in fact, it really is safe. β = *probability* that Frank thinks his rock climbing equipment may be safe when, in fact, it is not safe.

Notice that, in this case, the error with the greater consequence is the Type II error. (If Frank thinks his rock climbing equipment is safe, he will go ahead and use it.)

TRY IT 9.5

Suppose the null hypothesis, H_0, is: the blood cultures contain no traces of pathogen *X*. State the Type I and Type II errors.

EXAMPLE 9.6

Suppose the null hypothesis, H_0, is: a tomato plant is alive when a class visits the school garden.

Type I error: The null hypothesis claims that the tomato plant is alive, and it is true, but the students make the mistake of thinking that the plant is already dead.

Type II error: The tomato plant is already dead (the null hypothesis is false), but the students do not notice it, and believe that the tomato plant is alive.

α = *probability* that the class thinks the tomato plant is dead when, in fact, it is alive = *P*(Type I error). β = *probability* that the class thinks the tomato plant is alive when, in fact, it is dead = *P*(Type II error).

The error with the greater consequence is the Type I error. (If the class thinks the plant is dead, they will not water it.)

TRY IT 9.6

Suppose the null hypothesis, H_0, is: a patient is not sick. Which type of error has the greater consequence, Type I or Type II?

EXAMPLE 9.7

It's a Boy Genetic Labs, a genetics company, claims to be able to increase the likelihood that a pregnancy will result in a boy being born. Statisticians want to test the claim. Suppose that the null hypothesis, H_0, is: It's a Boy Genetic Labs has no effect on gender outcome.

Type I error: This error results when a true null hypothesis is rejected. In the context of this scenario, we would state that we believe that It's a Boy Genetic Labs influences the gender outcome, when in fact it has no effect. The probability of this error occurring is denoted by the Greek letter alpha, α.

Type II error: This error results when we fail to reject a false null hypothesis. In context, we would state that It's a Boy Genetic Labs does not influence the gender outcome of a pregnancy when, in fact, it does. The probability of this error occurring is denoted by the Greek letter beta, β.

The error with the greater consequence would be the Type I error since couples would use the It's a Boy Genetic Labs product in hopes of increasing the chances of having a boy.

TRY IT 9.7

Red tide is a bloom of poison-producing algae—a few different species of a class of plankton called dinoflagellates. When the weather and water conditions cause these blooms, shellfish such as clams living in the area develop dangerous levels of a paralysis-inducing toxin. In Massachusetts, the Division of Marine Fisheries montors levels of the toxin in shellfish by regular sampling of shellfish along the coastline. If the mean level of toxin in clams exceeds 800 μg (micrograms) of toxin per kilogram of clam meat in any area, clam harvesting is banned there until the bloom is over and levels of toxin in clams subside. Describe both a Type I and a Type II error in this context, and state which error has the greater consequence.

EXAMPLE 9.8

A certain experimental drug claims a cure rate of at least 75 percent for males with a disease. Describe both the Type I and Type II errors in context. Which error is the more serious?

Type I: A patient believes the cure rate for the drug is less than 75 percent when it actually is at least 75 percent.

Type II: A patient believes the experimental drug has at least a 75 percent cure rate when it has a cure rate that is less than 75 percent.

In this scenario, the Type II error contains the more severe consequence. If a patient believes the drug works at least 75 percent of the time, this most likely will influence the patient's (and doctor's) choice about whether to use the drug as a treatment option.

TRY IT 9.8

Determine both Type I and Type II errors for the following scenario:

Assume a null hypothesis, H_0, that states the percentage of adults with jobs is at least 88 percent.

Identify the Type I and Type II errors from these four possible choices.

a. Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88 percent when that percentage is actually less than 88 percent
b. Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88 percent when the percentage is actually at least 88 percent
c. Reject the null hypothesis that the percentage of adults who have jobs is at least 88 percent when the percentage is actually at least 88 percent
d. Reject the null hypothesis that the percentage of adults who have jobs is at least 88 percent when that percentage is actually less than 88 percent

9.3 Distribution Needed for Hypothesis Testing

Earlier in the course, we discussed sampling distributions. *Particular distributions are associated with hypothesis testing.* Perform tests of a population mean using a **normal distribution** or a **Student's *t*-distribution**. (Remember, use a Student's *t*-distribution when the population **standard deviation** is unknown and the distribution of the sample mean is approximately normal.) We perform tests of a population proportion using a normal distribution (usually n is large).

Assumptions

When you perform *a hypothesis test of a single population mean* μ using a **Student's *t*-distribution** (often called a *t*-test), there are fundamental assumptions that need to be met in order for the test to work properly. Your data should be a **simple random sample** that comes from a population that is approximately **normally distributed**. You use the sample **standard deviation** to approximate the population standard deviation. Note that if the sample size is sufficiently large, a *t*-test will work even if the population is not approximately normally distributed.

When you perform a *hypothesis test of a single population mean* μ using a normal distribution (often called a *z*-test), you take a simple random sample from the population. The population you are testing is normally distributed or your sample size is sufficiently large. You know the value of the population standard deviation which, in reality, is rarely known.

When you perform a *hypothesis test of a single population proportion p*, you take a simple random sample from the population. You must meet the conditions for a **binomial distribution**, which are the following: there are a certain number *n* of independent trials, the outcomes of any trial are success or failure, and each trial has the same probability of a success *p*. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities *np* and *nq* must both be greater than five ($np > 5$ and $nq > 5$). Then the binomial distribution of a sample (estimated) proportion can be approximated by the normal distribution with $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$. Remember that $q = 1 - p$.

9.4 Rare Events, the Sample, and the Decision and Conclusion

Establishing the type of distribution, sample size, and known or unknown standard deviation can help you figure out how to go about a hypothesis test. However, there are several other factors you should consider when working out a hypothesis test.

Rare Events

The thinking process in hypothesis testing can be summarized as follows: You want to test whether or not a particular property of the population is true. You make an assumption about the true population mean for numerical data or the true population proportion for categorical data. This assumption is the null hypothesis. Then you gather sample data that is representative of the population. From this sample data you compute the sample mean (or the sample proportion). If the value that you observe is very unlikely to occur (a rare event) if the null hypothesis is true, then you wonder why this is happening. A plausible explanation is that the null hypothesis is false.

For example, Didi and Ali are at a birthday party of a very wealthy friend. They hurry to be first in line to grab a prize from a tall basket that they cannot see inside because they will be blindfolded. There are 200 plastic bubbles in the basket, and Didi and Ali have been told that there is only one with a $100 bill. Didi is the first person to reach into the basket and pull out a bubble. Her bubble contains a $100 bill. The probability of this happening is $\frac{1}{200} = 0.005$. Because this is so unlikely, Ali is hoping that what the two of them were told is wrong and there are more $100 bills in the basket. A *rare event* has occurred (Didi getting the $100 bill) so Ali doubts the assumption about only one $100 bill being in the basket.

Using the Sample to Test the Null Hypothesis

After you collect data and obtain the test statistic (the sample mean, sample proportion, or other test statistic), you can determine the probability of obtaining that test statistic when the null hypothesis is true. This probability is called the ***p*-value**.

When the *p*-value is very small, it means that the observed test statistic is very unlikely to happen if the null hypothesis is true. This gives significant evidence to suggest that the null hypothesis is false, and to reject it in favor of the alternative hypothesis. In practice, to reject the null hypothesis we want the *p*-value to be smaller than 0.05 (5 percent) or sometimes even smaller than 0.01 (1 percent).

EXAMPLE 9.9

Suppose a baker claims that his bread height is more than 15 cm, on average. Several of his customers do not believe him. To persuade his customers that he is right, the baker decides to do a hypothesis test. He bakes 10 loaves of bread. The mean height of the sample loaves is 17 cm. The baker knows from baking hundreds of loaves of bread that the **standard deviation** for the height is 0.5 cm and the distribution of heights is normal.

The null hypothesis could be H_0: $\mu \leq 15$. The alternate hypothesis is H_a: $\mu > 15$.

The words *is more than* translates as a ">" so "$\mu > 15$" goes into the alternate hypothesis. The null hypothesis must contradict the alternate hypothesis.

Since σ *is known* ($\sigma = 0.5$ cm), the distribution for the population is known to be normal with mean $\mu = 15$ and standard deviation $\frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{10}} = 0.16$.

Suppose the null hypothesis is true (which is that the mean height of the loaves is no more than 15 cm). Then is the mean height (17 cm) calculated from the sample unexpectedly large? The hypothesis test works by asking the question

how *unlikely* the sample mean would be if the null hypothesis were true. The graph shows how far out the sample mean is on the normal curve. The *p*-value is the probability that, if we were to take other samples, any other sample mean would fall at least as far out as 17 cm.

The p-value, then, is the probability that a sample mean is the same or greater than 17 cm when the population mean is, in fact, 15 cm. We can calculate this probability using the normal distribution for means. In Figure 9.2, the *p*-value is the area under the normal curve to the right of 17. Using a normal distribution table or a calculator, we can compute that this probability is practically zero.

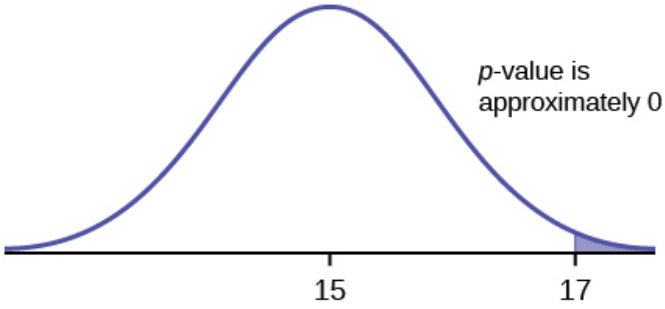


Figure 9.2

p-value = $P(\bar{x} > 17)$, which is approximately zero.

Because the *p*-value is almost 0, we conclude that obtaining a sample height of 17 cm or higher from 10 loaves of bread is very unlikely if the true mean height is 15 cm. We reject the null hypothesis and conclude that there is sufficient evidence to claim that the true population mean height of the baker's loaves of bread is higher than 15 cm.

TRY IT 9.9

A normal distribution has a standard deviation of 1. We want to verify a claim that the mean is greater than 12. A sample of 36 is taken with a sample mean of 12.5.

H_0: $\mu \leq 12$
H_a: $\mu > 12$
The *p*-value is 0.0013.
Draw a graph that shows the *p*-value.

Decision and Conclusion

A systematic way to make a decision of whether to reject or not reject the **null hypothesis** is to compare the *p*-value and a preset or preconceived α, also called the **level of significance of the test**. A preset α is the probability of a **Type I error** (rejecting the null hypothesis when the null hypothesis is true). It may or may not be given to you at the beginning of the problem.

When you make a *decision* to reject or not reject H_0, do as follows:

- If *p*-value $< \alpha$, reject H_0. The results of the sample data are significant. There is sufficient evidence to conclude that H_0 is an incorrect belief and that the **alternative hypothesis**, H_a, may be correct.
- If *p*-value $\geq \alpha$, do not reject H_0. The results of the sample data are not significant.There is not sufficient evidence to conclude that the alternative hypothesis, H_a, may be correct.
- When you *do not reject* H_0, it does not mean that you should believe that H_0 is true. It simply means that the sample data have *failed* to provide sufficient evidence to cast serious doubt about the truthfulness of H_0.

Conclusion: After you make your decision, write a thoughtful *conclusion* about the hypotheses in terms of the given problem.

EXAMPLE 9.10

When using the *p*-value to evaluate a hypothesis test, you might find it useful to use the following mnemonic device:

If the *p*-value is low, the null must go.

If the *p*-value is high, the null must fly.

This memory aid relates a *p*-value less than the established alpha (the *p* is low) as rejecting the null hypothesis and, likewise, relates a *p*-value higher than the established alpha (the *p* is high) as not rejecting the null hypothesis.

Problem

Fill in the blanks.

Reject the null hypothesis when _____.

The results of the sample data _____.

Do not reject the null hypothesis when _____.

The results of the sample data _____.

Solution

Reject the null hypothesis when *the p-value is less than the established alpha value*. The results of the sample data **support the alternative hypothesis**.

Do not reject the null hypothesis when *the p-value is greater or equal to the established alpha value*. The results of the sample data **do not support the alternative hypothesis**.

TRY IT 9.10

It's a Boy Genetics Labs, a genetics company, claims their procedures improve the chances of a boy being born. The results for a test of a single population proportion are as follows:

H_0: $p = 0.50$, H_a: $p > 0.50$

$\alpha = 0.01$

p-value = 0.025

Interpret the results and state a conclusion in simple, nontechnical terms.

9.5 Additional Information and Full Hypothesis Test Examples

- In a **hypothesis test** problem, you may see words such as "the level of significance is 1 percent". The "1 percent" is the preconceived or preset α.
- The statistician setting up the hypothesis test selects the value of α to use *before* collecting the sample data.
- *If no level of significance is given, a common standard to use is $\alpha = 0.05$.*
- When you calculate the *p*-value and draw the picture, the *p*-value is the area in the left tail, the right tail, or split evenly between the two tails. For this reason, we call the hypothesis test left, right, or two tailed.
- The **alternative hypothesis**, H_a, tells you if the test is left, right, or two-tailed. It is the *key* to conducting the appropriate test.
- H_a never has a symbol that contains an equal sign.
- *Thinking about the meaning of the* ***p*-value**: A data analyst should have more confidence that he made the correct decision to reject the null hypothesis with a smaller *p*-value (for example, 0.001 as opposed to 0.04) even if using the 0.05 level for alpha. Similarly, for a large *p*-value such as 0.4, as opposed to a *p*-value of 0.056 (alpha = 0.05 is less than either number), a data analyst should have more confidence that she made the correct decision in not rejecting the null hypothesis. This makes the data analyst use judgment rather than mindlessly applying rules.

The following examples illustrate a left-, right-, and two-tailed test.

EXAMPLE 9.11

H_0: $\mu = 5$ H_a: $\mu < 5$

Test of a single population mean. H_a tells you the test is left-tailed. The picture of the *p*-value is as follows:

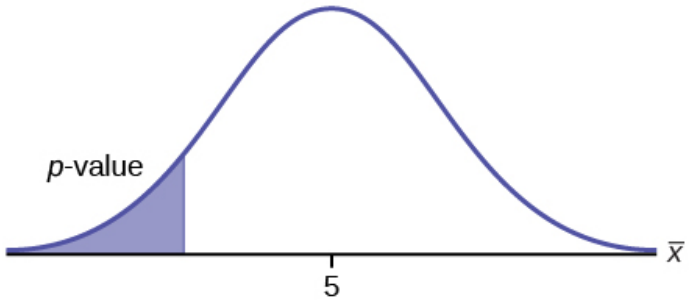


Figure 9.3

TRY IT 9.11

H_0: $\mu = 10$ H_a: $\mu < 10$

Assume the *p*-value is 0.0935. What type of test is this? Draw the picture of the *p*-value.

EXAMPLE 9.12

H_0: $p \leq 0.2$ H_a: $p > 0.2$

This is a test of a single population proportion. H_a tells you the test is **right-tailed**. The picture of the *p*-value is as follows:

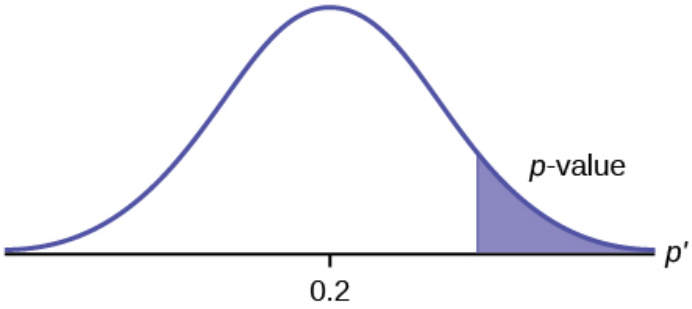


Figure 9.4

TRY IT 9.12

H_0: $\mu \leq 1$ H_a: $\mu > 1$

Assume the *p*-value is 0.1243. What type of test is this? Draw the picture of the *p*-value.

EXAMPLE 9.13

H_0: $p = 50$ H_a: $p \neq 50$

This is a test of a single population mean. H_a tells you the test is **two-tailed**. The picture of the *p*-value is as follows.

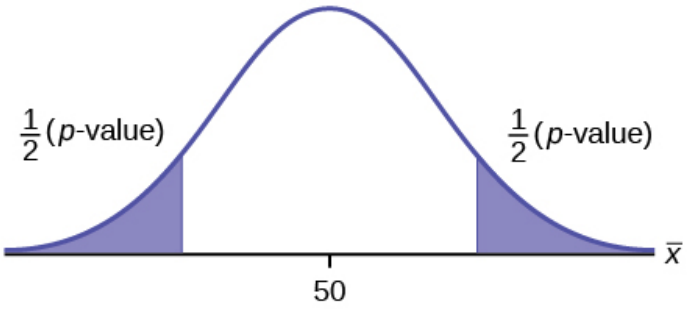


Figure 9.5

TRY IT 9.13

H_0: $p = 0.5$ H_a: $p \neq 0.5$

Assume the *p*-value is 0.2564. What type of test is this? Draw the picture of the *p*-value.

Full Hypothesis Test Examples

EXAMPLE 9.14

Problem

Jeffrey, as an eight-year-old, established a mean time of 16.43 seconds for swimming the 25-yard freestyle, with a standard deviation of 0.8 seconds. His dad, Frank, thought that Jeffrey could swim the 25-yard freestyle faster using goggles. Frank bought Jeffrey a new pair of expensive goggles and timed Jeffrey for 15 25-yard freestyle swims. For the 15 swims, Jeffrey's mean time was 16 seconds. Frank thought that the goggles helped Jeffrey to swim faster than the 16.43 seconds. Conduct a hypothesis test using a preset $\alpha = 0.05$. Assume that the swim times for the 25-yard freestyle are normal.

Solution

Set up the hypothesis test:

Since the problem is about a mean, this is a *test of a single population mean*.

H_0: $\mu = 16.43$ H_a: $\mu < 16.43$

For Jeffrey to swim faster, his time will be less than 16.43 seconds. The "<" tells you this is left-tailed.

Determine the distribution needed:

Random variable: X = the mean time to swim the 25-yard freestyle.

Distribution for the test: X is normal (population standard deviation is known: $\sigma = 0.8$)

with mean $\mu = 16.43$ and standard error of $\frac{0.8}{\sqrt{15}}$;

$\mu = 16.43$ comes from H_0 and not the data. $\sigma = 0.8$, and $n = 15$.

Using a table or a calculator, we can calculate the *p*-value as the area to the left of 16 under the normal curve:

p-value = $P(\bar{x} < 16) = 0.0187$ where the sample mean in the problem is given as 16.

p-value = 0.0187. The *p*-value is the area to the left of the sample mean given as 16.

Graph:

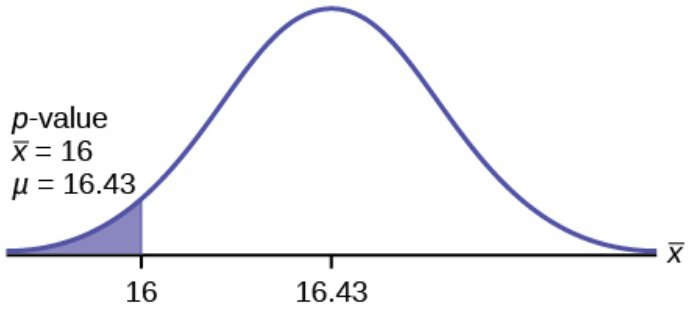


Figure 9.6

μ = 16.43 comes from H_0. Our assumption is μ = 16.43.

Interpretation of the *p*-value: *If H_0 is true*, there is a 0.0187 probability (1.87 percent), that Jeffrey's mean time to swim the 25-yard freestyle is 16 seconds or less. Because a 1.87 percent chance is small, the mean time of 16 seconds or less is unlikely to have happened randomly. It is a rare event.

Compare α and the *p*-value:

α = 0.05 *p*-value = 0.0187 α > *p*-value

Make a decision: Since α > *p*-value, reject H_0.

An alternative approach is to find the *z*-test corresponding to the sample mean $\overline{x} = 16$. This is

$$z\text{-test} = \frac{\overline{x} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} = \frac{16 - 16.43}{\frac{0.8}{\sqrt{15}}} = -2.081729.$$

The critical *z*-value = –1.645 for this test has probability 0.05 to its left tail, according to the Normal Table (see Appendices). Because the *z*-test is *to the left* of the critical *z*-value, we reject the null hypothesis.

This means that you reject μ = 16.43. In other words, you do not think Jeffrey swims the 25-yard freestyle in 16.43 seconds but instead that he swims faster with the new goggles.

Conclusion: At the 5 percent significance level, we conclude that Jeffrey swims faster using the new goggles. The sample data show there is sufficient evidence that Jeffrey's mean time to swim the 25-yard freestyle is less than 16.43 seconds.

The *p*-value can easily be calculated.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Press STAT and arrow over to TESTS. Press `1:z-Test`. Arrow over to `Stats` and press ENTER. Arrow down and enter 16.43 for μ_0 (null hypothesis), .8 for σ, 16 for the sample mean, and 15 for *n*. Arrow down to μ : (alternate hypothesis) and arrow over to < μ_0. Press ENTER. Arrow down to `Calculate` and press ENTER. The calculator not only calculates the *p*-value (*p* = 0.0187) but it also calculates the test statistic (*z*-score) for the sample mean. μ < 16.43 is the alternative hypothesis. Do this set of instructions again except arrow to `Draw`(instead of `Calculate`). Press ENTER. A shaded graph appears with *z* = -2.08 (test statistic) and *p* = 0.0187 (*p*-value). Make sure when you use `Draw` that no other equations are highlighted in *Y* = and the plots are turned off.

When the calculator does a *z*-Test, the `z-Test` function finds the *p*-value by doing a normal probability calculation:

$P(\overline{x} < 16) =$ `2nd DISTR normcdf` $\left(-10\ \text{^}\ 99, 16, 16.43, 0.8/\sqrt{15}\right)$.

The Type I and Type II errors for this problem are as follows:

The Type I error is to conclude that Jeffrey swims the 25-yard freestyle, on average, in less than 16.43 seconds when, in fact, he actually swims the 25-yard freestyle, on average, in 16.43 seconds. (Reject the null hypothesis when the null hypothesis is true.)

The Type II error is that there is not evidence to conclude that Jeffrey swims the 25-yard freestyle, on average, in less than 16.43 seconds when, in fact, he actually does swim the 25-yard freestyle, on average, in less than 16.43 seconds. (Do not reject the null hypothesis when the null hypothesis is false.)

Historical Note (Example 9.11)

The traditional way to compare the two probabilities, α and the p-value, is to compare the critical value (z-score from α) to the test statistic (z-score from data). The calculated test statistic for the p-value is –2.08. (From the central limit theorem, the test statistic formula is $z = \frac{\overline{x} - \mu_X}{(\frac{\sigma_X}{\sqrt{n}})}$. For this problem, $\overline{x}$ = 16, μ_X = 16.43 from the null hypothesis, σ_X = 0.8, and n = 15.) You can find the critical value for α = 0.05 in the normal table (see Appendix H: Tables). The z-score for an area to the left equal to 0.05 is midway between –1.65 and –1.64 (0.05 is midway between 0.0505 and 0.0495). The z-score is –1.645. Since –1.645 > –2.08 (which demonstrates that α > p-value), reject H_0. Traditionally, the decision to reject or not reject was done in this way. Today, comparing the two probabilities α and the p-value is very common. For this problem, the p-value, 0.0187, is considerably smaller than α, 0.05. You can be confident about your decision to reject. The graph shows α, the p-value, and the test statistic and the critical value.

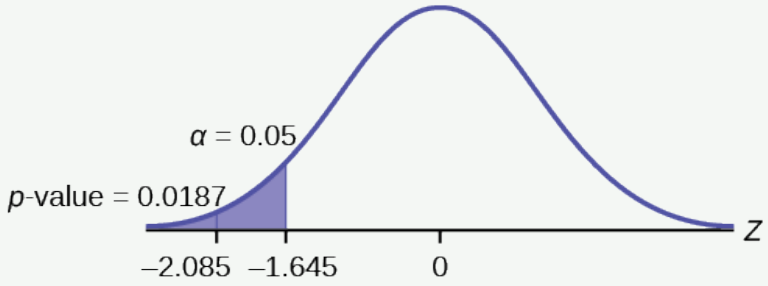


Figure 9.7

TRY IT 9.14

The mean throwing distance of a football by Marco, a high school freshman quarterback, is 40 yards, with a standard deviation of two yards. The team coach tells Marco to adjust his grip to get more distance. The coach records the distances for 20 throws. For the 20 throws, Marco's mean distance was 45 yards. The coach thought the different grip helped Marco throw farther than 40 yards. Conduct a hypothesis test using a preset α = 0.05. Assume the throw distances for footballs are normal.

First, determine what type of test this is, set up the hypothesis test, find the p-value, sketch the graph, and state your conclusion.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Press STAT and arrow over to TESTS. Press `1: z-Test`. Arrow over to `Stats` and press ENTER. Arrow down and enter 40 for μ_0 (null hypothesis), 2 for σ, 45 for the sample mean, and 20 for n. Arrow down to μ: (alternative hypothesis) and set it either as <, ≠, or >. Press ENTER. Arrow down to `Calculate` and press ENTER. The calculator not only calculates the p-value but it also calculates the test statistic (z-score) for the sample mean. Select <, ≠, or >

for the alternative hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press ENTER. A shaded graph appears with test statistic and *p*-value. Make sure when you use `Draw` that no other equations are highlighted in *Y* = and the plots are turned off.

EXAMPLE 9.15

Problem

A college football coach records the mean weight that his players can bench press as 275 pounds, with a standard deviation of 55 pounds. Three of his players thought that the mean weight was more than that amount. They asked 30 of their teammates for their estimated maximum lift on the bench press exercise. The data ranged from 205 pounds to 385 pounds. The actual different weights were (frequencies are in parentheses) 205(3); 215(3); 225(1); 241(2); 252(2); 265(2); 275(2); 313(2); 316(5); 338(2); 341(1); 345(2); 368(2); 385(1).

Conduct a hypothesis test using a 2.5 percent level of significance to determine if the bench press mean is more than 275 pounds.

Solution

Set up the hypothesis test:

Since the problem is about a mean weight, this is a *test of a single population mean*.

H_0: $\mu = 275$ H_a: $\mu > 275$ This is a right-tailed test.

Calculating the distribution needed:

Random variable: $\overline{X}$ = the mean weight, in pounds, lifted by the football players.

Distribution for the test: It is normal because σ is known.

$$\overline{X} \sim N\left(275, \frac{55}{\sqrt{30}}\right)$$

$\overline{x} = 286.2$ pounds (from the data).

$\sigma = 55$ pounds. *Always use σ if you know it.* We assume $\mu = 275$ pounds unless our data shows us otherwise.

First, we compute the sample mean:

$$\overline{x} = \frac{205 + 205 + 205 + 215 + \cdots + 385}{30} = 286.2.$$

Next, we compute the *z*-test:

$$z\text{-test} = \frac{286.2 - 275}{\frac{55}{\sqrt{30}}} = 1.115362$$

Finally, the *p*-value is the probability to the right tail of the *z*-test, which we can compute from the table of *z*-scores as 0.5 -- 0.36650 = 0.1335.

$$p\text{-value} = P(\overline{x} > 286.2) = 0.1323$$

Interpretation of the *p*-value: If H_0 is true, then there is a 0.1331 probability, 13.23 percent, that the football players can lift a mean weight of 286.2 pounds or more. Because a 13.23 percent chance is large enough, a mean weight lift of 286.2 pounds or more is not a rare event.

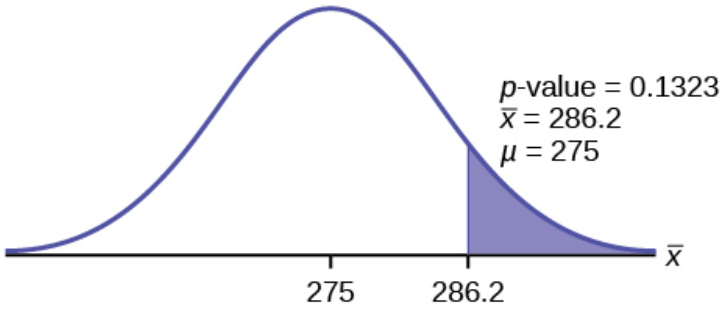


Figure 9.8

Compare α and the p-value:

$$\alpha = 0.025$$
$$p\text{-value} = 0.1323$$

Make a decision: Since $\alpha < p$-value, do not reject H_0.

Conclusion: At the 2.5 percent level of significance, from the sample data, there is not sufficient evidence to conclude that the true mean weight lifted is more than 275 pounds.

The p-value can easily be calculated.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Put the data and frequencies into lists. Press STAT and arrow over to TESTS. Press `1:z-Test`. Arrow over to `Data` and press ENTER. Arrow down and enter 275 for μ_0, 55 for σ, the name of the list where you put the data, and the name of the list where you put the frequencies. Arrow down to μ: and arrow over to $> \mu_0$. Press ENTER. Arrow down to `Calculate` and press ENTER. The calculator not only calculates the p-value ($p = 0.1331$, a little different from the previous calculation—in it we used the sample mean rounded to one decimal place instead of the data), but also the test statistic (z-score) for the sample mean, the sample mean, and the sample standard deviation. $\mu > 275$ is the alternative hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press ENTER. A shaded graph appears with $z = 1.112$ (test statistic) and $p = 0.1331$ (p-value). Make sure when you use `Draw` that no other equations are highlighted in $Y =$ and the plots are turned off.

EXAMPLE 9.16

Problem

Statistics students believe that the mean score on the first statistics test is 65. A statistics instructor thinks the mean score is higher than 65. He samples 10 statistics students and obtains the scores 65; 65; 70; 67; 66; 63; 63; 68; 72; 71. He performs a hypothesis test using a 5 percent level of significance. The data are assumed to be from a normal distribution.

Solution

Set up the hypothesis test:

A 5 percent level of significance means that $\alpha = 0.05$. This is a test of a **single population mean**.

H_0: $\mu = 65$ H_a: $\mu > 65$

Since the instructor thinks the average score is higher, use a ">". The ">" means the test is right-tailed.

Determine the distribution needed:

Random variable: $\overline{X}$ = average score on the first statistics test.

Distribution for the test: If you read the problem carefully, you will notice that there is *no population standard deviation given*. You are only given $n = 10$ sample data values. Notice also that the data come from a normal distribution. This means that the distribution for the test is a Student's *t*-distribution.

Use *t*-distribution. Therefore, the distribution for the test is *t* with nine degrees of freedom.

Calculate the *p*-value using the Student's *t*-distribution:

First, we compute the sample mean as

$$\overline{x} = \frac{65 + 65 + \cdots + 71}{10} = 67.$$

Next, we compute the *t*-test as

$$t\text{-test} = \frac{\overline{x} - \mu_X}{\frac{s_X}{\sqrt{n}}} = \frac{67 - 65}{\frac{3.12}{\sqrt{10}}} \approx 1.98.$$

The *p*-value is the probability to the right tail of 1.98 in a *t*-distribution with nine degrees of freedom.

p-value = $P(\overline{x} > 67) = 0.0396$ where the sample mean and sample standard deviation are calculated as 67 and 3.1972 from the data.

Interpretation of the *p*-value: If the null hypothesis is true, then there is a 0.0396 probability— (3.96 percent—) that the sample mean is 65 or more.

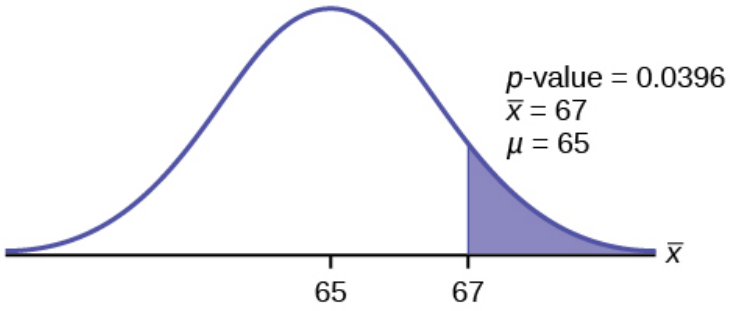


Figure 9.9

Compare α and the *p*-value:

Since $\alpha = 0.05$ and *p*-value = 0.0396, $\alpha >$ *p*-value.

Make a decision: Since $\alpha >$ *p*-value, reject H_0.

Alternatively, according to a Student's *t*-distribution table (see Appendices), the critical *t*-value is 1.833. Since the *t*-test (1.98) is to the right of the critical *t*-value 1.833, we reject the null hypothesis.

This decision means we reject $\mu = 65$. In other words, we believe the average test score is more than 65.

Conclusion: At a 5 percent level of significance, the sample data show sufficient evidence that the mean (average) test score is more than 65, just as the math instructor thinks.

The *p*-value can easily be calculated.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Put the data into a list. Press STAT and arrow over to TESTS. Press `2:T-Test`. Arrow over to `Data` and press ENTER. Arrow down and enter 65 for μ_0, the name of the list where you put the data, and 1 for `Freq:`. Arrow down to μ: and arrow over to $> \mu_0$. Press ENTER. Arrow down to `Calculate` and press ENTER. The calculator not only calculates the *p*-value ($p = 0.0396$) but it also calculates the test statistic (*t*-score) for the sample mean, the sample mean, and the sample standard deviation. $\mu > 65$ is the alternative hypothesis. Do this set of instructions again except arrow to `Draw`

(instead of Calculate). Press ENTER. A shaded graph appears with $t = 1.9781$ (test statistic) and $p = 0.0396$ (p-value). Make sure when you use Draw that no other equations are highlighted in $Y =$ and the plots are turned off.

TRY IT 9.16

It is believed that a stock price for a particular company will grow at a rate of $5 per week with a standard deviation of $1. An investor believes the stock won't grow as quickly. The changes in stock price are recorded for 10 weeks and are as follows: $4, $3, $2, $3, $1, $7, $2, $1, $1, $2. Perform a hypothesis test using a 5 percent level of significance. State the null and alternative hypotheses, find the p-value, state your conclusion, and identify the Type I and Type II errors.

EXAMPLE 9.17

Problem

Joon believes that 50 percent of first-time brides in the United States are younger than their grooms. She performs a hypothesis test to determine if the percentage is the same or different from 50 percent. Joon samples 100 first-time brides and 53 reply that they are younger than their grooms. For the hypothesis test, she uses a 1 percent level of significance.

Solution

Set up the hypothesis test:

The 1 percent level of significance means that $\alpha = 0.01$. This is a *test of a single population proportion*.

H_0: $p = 0.50$ H_a: $p \neq 0.50$

The words ***is the same or different from*** tell you this is a two-tailed test.

Calculate the distribution needed:

Random variable: P' = the percentage of first-time brides who are younger than their grooms.

Distribution for the test: The problem contains no mention of a mean. The information is given in terms of percentages. Use the distribution for P', the estimated proportion.

P' follows a normal distribution with mean value $\mu = p$, and standard error $\sigma = \sqrt{\frac{p \cdot q}{n}}$.

In our example, $p = q = 0.5$, and $n = 100$,
where $p = 0.50$, $q = 1 - p = 0.50$, and $n = 100$.

Calculate the p-value using the normal distribution for proportions:

First, we compute the sample proportion as $\hat{p} = \frac{53}{100} = 0.53$.

Next, the z-test is given by

$$z\text{-test} = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{0.53 - 0.50}{\sqrt{\frac{0.50 \times 0.50}{100}}} = 0.6.$$

Since the z-test is positive, we compute the area to the right tail of 0.6 in a normal distribution, $P(Z > 0.6) = 0.2742531$. Finally, because this is a two-sided test of significance, we multiply this probability times two to account for the left tail, and obtain

$$p\text{-value} = 2 \times 0.2742531 = 0.5485062$$

where $x = 53$, $p' = \frac{x}{n} = \frac{53}{100} = 0.53$.

Interpretation of the *p*-value: If the null hypothesis is true, there is 0.5485 probability, (54.85 percent) that the sample

(estimated) proportion p' is 0.53 or more OR 0.47 or less (see the graph in Figure 9.10).

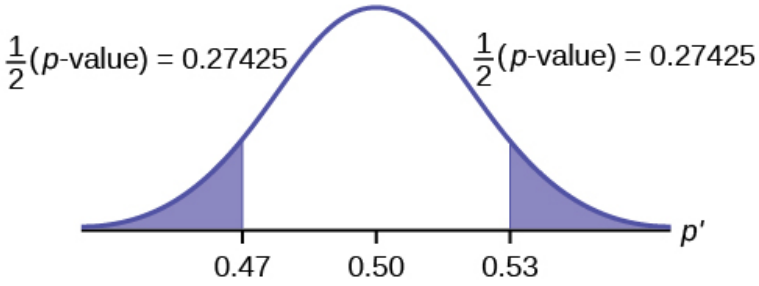


Figure 9.10

$\mu = p = 0.50$ comes from H_0, the null hypothesis.

$p' = 0.53$. Since the curve is symmetrical and the test is two-tailed, the p' for the left tail is equal to 0.50 – 0.03 = 0.47 where $\mu = p = 0.50$. (0.03 is the difference between 0.53 and 0.50.)

Compare α and the p-value:

Since $\alpha = 0.01$ and p-value = 0.5485, $\alpha < p$-value.

Make a decision: Since $\alpha < p$-value, you cannot reject H_0.

Conclusion: At the 1 percent level of significance, the sample data do not show sufficient evidence that the percentage of first-time brides who are younger than their grooms is different from 50 percent.

The p-value can easily be calculated.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Press STAT and arrow over to TESTS. Press `5:1-PropZTest`. Enter .5 for p_0, 53 for x and 100 for n. Arrow down to `Prop` and arrow to `not equals` p_0. Press ENTER. Arrow down to `Calculate` and press ENTER. The calculator calculates the p-value ($p = 0.5485$) and the test statistic (z-score). `Prop not equals` .5 is the alternate hypothesis. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press ENTER. A shaded graph appears with $z = 0.6$ (test statistic) and $p = 0.5485$ (p-value). Make sure when you use `Draw` that no other equations are highlighted in $Y =$ and the plots are turned off.

The Type I and Type II errors are as follows:

The Type I error is to conclude that the proportion of first-time brides who are younger than their grooms is different from 50 percent when, in fact, the proportion is actually 50 percent. Reject the null hypothesis when the null hypothesis is true.

The Type II error is there is not enough evidence to conclude that the proportion of first-time brides who are younger than their grooms differs from 50 percent when, in fact, the proportion does differ from 50 percent. Do not reject the null hypothesis when the null hypothesis is false.

TRY IT 9.17

A teacher believes that 85 percent of students in the class will want to go on a field trip to the local zoo. She performs a hypothesis test to determine if the percentage is the same or different from 85 percent. The teacher samples 50 students and 39 reply that they would want to go to the zoo. For the hypothesis test, use a 1 percent level of significance.

First, determine what type of test this is, set up the hypothesis test, find the p-value, sketch the graph, and state your conclusion.

EXAMPLE 9.18

Problem

Suppose a consumer group suspects that the proportion of households that have three cell phones is 30 percent. A cell phone company has reason to believe that the proportion is not 30 percent. Before the cell phone company starts a big advertising campaign, it conducts a hypothesis test. The company's marketing people survey 150 households with the result that 43 of the households have three cell phones.

a. The value that helps determine the *p*-value is p'. Calculate p'.
b. What is a *success* for this problem?
c. What is the level of significance?
d. Draw the graph for this problem. Draw the horizontal axis. Label and shade appropriately. Calculate the *p*-value.
e. Make a decision. _____(Reject/Do not reject) H_0 because_____.

Solution

Set up the hypothesis test:

H_0: $p = 0.30$ H_a: $p \neq 0.30$

Determine the distribution needed:

The **random variable** is P' = proportion of households that have three cell phones.

The **distribution** for the hypothesis test is $P' \sim N\left(0.30, \sqrt{\frac{(0.30)\cdot(0.70)}{150}}\right)$.

a. $p' = \frac{x}{n}$ where x is the number of successes and n is the total number in the sample.

$$x = 43, \; n = 150$$

$$p' = \frac{43}{150.}$$

b. A success is having three cell phones in a household.
c. The level of significance is the preset α. Since α is not given, assume that $\alpha = 0.05$.
d. First we compute the sample proportion $\hat{p} = \frac{43}{150} = 0.287$.

Next, the *z*-test is given by

$$z\text{-test} = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{0.287 - 0.30}{\sqrt{\frac{0.30 \times 0.70}{150}}} \approx -0.36.$$

Since the *z*-test is negative, we compute the area to the left tail of –0.36 in a normal distribution, $P(Z < -0.36) \approx 0.3607902$. Finally, because this is a two-sided test of significance, we multiply this probability times two to account for the right tail, and obtain $p\text{-value} = 2 \times 0.3607902 = 0.7215804$.

e. Assuming that $\alpha = 0.05$, $\alpha < p$-value. The decision is *do not reject* H_0 because there is not sufficient evidence to conclude that the proportion of households that have three cell phones is not 30 percent.

TRY IT 9.18

Marketers believe that 92 percent of adults in the United States own a cell phone. A cell phone manufacturer believes that number is actually lower. Two hundred American adults are surveyed, of which 174 report having cell phones. Use a 5 percent level of significance. State the null and alternative hypotheses, find the *p*-value, state your conclusion, and identify the Type I and Type II errors.

The next example is a poem written by a statistics student named Nicole Hart. The solution to the problem follows the poem. Notice that the hypothesis test is for a single population proportion. This means that the null and alternate hypotheses use the parameter *p*. The distribution for the test is normal. The estimated proportion p' is the proportion of

fleas killed to the total fleas found on Fido. This is sample information. The problem gives a preconceived $\alpha = 0.01$, for comparison, and a 95 percent confidence interval computation. The poem is clever and humorous, so please enjoy it!

EXAMPLE 9.19

Problem

My dog has so many fleas,
They do not come off with ease.
As for shampoo, I have tried many types
Even one called Bubble Hype,
Which only killed 25 percent of the fleas,
Unfortunately I was not pleased.

I've used all kinds of soap,
Until I had given up hope
Until one day I saw
An ad that put me in awe.

A shampoo used for dogs
Called GOOD ENOUGH to Clean a Hog
Guaranteed to kill more fleas.

I gave Fido a bath
And after doing the math
His number of fleas
Started dropping by 3's!

Before his shampoo
I counted 42.
At the end of his bath,
I redid the math
And the new shampoo had killed 17 fleas.
So now I was pleased.

Now it is time for you to have some fun
With the level of significance being .01,
You must help me figure out
Use the new shampoo or go without?

Solution

Set up the hypothesis test:

H_0: $p \leq 0.25$ H_a: $p > 0.25$

Determine the distribution needed:

In words, *clearly* state what your random variable $\overline{X}$ or P' represents.

P' = The proportion of fleas that are killed by the new shampoo

State the distribution to use for the test.

Normal:

$$N\left(0.25, \sqrt{\frac{(0.25)(1-0.25)}{42}}\right)$$

The z-test is given by

$$z\text{-test} = \frac{\overline{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{0.4048 - 0.25}{\sqrt{42}} \approx 2.316834.$$

Because this is a hypothesis test one-sided to the right, we compute the *p*-value as the area to the right tail of the *z*-test in a standard normal distribution, $P(Z > 3.32) \approx 0.0103$.

In one to two complete sentences, explain what the *p*-value means for this problem.

If the null hypothesis is true (the proportion is 0.25), then there is a 0.0103 probability that the sample (estimated) proportion is 0.4048 $\left(\frac{17}{42}\right)$ or more.

Use the previous information to sketch a picture of this situation. *Clearly* label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.

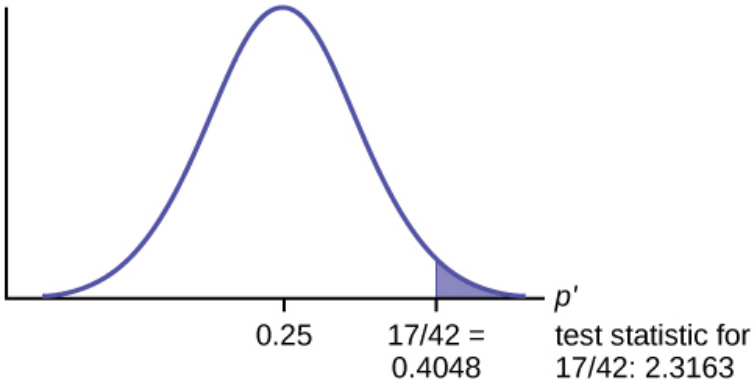


Figure 9.11

Compare α and the *p*-value:

Indicate the correct decision (*reject* or *do not reject* the null hypothesis) and the reason for it, and write an appropriate conclusion, using complete sentences.

Alpha	Decision	Reason for Decision
0.01	Do not reject H_0	α < *p*-value

Table 9.3

Conclusion: At the 1 percent level of significance, the sample data do not show sufficient evidence that the percentage of fleas that are killed by the new shampoo is more than 25 percent.

Construct a 95 percent confidence interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the confidence interval.

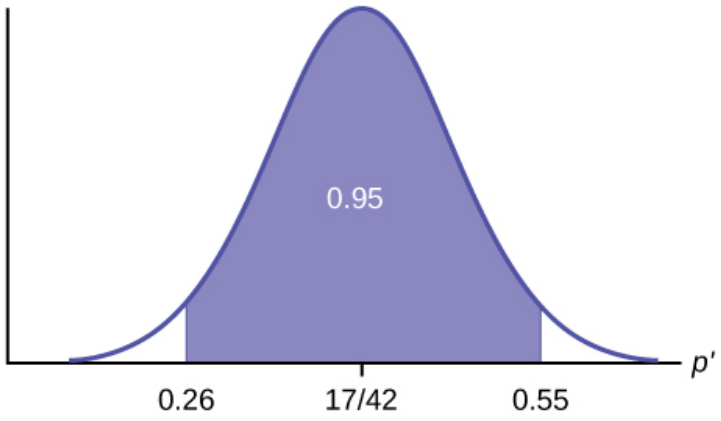


Figure 9.12

Confidence Interval: (0.26, 0.55). We are 95 percent confident that the true population proportion *p* of fleas that are

killed by the new shampoo is between 26 percent and 55 percent.

Note

This test result is not very definitive since the *p*-value is very close to alpha. In reality, one would probably do more tests by giving the dog another bath after the fleas have had a chance to return.

EXAMPLE 9.20

Problem

The National Institute of Standards and Technology provides exact data on conductivity properties of materials. Following are conductivity measurements for 11 randomly selected pieces of a particular type of glass:

1.11, 1.07, 1.11, 1.07, 1.12, 1.08, 0.98, 0.98, 1.02, 0.95, 0.95

Is there convincing evidence that the average conductivity of this type of glass is greater than one? Use a significance level of 0.05. Assume the population is normal.

Solution

Let's follow a four-step process to answer this statistical question.

1. **State the question**: We need to determine if, at a 0.05 significance level, the average conductivity of the selected glass is greater than one. Our hypotheses will be as follows:
   a. H_0: $\mu \leq 1$
   b. H_a: $\mu > 1$
2. **Plan**: We are testing a sample mean without a known population standard deviation. Therefore, we need to use a Student's *t*-distribution. Assume the underlying population is normal.
3. **Do the calculations**: We will input the sample data into the TI-83 as follows.

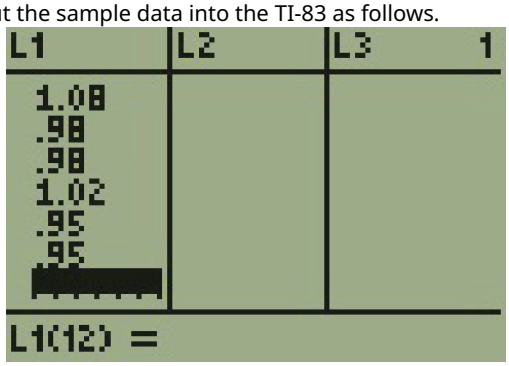


Figure 9.13

Figure 9.14

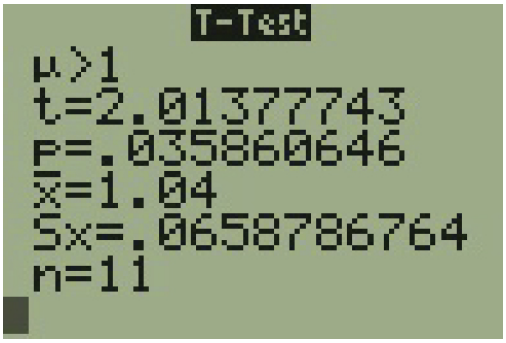


Figure 9.15

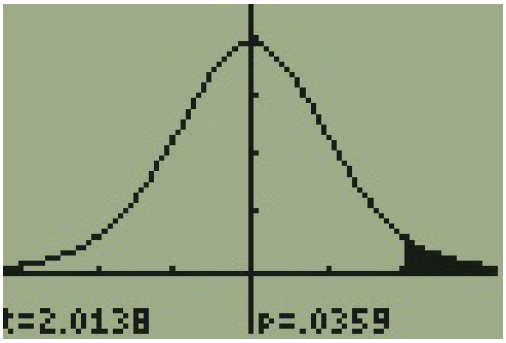


Figure 9.16

4. **State the conclusions**: Since the *p*-value ($p = 0.036$) is less than our alpha value, we will reject the null hypothesis. It is reasonable to state that the data support the claim that the average conductivity level is greater than one.

EXAMPLE 9.21

Problem

In a study of 420,019 cell phone users, 172 of the subjects developed brain cancer. Test the claim that cell phone users developed brain cancer at a greater rate than that for non-cell phone users. The rate of brain cancer for non-cell phone users is 0.0340 percent. Since this is a critical issue, use a 0.005 significance level. Explain why the significance level should be so low in terms of a Type I error.

Solution

We will follow the four-step process.

1. We need to conduct a hypothesis test on the claimed cancer rate. Our hypotheses will be as follows:
   a. H_0: $p \leq 0.00034$
   b. H_a: $p > 0.00034$

   If we commit a Type I error, we are essentially accepting a false claim. Since the claim describes cancer-causing environments, we want to minimize the chances of incorrectly identifying causes of cancer.
2. We will be testing a sample proportion with $x = 172$ and $n = 420{,}019$. The sample is sufficiently large because we have $np = 420{,}019(0.00034) = 142.8$, $nq = 420{,}019(0.99966) = 419{,}876.2$, two independent outcomes, and a fixed probability of success $p = 0.00034$. Thus we will be able to generalize our results to the population.
3. The associated TI results are shown in the following figures.

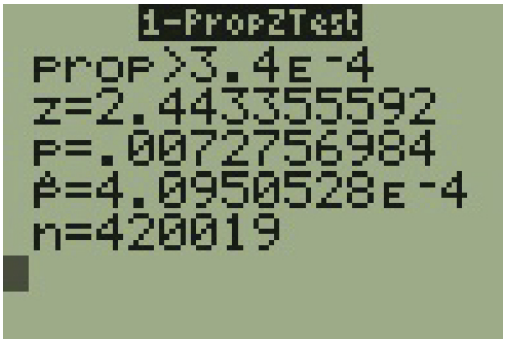


Figure 9.17

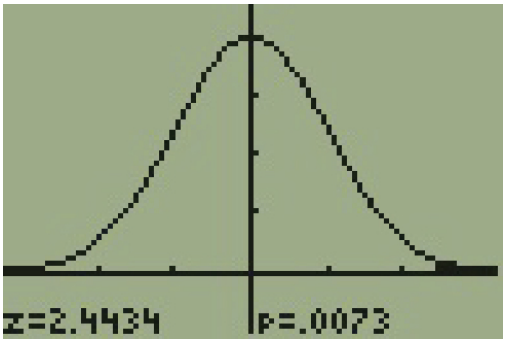


Figure 9.18

4. Since the *p*-value = 0.0073 is greater than our alpha value = 0.005, we cannot reject the null. Therefore, we conclude that there is not enough evidence to support the claim of higher brain cancer rates for the cell phone users.

9.6 Hypothesis Testing of a Single Mean and Single Proportion

Stats Lab

Hypothesis Testing of a Single Mean and Single Proportion

Student Learning Outcomes

- The student will select the appropriate distributions to use in each case.
- The student will conduct hypothesis tests and interpret the results.

Television Survey

In a recent survey, it was stated that Americans watch television on average four hours per day. Assume that $\sigma = 2$. Using your class as the sample, conduct a hypothesis test to determine if the average for students at your school is lower.

1. H_0: _____
2. H_a: _____
3. In words, define the random variable. _____ = _____
4. The distribution to use for the test is _____.
5. Determine the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.
   a. Graph:

Figure 9.19

b. Determine the *p*-value.

7. Do you or do you not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Language Survey

About 42.3 percent of Californians and 19.6 percent of all Americans over age five speak a language other than English at home. Using your class as the sample, conduct a hypothesis test to determine if the percentage of the students at your school who speak a language other than English at home is different from 42.3 percent.

1. H_0: _____
2. H_a: _____
3. In words, define the random variable. _____ = _____
4. The distribution to use for the test is _____.
5. Determine the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.
   a. Graph:

Figure 9.20

b. Determine the *p*-value.

7. Do you or do you not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Jeans Survey

You've read in an article that young adults own an average of three pairs of jeans. Survey eight people from your class to determine if the average is higher than three. Assume the population is normal.

1. H_0: _____
2. H_a: _____
3. In words, define the random variable. _____ = _____
4. The distribution to use for the test is _____.

5. Determine the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.
   a. Graph:

Figure 9.21

   b. Determine the *p*-value.
7. Do you or do you not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Key Terms

binomial distribution a discrete random variable (RV) that arises from Bernoulli trials; there are a fixed number, *n*, of independent trials
Independent means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in *n* trials. The notation is: $X \sim B(n, p)$ $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly *x* successes in *n* trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$.

confidence interval (*CI*) an interval estimate for an unknown population parameter
This depends on the following:
- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

hypothesis a statement about the value of a population parameter; in the case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternative hypothesis (notation H_a)

hypothesis testing based on sample evidence, a procedure for determining whether the hypothesis stated is a reasonable statement and should not be rejected, or is unreasonable and should be rejected

level of significance of the test probability of a Type I error (reject the null hypothesis when it is true)
Notation: α. In hypothesis testing, the level of significance is called the preconceived α or the preset α.

normal distribution
a bell-shaped continuous random variable *X*, with center at the mean value (μ) and distance from the center to the inflection points of the bell curve given by the standard deviation (σ)
We write $X \sim N(\mu, \sigma)$. If the mean value is 0 and the standard deviation is 1, the random variable is called the standard normal distribution, and it is denoted with the letter *Z*.

***p*-value** the probability that an event will happen purely by chance assuming the null hypothesis is true; the smaller the *p*-value, the stronger the evidence is against the null hypothesis

standard deviation a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: *s* for sample standard deviation and σ for population standard deviation

Student's *t*-distribution investigated and reported by William S. Gosset in 1908 and published under the pseudonym Student
The major characteristics of the random variable (RV) are as follows
- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as *n* gets larger.
- There is a *family* of *t*-distributions: every representative of the family is completely defined by the number of degrees of freedom, which is one less than the number of data items.

Type 1 error the decision is to reject the null hypothesis when, in fact, the null hypothesis is true

Type 2 error the decision is not to reject the null hypothesis when, in fact, the null hypothesis is false

Chapter Review

9.1 Null and Alternative Hypotheses

In a **hypothesis test**, sample data are evaluated in order to arrive at a decision about some type of claim. If certain conditions about the sample are satisfied, then the claim can be evaluated for a population. In a hypothesis test, we do the following:

1. Evaluate the **null hypothesis**, typically denoted with H_0. The null is not rejected unless the hypothesis test shows otherwise. The null statement must always contain some form of equality (=, ≤, or ≥).
2. Always write the **alternative hypothesis**, typically denoted with H_a or H_1, using less than, greater than, or not equals symbols, i.e., (≠, >, or <).
3. If we reject the null hypothesis, then we can assume there is enough evidence to support the alternative hypothesis.
4. Never state that a claim is proven true or false. Keep in mind the underlying fact that hypothesis testing is based on probability laws; therefore, we can talk only in terms of non-absolute certainties.

9.2 Outcomes and the Type I and Type II Errors

In every hypothesis test, the outcomes are dependent on a correct interpretation of the data. Incorrect calculations or misunderstood summary statistics can yield errors that affect the results. A **Type I** error occurs when a true null hypothesis is rejected. A **Type II error** occurs when a false null hypothesis is not rejected.

The probabilities of these errors are denoted by the Greek letters α and β, for a Type I and a Type II error respectively. The power of the test, $1 - \beta$, quantifies the likelihood that a test will yield the correct result of a true alternative hypothesis being accepted. A high power is desirable.

9.3 Distribution Needed for Hypothesis Testing

In order for a hypothesis test's results to be generalized to a population, certain requirements must be satisfied.

When testing for a single population mean:

1. A Student's t-test should be used if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large, with an unknown standard deviation.
2. The normal test will work if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large, with a known standard deviation.

When testing a single population proportion use a normal test for a single population proportion if the data come from a simple, random sample, fill the requirements for a binomial distribution, and the mean number of success and the mean number of failures satisfy the conditions: $np > 5$ and $nq > n$ where n is the sample size, p is the probability of a success, and q is the probability of a failure.

9.4 Rare Events, the Sample, and the Decision and Conclusion

When the probability of an event occurring is low, and it happens, it is called a rare event. Rare events are important to consider in hypothesis testing because they can inform your willingness not to reject or to reject a null hypothesis. To test a null hypothesis, find the p-value for the sample data and graph the results. When deciding whether or not to reject the null the hypothesis, keep these two parameters in mind:

1. $\alpha > p$-value, reject the null hypothesis.
2. $\alpha \leq p$-value, do not reject the null hypothesis.

9.5 Additional Information and Full Hypothesis Test Examples

The **hypothesis test** itself has an established process. This can be summarized as follows:

1. Determine H_0 and H_a. Remember, they are contradictory.
2. Determine the random variable.
3. Determine the distribution for the test.
4. Draw a graph, calculate the test statistic, and use the test statistic to calculate the p-value. (A z-score and a t-score are examples of test statistics.)
5. Compare the preconceived α with the p-value, make a decision (reject or do not reject H_0), and write a clear conclusion using English sentences.

Notice that in performing the hypothesis test, you use α and not β. β is needed to help determine the sample size of the data that are used in calculating the p-value. Remember that the quantity $1 - \beta$ is called the **Power of the Test**. A high power is desirable. If the power is too low, statisticians typically increase the sample size while keeping α the same. If the power is low, the null hypothesis might not be rejected when it should be.

Formula Review

9.1 Null and Alternative Hypotheses

H_0 and H_a are contradictory.

If H_0 has:	equal (=)	greater than or equal to (≥)	less than or equal to (≤)
then H_a has:	not equal (≠) **or** greater than (>) **or** less than (<)	less than (<)	greater than (>)

Table 9.4

If $\alpha \leq p$-value, then do not reject H_0.

If $\alpha > p$-value, then reject H_0.

α is preconceived. Its value is set before the hypothesis test starts. The p-value is calculated from the data.

9.2 Outcomes and the Type I and Type II Errors

α = probability of a Type I error = P(Type I error) = probability of rejecting the null hypothesis when the null hypothesis is true.

β = probability of a Type II error = P(Type II error) = probability of not rejecting the null hypothesis when the null hypothesis is false.

9.3 Distribution Needed for Hypothesis Testing

If there is no given preconceived α, then use $\alpha = 0.05$.

Types of Hypothesis Tests

- Single population mean, *known* population variance (or standard deviation): **Normal test**.
- Single population mean, *unknown* population variance (or standard deviation): **Student's *t*-test**.
- Single population proportion: **Normal test**.
- For a *single population mean*, we may use a normal distribution with the following mean and standard deviation. Means: $\mu = \mu_{\bar{x}}$ and $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$.
- For a *single population proportion*, we may use a normal distribution with the following mean and standard deviation. Proportions: $\mu = \boldsymbol{p}$ and $\sigma = \sqrt{\frac{pq}{n}}$.

Practice

9.1 Null and Alternative Hypotheses

1. You are testing that the mean speed of your cable internet connection is more than three megabits per second. What is the random variable? Describe it in words.

2. You are testing that the mean speed of your cable internet connection is more than three megabits per second. State the null and alternative hypotheses.

3. The American family has an average of two children. What is the random variable? Describe in words.

4. The mean entry level salary of an employee at a company is $58,000. You believe it is higher for IT professionals in the company. State the null and alternative hypotheses.

5. A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the proportion is actually less. What is the random variable? Describe in words.

6. A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the claim is correct. State the null and alternative hypotheses.

7. In a population of fish, approximately 42 percent are female. A test is conducted to see if, in fact, the proportion is less. State the null and alternative hypotheses.

8. Suppose that a recent article stated that the mean time students spend doing homework each week is 2.5 hours. A study was then done to see if the mean time has increased in the new century. A random sample of 26 students. The mean length of time the students spent on homework was 3 hours with a standard deviation of 1.8 hours. Suppose that it is somehow known that the population standard deviation is 1.5. If you were conducting a hypothesis test to determine if the mean length of homework has increased, what would the null and alternative hypotheses be? The distribution of the population is normal.

a. H_0: _____
b. H_a: _____

9. A random survey of 75 long-term marathon runners revealed that the mean length of time they've been running is 17.4 years with a standard deviation of 6.3 years. If you were conducting a hypothesis test to determine if the population mean time for these runners could likely be 15 years, what would the null and alternative hypotheses be?

a. H_0: _____
b. H_a: _____

10. Researchers published an article stating that in any one-year period, approximately 9.5 percent of American adults suffer from a particular type of disease. Suppose that in a survey of 100 people in a certain town, seven of them suffered from this disease. If you were conducting a hypothesis test to determine if the true proportion of people in that town suffering from this disease is lower than the percentage in the general adult American population, what would the null and alternative hypotheses be?

a. H_0: _____
b. H_a: _____

9.2 Outcomes and the Type I and Type II Errors

11. The mean price of mid-sized cars in a region is $32,000. A test is conducted to see if the claim is true. State the Type I and Type II errors in complete sentences.

12. A sleeping bag is tested to withstand temperatures of –15 °F. You think the bag cannot stand temperatures that low. State the Type I and Type II errors in complete sentences.

13. For Exercise 9.12, what are α and β in words?

14. In words, describe 1 – β for Exercise 9.12.

15. A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis, H_0, is: the surgical procedure will go well. State the Type I and Type II errors in complete sentences.

16. A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis, H_0, is: the surgical procedure will go well. Which is the error with the greater consequence?

17. The power of a test is 0.981. What is the probability of a Type II error?

18. A group of divers is exploring an old sunken ship. Suppose the null hypothesis, H_0, is the sunken ship does not contain buried treasure. State the Type I and Type II errors in complete sentences.

19. A microbiologist is testing a water sample for E. coli. Suppose the null hypothesis, H_0, is the sample does not contain E. coli. The probability that the sample does not contain E. coli, but the microbiologist thinks it does is 0.012. The probability that the sample does contain E. coli, but the microbiologist thinks it does not is 0.002. What is the power of this test?

20. A microbiologist is testing a water sample for E. coli. Suppose the null hypothesis, H_0, is the sample contains E-coli. Which is the error with the greater consequence?

9.3 Distribution Needed for Hypothesis Testing

21. Which two distributions can you use for hypothesis testing for this chapter?

22. Which distribution do you use when the standard deviation is not known? Assume sample size is large.

23. Which distribution do you use when the standard deviation is not known and you are testing one population mean? Assume sample size is large.

24. A population mean is 13. The sample mean is 12.8, and the sample standard deviation is two. The sample size is 20. What distribution should you use to perform a hypothesis test? Assume the underlying population is normal.

25. A population has a mean of 25 and a standard deviation of five. The sample mean is 24, and the sample size is 108. What distribution should you use to perform a hypothesis test?

26. It is thought that 42 percent of respondents in a taste test would prefer Brand *A*. In a particular test of 100 people, 39 percent preferred Brand *A*. What distribution should you use to perform a hypothesis test?

27. You are performing a hypothesis test of a single population mean using a Student's *t*-distribution. What must you assume about the distribution of the data?

28. You are performing a hypothesis test of a single population mean using a Student's *t*-distribution. The data are not from a simple random sample. Can you accurately perform the hypothesis test?

29. You are performing a hypothesis test of a single population proportion. What must be true about the quantities of np and nq?

30. You are performing a hypothesis test of a single population proportion. You find out that np is less than five. What must you do to be able to perform a valid hypothesis test?

31. You are performing a hypothesis test of a single population proportion. The data come from which distribution?

9.4 Rare Events, the Sample, and the Decision and Conclusion

32. When do you reject the null hypothesis?

33. The probability of winning the grand prize at a particular carnival game is 0.005. Is the outcome of winning very likely or very unlikely?

34. The probability of winning the grand prize at a particular carnival game is 0.005. Michele wins the grand prize. Is this considered a rare or common event? Why?

35. It is believed that the mean height of high school students who play basketball on the school team is 73 inches with a standard deviation of 1.8 inches. A random sample of 40 players is chosen. The sample mean was 71 inches, and the sample standard deviation was 1.5 inches. Do the data support the claim that the mean height is less than 73 inches? The *p*-value is almost zero. State the null and alternative hypotheses and interpret the *p*-value.

36. The mean age of graduate students at a university is at most 31 years with a standard deviation of two years. A random sample of 15 graduate students is taken. The sample mean is 32 years and the sample standard deviation is three years. Are the data significant at the 1 percent level? The *p*-value is 0.0264. State the null and alternative hypotheses and interpret the *p*-value.

37. Does the shaded region represent a low or a high *p*-value compared to a level of significance of 1 percent?

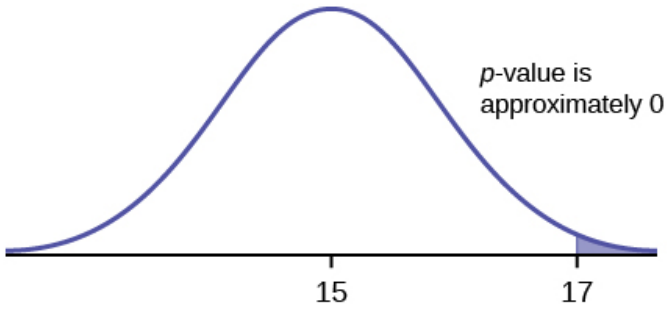


Figure 9.22

38. What should you do when $\alpha > p$-value?

39. What should you do if $\alpha = p$-value?

40. If you do not reject the null hypothesis, then it must be true. Is that statement correct? State why or why not in complete sentences.

Use the following information to answer the next seven exercises: Suppose that a recent article stated that the mean time students spend doing homework each week is 2.5 hours. A study was then done to see if the mean time has increased in the new century. A random sample of 26 students was taken. The mean length of time they did homework each week was three hours with a standard deviation of 1.8 hours. Suppose that it is somehow known that the population standard deviation is 1.5. Conduct a hypothesis test to determine if the mean length of time doing homework each week has increased. Assume the distribution of homework times is approximately normal.

41. Is this a test of means or proportions?

42. What symbol represents the random variable for this test?

43. In words, define the random variable for this test.

44. Is σ known and, if so, what is it?

45. Calculate the following:
a. $\overline{x}$ _____
b. σ _____
c. s_x _____
d. n _____

46. Since both σ and s_x are given, which should be used? In one to two complete sentences, explain why.

47. State the distribution to use for the hypothesis test.

48. A random survey of 75 long-term marathon runners revealed that the mean length of time they have been

running is 17.4 years with a standard deviation of 6.3 years. Conduct a hypothesis test to determine if the population mean time is likely to be 15 years.

a. Is this a test of one mean or proportion?
b. State the null and alternative hypotheses.
   H_0: _____ H_a : _____
c. Is this a right-tailed, left-tailed, or two-tailed test?
d. What symbol represents the random variable for this test?
e. In words, define the random variable for this test.
f. Is the population standard deviation known and, if so, what is it?
g. Calculate the following:
   i. $\overline{x}$ = _____
   ii. s = _____
   iii. n = _____
h. Which test should be used?
i. State the distribution to use for the hypothesis test.
j. Find the p-value.
k. At a pre-conceived $\alpha = 0.05$, give your answer for each of the following:
   i. Decision:
   ii. Reason for the decision:
   iii. Conclusion (write out in a complete sentence):

9.5 Additional Information and Full Hypothesis Test Examples

49. Assume H_0: $\mu = 9$ and H_a: $\mu < 9$. Is this a left-tailed, right-tailed, or two-tailed test?

50. Assume H_0: $\mu \leq 6$ and H_a: $\mu > 6$. Is this a left-tailed, right-tailed, or two-tailed test?

51. Assume H_0: $p = 0.25$ and H_a: $p \neq 0.25$. Is this a left-tailed, right-tailed, or two-tailed test?

52. Draw the general graph of a left-tailed test.

53. Draw the graph of a two-tailed test.

54. A bottle of water is labeled as containing 16 fluid ounces of water. You believe it is less than that. What type of test would you use?

55. Your friend claims that his mean golf score is 63. You want to show that it is higher than that. What type of test would you use?

56. A bathroom scale claims to be able to identify correctly any weight within a pound. You think that it cannot be that accurate. What type of test would you use?

57. You flip a coin and record whether it shows heads or tails. You know the probability of getting heads is 50 percent, but you think it is less for this particular coin. What type of test would you use?

58. If the alternative hypothesis has a not equals ($\neq$) symbol, you know to use which type of test?

59. Assume the null hypothesis states that the mean is at least 18. Is this a left-tailed, right-tailed, or two-tailed test?

60. Assume the null hypothesis states that the mean is at most 12. Is this a left-tailed, right-tailed, or two-tailed test?

61. Assume the null hypothesis states that the mean is equal to 88. The alternative hypothesis states that the mean is not equal to 88. Is this a left-tailed, right-tailed, or two-tailed test?

Homework

9.1 Null and Alternative Hypotheses

62. Some of the following statements refer to the null hypothesis, some to the alternate hypothesis.
State the null hypothesis, H_0, and the alternative hypothesis. H_a, in terms of the appropriate parameter (μ or p).

a. The mean number of years Americans work before retiring is 34.
b. At most 60 percent of Americans vote in presidential elections.
c. The mean starting salary for San Jose State University graduates is at least $100,000 per year.
d. Twenty-nine percent of high school students take physical education daily.
e. Less than 5 percent of adults ride the bus to work in Los Angeles.

f. The mean number of cars a person owns in her lifetime is not more than 10.
g. About half of Americans prefer to live away from cities, given the choice.
h. Europeans have a mean paid vacation each year of six weeks.
i. The chance of developing breast cancer is under 11 percent for women.
j. Private universities' mean tuition cost is more than $20,000 per year.

63. A recent survey of 273 randomly selected teens living in Massachusetts asked about social media. Sixty-three said that they routinely use a certain app to share pictures. The researchers want to determine if there is good evidence that more than 30 percent of teens use this app. The alternative hypothesis is as follows:
a. $p < 0.30$
b. $p \leq 0.30$
c. $p \geq 0.30$
d. $p > 0.30$

64. A statistics instructor believes that fewer than 20 percent of Evergreen Valley College (EVC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 attended the midnight showing. An appropriate alternative hypothesis is as follows:
a. $p = 0.20$
b. $p > 0.20$
c. $p < 0.20$
d. $p \leq 0.20$

65. Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test. The null and alternative hypotheses are as follows:
a. $H_o: \overline{x} = 4.5$, $H_a: \overline{x} > 4.5$
b. $H_o: \mu \geq 4.5$, $H_a: \mu < 4.5$
c. $H_o: \mu = 4.75$, $H_a: \mu > 4.75$
d. $H_o: \mu = 4.5$, $H_a: \mu > 4.5$

9.2 Outcomes and the Type I and Type II Errors

66. State the Type I and Type II errors in complete sentences given the following statements.
a. The mean number of years Americans work before retiring is 34.
b. At most 60 percent of Americans vote in presidential elections.
c. The mean starting salary for San Jose State University graduates is at least $100,000 per year.
d. 29 percent of high school students take physical education every day.
e. Less than 5 percent of adults ride the bus to work in Los Angeles.
f. The mean number of cars a person owns in his or her lifetime is not more than 10.
g. About half of Americans prefer to live away from cities, given the choice.
h. Europeans have a mean paid vacation each year of six weeks.
i. The chance of developing breast cancer is under 11 percent for women.
j. Private universitie' mean tuition cost is more than $20,000 per year.

67. For Statements A–J in Exercise 9.66, answer the following in complete sentences.
a. State a consequence of committing a Type I error.
b. State a consequence of committing a Type II error.

68. When a new drug is created, the pharmaceutical company must subject it to testing before receiving the necessary permission from the U.S. Food and Drug Administration (FDA) to market the drug. Suppose the null hypothesis is *the drug is unsafe*. What is the Type II error?
a. To conclude the drug is safe when, in fact, it is unsafe.
b. Not to conclude the drug is safe when, in fact, it is safe.
c. To conclude the drug is safe when, in fact, it is safe.
d. Not to conclude the drug is unsafe when, in fact, it is unsafe.

69. A statistics instructor believes that fewer than 20 percent of Evergreen Valley College (EVC) students attended the opening midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of

them attended the midnight showing. The Type I error is to conclude that the percent of EVC students who attended is _____.

a. at least 20 percent, when, in fact, it is less than 20 percent.
b. 20 percent, when, in fact, it is 20 percent.
c. less than 20 percent, when, in fact, it is at least 20 percent.
d. less than 20 percent, when, in fact, it is less than 20 percent.

70. It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5 percent, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average?
The Type II error is not to reject that the mean number of hours of sleep LTCC students get per night is at least seven when, in fact, the mean number of hours

a. is more than seven hours.
b. is at most seven hours.
c. is at least seven hours.
d. is less than seven hours.

71. Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test. The Type I error is

a. to conclude that the current mean hours per week is higher than 4.5, when, in fact, it is higher.
b. to conclude that the current mean hours per week is higher than 4.5, when, in fact, it is the same.
c. to conclude that the mean hours per week currently is 4.5, when, in fact, it is higher.
d. to conclude that the mean hours per week currently is no higher than 4.5, when, in fact, it is not higher.

9.3 Distribution Needed for Hypothesis Testing

72. It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5 percent, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average? The distribution to be used for this test is $\overline{X}$ ~ _____

a. $N(7.24, \frac{1.93}{\sqrt{22}})$
b. $N(7.24, 1.93)$
c. t_{22}
d. t_{21}

9.4 Rare Events, the Sample, and the Decision and Conclusion

73. The National Institute of Mental Health published an article stating that in any one-year period approximately 9.5 percent of American adults suffer from depression or a depressive illness. Suppose that in a survey of 100 people in a certain town, seven of them suffered from depression or a depressive illness. Conduct a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population.

a. Is this a test of one mean or proportion?
b. State the null and alternative hypotheses.
   H_0: _____ H_a: _____
c. Is this a right-tailed, left-tailed, or two-tailed test?
d. What symbol represents the random variable for this test?
e. In words, define the random variable for this test.
f. Calculate the following:
   i. x = _____
   ii. n = _____
   iii. p' = _____
g. Calculate σ_x = _____. Show the formula setup.

h. State the distribution to use for the hypothesis test.
i. Find the *p*-value.
j. At a pre-conceived $\alpha = 0.05$, give your answer for each of the following:
   i. Decision:
   ii. Reason for the decision:
   iii. Conclusion (write out in a complete sentence):

9.5 Additional Information and Full Hypothesis Test Examples

For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in Appendix E, Solution Sheets. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

Note

If you are using a Student's-*t*-distribution for one of the following homework problems, you may assume that the underlying population is normally distributed. In general, you must first prove that assumption, however.

74. A particular brand of tires claims that its deluxe tire averages at least 50,000 miles before it needs to be replaced. From past studies of this tire, the standard deviation is known to be 8,000. A survey of owners of that tire design is conducted. From the 28 tires surveyed, the mean lifespan was 46,500 miles with a standard deviation of 9,800 miles. Using alpha = 0.05, are the data highly inconsistent with the claim?

75. In 2009, President Barack Obama announced a new national fuel economy and emissions policy for cars and light trucks. It stated that the combined fleet fuel economy for an auto manufacturer of cars and light trucks will have to average 35.5 mpg or better by 2016. From past studies on fuel economy, it is known that the standard deviation of a typical fleet is 7.6 mpg. An auto manufacturer selects a random sample of 55 cars and light trucks and finds the sample mean fuel economy to be 34.6 mpg with a standard deviation of 10.3 mpg. Can the manufacturer claim that their fleet meets the fuel economy standard in the 2016 policy at the 5 percent level?

76. The cost of a daily newspaper varies from city to city. However, the variation among prices remains steady with a standard deviation of 20¢. A study was done to test the claim that the mean cost of a daily newspaper is $1.00. Twelve costs yield a mean cost of 95¢ with a standard deviation of 18¢. Do the data support the claim at the 1 percent level?

77. An article in the *San Jose Mercury News* stated that students in the California state university system take 4.5 years, on average, to finish their undergraduate degrees. Suppose you believe that the mean time is longer. You conduct a survey of 49 students and obtain a sample mean of 5.1 with a sample standard deviation of 1.2. Do the data support your claim at the 1 percent level?

78. The mean number of sick days an employee takes per year is believed to be about 10. Members of a personnel department do not believe this figure. They randomly survey eight employees. The number of sick days they took for the past year are as follows: 12; 4; 15; 3; 11; 8; 6; 8. Let x = the number of sick days they took for the past year. Should the personnel team believe that the mean number is 10?

79. In 1955, *Life Magazine* reported that the 25-year-old mother of three worked, on average, an 80-hour week. Recently, many groups have been studying whether or not the women's movement has, in fact, resulted in an increase in the average work week for women (combining employment and at-home work). Suppose a study was done to determine if the mean work week has increased. Eighty-one women were surveyed with the following results. The sample mean was 83; the sample standard deviation was 10. Does it appear that the mean work week has increased for women at the 5 percent level?

80. Your statistics instructor claims that 60 percent of the students who take her Elementary Statistics class go through life feeling more enriched. For some reason that she can't quite figure out, most people don't believe her. You decide to check this out on your own. You randomly survey 64 of her past Elementary Statistics students and find that 34 feel more enriched as a result of her class. Now, what do you think?

81. A Nissan Motor Corporation advertisement read, “The average man’s I.Q. is 107. The average brown trout’s I.Q. is 4. So why can’t man catch brown trout?” Suppose you believe that the brown trout’s mean I.Q. is greater than four. You catch 12 brown trout. A fish psychologist determines the I.Q.s as follows: 5, 4, 7, 3, 6, 4, 5, 3, 6, 3, 8, 5. Conduct a hypothesis test of your belief.

82. Refer to Exercise 9.81. Conduct a hypothesis test to see if your decision and conclusion would change if your belief were that the brown trout's mean I.Q. is **not** four.

83. According to an article in *Newsweek*, the natural ratio of girls to boys is 100:105. In China, the birth ratio is 100:114 (46.7 percent girls). Suppose you don't believe the reported figures of the percentage of girls born in China. You conduct a study. In this study, you count the number of girls and boys born in 150 randomly chosen recent births. There are 60 girls and 90 boys born of the 150. Based on your study, do you believe that the percentage of girls born in China is 46.7?

84. A group of researchers research a common contagious disease. A newspaper found that 13 percent of Americans have been diagnosed with the disease in the last year. The researchers doubt that the percentage is really that high. It conducts its own survey. Out of 76 Americans surveyed, only two had been diagnosed with the disease. Would you agree with the newspaper's poll? In complete sentences, give three reasons why polls might give different results.

85. The mean work week for engineers in a start-up company is believed to be about 60 hours. A newly hired engineer hopes that it's shorter. She asks 10 engineering friends in start-ups for the lengths of their mean work weeks. Based on the results that follow, should she count on the mean work week to be shorter than 60 hours? Data (length of mean work week): 70, 45, 55, 60, 65, 55, 55, 60, 50, 55.

86. Use the *Lap time* data for Lap 4 (see Appendix C: Data Sets) to test the claim that Terri finishes Lap 4, on average, in less than 129 seconds. Use all 20 races given.

87. Use the *Initial Public Offering* data (see Appendix C: Data Sets) to test the claim that the mean offer price was $18 per share. Do not use all the data. Use your random number generator to randomly survey 15 prices.

Note

The following questions were written by past students. They are excellent problems!

88. "Asian Family Reunion," by Chau Nguyen
Every two years it comes around.
We all get together from different towns.
In my honest opinion,
It's not a typical family reunion.
Not forty, or fifty, or sixty,
But how about seventy companions!
The kids would play, scream, and shout
One minute they're happy, another they'll pout.
The teenagers would look, stare, and compare
From how they look to what they wear.
The men would chat about their business
That they make more, but never less.
Money is always their subject
And there's always talk of more new projects.
The women get tired from all of the chats
They head to the kitchen to set out the mats.
Some would sit and some would stand
Eating and talking with plates in their hands.
Then come the games and the songs
And suddenly, everyone gets along!
With all that laughter, it's sad to say
That it always ends in the same old way.
They hug and kiss and say "good-bye"
And then they all begin to cry!
I say that 60 percent shed their tears
But my mom counted 35 people this year.
She said that boys and men will always have their pride,
So we won't ever see them cry.

I myself don't think she's correct,
So could you please try this problem to see if you object?

89. "Blowing Bubbles," by Sondra Prull
Studying stats just made me tense,
I had to find some sane defense.
Some light and lifting simple play
To float my math anxiety away.
Blowing bubbles lifts me high
Takes my troubles to the sky.
POIK! They're gone, with all my stress
Bubble therapy is the best.
The label said each time I blew
The average number of bubbles would be at least 22.
I blew and blew and this I found
From 64 blows, they all are round!
But the number of bubbles in 64 blows
Varied widely, this I know.
20 per blow became the mean
They deviated by 6, and not 16.
From counting bubbles, I sure did relax
But now I give to you your task.
Was 22 a reasonable guess?
Find the answer and pass this test!

90. "Dalmatian Darnation," by Kathy Sparling
A greedy dog breeder named Spreckles
Bred puppies with numerous freckles
The Dalmatians he sought
Possessed spot upon spot
The more spots, he thought, the more shekels.
His competitors did not agree
That freckles would increase the fee.
They said, “Spots are quite nice
But they don't affect price;
One should breed for improved pedigree.”
The breeders decided to prove
This strategy was a wrong move.
Breeding only for spots
Would wreak havoc, they thought.
His theory they want to disprove.
They proposed a contest to Spreckles
Comparing dog prices to freckles.
In records they looked up
One hundred one pups:
Dalmatians that fetched the most shekels.
They asked Mr. Spreckles to name
An average spot count he'd claim
To bring in big bucks.
Said Spreckles, “Well, shucks,
It's for one hundred one that I aim.”
Said an amateur statistician
Who wanted to help with this mission.
“Twenty-one for the sample
Standard deviation's ample.”
They examined one hundred and one
Dalmatians that fetched a good sum.
They counted each spot,
Mark, freckle, and dot

And tallied up every one.
Instead of one hundred one spots
They averaged ninety-six dots
Can they muzzle Spreckles'
Obsession with freckles
Based on all the dog data they've got?

91. *Macaroni and Cheese, please!!* by Nedda Misherghi and Rachelle Hall
As a poor starving student I don't have much money to spend for even the bare necessities. So my favorite and main staple food is macaroni and cheese. It's high in taste and low in cost and nutritional value.
One day, as I sat down to determine the meaning of life, I got a serious craving for this, oh, so important, food of my life. So I went down the street to Greatway to get a box of macaroni and cheese, but it was SO expensive! $2.02 !!! Can you believe it? It made me stop and think. The world is changing fast. I had thought that the mean cost of a box (the normal size, not some super-gigantic-family-value-pack) was at most $1, but now I wasn't so sure. However, I was determined to find out. I went to 53 of the closest grocery stores and surveyed the prices of macaroni and cheese. Here are the data I wrote in my notebook:
Price per box of Mac and Cheese

- 5 stores @ $2.02
- 15 stores @ $0.25
- 3 stores @ $1.29
- 6 stores @ $0.35
- 4 stores @ $2.27
- 7 stores @ $1.50
- 5 stores @ $1.89
- 8 stores @ $0.75

I could see that the cost varied but I had to sit down to figure out whether or not I was right. If it does turn out that this mouth-watering dish is at most $1, then I'll throw a big cheesy party in our next statistics lab, with enough macaroni and cheese for just me. After all, as a poor starving student I can't be expected to feed our class of animals!

92. "William Shakespeare: The Tragedy of Hamlet, Prince of Denmark," by Jacqueline Ghodsi
THE CHARACTERS (in order of appearance):

- HAMLET, Prince of Denmark and student of statistics
- POLONIUS, Hamlet's tutor
- HORATIO, friend to Hamlet and fellow student

Scene: The great library of the castle, in which Hamlet does his lessons
Act I
The day is fair, but the face of Hamlet is clouded. He paces the large room. His tutor, Polonius, is reprimanding Hamlet regarding the latter's recent experience. Horatio is seated at the large table at right stage.
POLONIUS: My Lord, how cans't thou admit that thou hast seen a ghost! It is but a figment of your imagination!
HAMLET: I beg to differ; I know of a certainty that five-and-seventy in one hundred of us, condemned to the whips and scorns of time as we are, have gazed upon a spirit of health, or goblin damn'd, be their intents wicked or charitable.
POLONIUS: If thou dost insist upon thy wretched vision then let me invest your time; be true to thy work and speak to me through the reason of the null and alternate hypotheses. (He turns to Horatio.) Did not Hamlet himself say, "What a piece of work is man, how noble in reason, how infinite in faculties"? Then let not this foolishness persist. Go, Horatio, make a survey of three-and-sixty and discover what the true proportion be. For my part, I will never succumb to this fantasy, but deem man to be devoid of all reason should thy proposal of at least five-and-seventy in one hundred hold true.
HORATIO (to Hamlet): What should we do, my Lord?
HAMLET: Go to thy purpose, Horatio.
HORATIO: To what end, my Lord?
HAMLET: That you must teach me. But let me conjure you by the rights of our fellowship, by the consonance of our youth, but the obligation of our ever-preserved love, be even and direct with me, whether I am right or no.
Horatio exits, followed by Polonius, leaving Hamlet to ponder alone.
Act II
The next day, Hamlet awaits anxiously the presence of his friend, Horatio. Polonius enters and places some books upon the table just a moment before Horatio enters.

POLONIUS: So, Horatio, what is it thou didst reveal through thy deliberations?
HORATIO: In a random survey, for which purpose thou thyself sent me forth, I did discover that one-and-forty believe fervently that the spirits of the dead walk with us. Before my God, I might not this believe, without the sensible and true avouch of mine own eyes.
POLONIUS: Give thine own thoughts no tongue, Horatio. (Polonius turns to Hamlet.) But look to't I charge you, my Lord. Come Horatio, let us go together, for this is not our test. (Horatio and Polonius leave together.)
HAMLET: To reject, or not reject, that is the question: whether 'tis nobler in the mind to suffer the slings and arrows of outrageous statistics, or to take arms against a sea of data, and, by opposing, end them. (Hamlet resignedly attends to his task.)
(Curtain falls)

93. "Untitled," by Stephen Chen
I've often wondered how software is released and sold to the public. Ironically, I work for a company that sells products with known problems. Unfortunately, most of the problems are difficult to create, which makes them difficult to fix. I usually use the test program X, which tests the product, to try to create a specific problem. When the test program is run to make an error occur, the likelihood of generating an error is 1 percent.
So, armed with this knowledge, I wrote a new test program Y that will generate the same error that test program X creates, but more often. To find out if my test program is better than the original, so that I can convince the management that I'm right, I ran my test program to find out how often I can generate the same error. When I ran my test program 50 times, I generated the error twice. While this may not seem much better, I think that I can convince the management to use my test program instead of the original test program. Am I right?

94. "Japanese Girls' Names"
by Kumi Furuichi
It used to be very typical for Japanese girls' names to end with "ko." The trend might have started around my grandmothers' generation and its peak might have been around my mother's generation. "Ko" means "child" in Chinese characters. Parents would name their daughters with "ko" attaching to other Chinese characters that have meanings that they want their daughters to become, such as Sachiko—happy child, Yoshiko—a good child, Yasuko—a healthy child, and so on.
However, I noticed recently that only two out of nine of my Japanese girlfriends at this school have names that end with "ko." More and more, parents seem to have become creative, modernized, and, sometimes, westernized in naming their children.
I have a feeling that, while 70 percent or more of my mother's generation would have names with "ko" at the end, the proportion has dropped among my peers. I wrote down all my Japanese friends', ex-classmates', coworkers', and acquaintances' names that I could remember. Following are the names. Some are repeats. Test to see if the proportion has dropped for this generation.
Ai, Akemi, Akiko, Ayumi, Chiaki, Chie, Eiko, Eri, Eriko, Fumiko, Harumi, Hitomi, Hiroko, Hiroko, Hidemi, Hisako, Hinako, Izumi, Izumi, Junko, Junko, Kana, Kanako, Kanayo, Kayo, Kayoko, Kazumi, Keiko, Keiko, Kei, Kumi, Kumiko, Kyoko, Kyoko, Madoka, Maho, Mai, Maiko, Maki, Miki, Miki, Mikiko, Mina, Minako, Miyako, Momoko, Nana, Naoko, Naoko, Naoko, Noriko, Rieko, Rika, Rika, Rumiko, Rei, Reiko, Reiko, Sachiko, Sachiko, Sachiyo, Saki, Sayaka, Sayoko, Sayuri, Seiko, Shiho, Shizuka, Sumiko, Takako, Takako, Tomoe, Tomoe, Tomoko, Touko, Yasuko, Yasuko, Yasuyo, Yoko, Yoko, Yoko, Yoshiko, Yoshiko, Yoshiko, Yuka, Yuki, Yuki, Yukiko, Yuko, Yuko.

95. "Phillip's Wish," by Suzanne Osorio
My nephew likes to play
Chasing the girls makes his day.
He asked his mother
If it is okay
To get his ear pierced.
She said, "No way!"
To poke a hole through your ear,
Is not what I want for you, dear.
He argued his point quite well,
Says even my macho pal, Mel,
Has gotten this done.
It's all just for fun.
C'mon please, mom, please, what the hell.
Again Phillip complained to his mother,
Saying half his friends (including their brothers)
Are piercing their ears

And they have no fears
He wants to be like the others.
She said, “I think it’s much less.
We must do a hypothesis test.
And if you are right,
I won’t put up a fight.
But, if not, then my case will rest.”
We proceeded to call fifty guys
To see whose prediction would fly.
Nineteen of the fifty
Said piercing was nifty
And earrings they’d occasionally buy.
Then there’s the other thirty-one,
Who said they’d never have this done.
So now this poem’s finished.
Will his hopes be diminished,
Or will my nephew have his fun?

96. "The Craven," by Mark Salangsang
Once upon a morning dreary
In stats class I was weak and weary.
Pondering over last night’s homework
Whose answers were now on the board
This I did and nothing more.
While I nodded nearly napping
Suddenly, there came a tapping.
As someone gently rapping,
Rapping my head as I snore.
Quoth the teacher, “Sleep no more.”
“In every class you fall asleep,”
The teacher said, his voice was deep.
“So a tally I’ve begun to keep
Of every class you nap and snore.
The percentage being forty-four.”
“My dear teacher I must confess,
While sleeping is what I do best.
The percentage, I think, must be less,
A percentage less than forty-four.”
This I said and nothing more.
“We’ll see,” he said and walked away,
And fifty classes from that day
He counted till the month of May
The classes in which I napped and snored.
The number he found was twenty-four.
At a significance level of 0.05,
Please tell me am I still alive?
Or did my grade just take a dive
Plunging down beneath the floor?
Upon thee I hereby implore.

97. Toastmasters International cites a report by Gallup Poll that 40 percent of Americans fear public speaking. A student believes that less than 40 percent of students at her school fear public speaking. She randomly surveys 361 schoolmates and finds that 135 report they fear public speaking. Conduct a hypothesis test to determine if the percentage at her school is less than 40.

98. Sixty-eight percent of online courses taught at community colleges nationwide were taught by full-time faculty. To test if 68 percent also represents California’s percent for full-time faculty teaching the online classes, Long Beach City College (LBCC) in California was randomly selected for comparison. In the same year, 34 of the 44 online courses LBCC offered were taught by full-time faculty. Conduct a hypothesis test to determine if 68 percent represents California. Note: For more accurate results, use more California community colleges and this

past year's data.

99. According to an article in a local poll, a city found that 14 percent of its residents walk for exercise. Suppose that a survey is conducted to determine this year's rate. Nine out of 70 randomly chosen city residents replied that they walk for exercise. Conduct a hypothesis test to determine if the rate is still 14 percent or if it has decreased.

100. The mean age of De Anza College students in a previous term was 26.6 years old. An instructor thinks the mean age for online students is older than 26.6. She randomly surveys 56 online students and finds that the sample mean is 29.4 with a standard deviation of 2.1. Conduct a hypothesis test.

101. Registered nurses earned an average annual salary of $69,110. For that same year, a survey was conducted of 41 California registered nurses to determine if the annual salary is higher than $69,110 for California nurses. The sample average was $71,121 with a sample standard deviation of $7,489. Conduct a hypothesis test.

102. La Leche League International reports that the mean age of weaning a child from breastfeeding is age four to five worldwide. In America, most nursing mothers wean their children much earlier. Suppose a random survey is conducted of 21 U.S. mothers who recently weaned their children. The mean weaning age was nine months (3/4 year) with a standard deviation of 4 months. Conduct a hypothesis test to determine if the mean weaning age in the United States is less than four years old.

103. Harley Davidson motorcycles are the largest selling motorcycle in the United States, with 14 percent of all motorcycles sold in 2012. Interestingly, a random sample of 1,945 stolen motorcycles was selected, and it was found that just 8 percent of them were Harleys. Is there good evidence that the proportion of Harleys among stolen motorcycles is significantly less than their share of all motorcycles? After conducting the test, what decision and conclusion would you make?

a. Reject H_0: There is sufficient evidence to conclude that the proportion of Harleys stolen is significantly less than their share of all motorcycles
b. Do not reject H_0: There is not sufficient evidence to conclude that the proportion of Harleys stolen is significantly less than their share of all motorcycles
c. Do not reject H_0: There is sufficient evidence to conclude that the proportion of Harleys stolen is significantly more than their share of all motorcycles
d. Reject H_0: There is sufficient evidence to conclude that the proportion of Harleys stolen is significantly more than their share of all motorcycles

104. A statistics instructor believes that fewer than 20 percent of Evergreen Valley College (EVC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of them attended the midnight showing.
At a 1 percent level of significance, what is an appropriate conclusion?

a. There is insufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is less than 20 percent.
b. There is sufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is more than 20 percent.
c. There is sufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is less than 20 percent.
d. There is insufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is at least 20 percent.

105. Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test.
At a significance level of $a = 0.05$, what is the correct conclusion?

a. There is enough evidence to conclude that the mean number of hours is more than 4.75.
b. There is enough evidence to conclude that the mean number of hours is more than 4.5.
c. There is not enough evidence to conclude that the mean number of hours is more than 4.5.
d. There is not enough evidence to conclude that the mean number of hours is more than 4.75.

Hypothesis testing: For the following 10 exercises, answer each question.

a. State the null and alternate hypotheses.
b. State the *p*-value.

c. State alpha.
d. What is your decision?
e. Write a conclusion.
f. Answer any other questions asked in the problem.

106. A research group is studying a particular infectious disease. In 2011 at least 18 percent of nursing home residents had the disease. An Introduction to Statistics class in Daviess County, KY, conducted a hypothesis test at the nursing home (approximately 1,200 residents) to determine if the local nursing home's incidence was lower. One hundred fifty residents were chosen at random and surveyed. Of the 150 residents surveyed, 82 have the disease. Use a significance level of 0.05 and, using appropriate statistical evidence, conduct a hypothesis test and state the conclusions.

107. A recent survey in the *New York Times Almanac* indicated that 48.8 percent of families own stock. A broker wanted to determine if this survey could be valid. He surveyed a random sample of 250 families and found that 142 owned some type of stock. At the 0.05 significance level, can the survey be considered to be accurate?

108. Driver error can be listed as the cause of approximately 54 percent of all fatal auto accidents, according to the American Automobile Association. Thirty randomly selected fatal accidents are examined, and it is determined that 14 were caused by driver error. Using $\alpha = 0.05$, is the AAA proportion accurate?

109. The U.S. Department of Energy reported that 51.7 percent of homes were heated by natural gas. A random sample of 221 homes in Kentucky found that 115 were heated by natural gas. Does the evidence support the claim for Kentucky at the $\alpha = 0.05$ level? Are the results applicable across the country? Why?

110. For Americans using library services, the American Library Association claims that at most 67 percent of patrons borrow books. The library director in Owensboro, KY, feels this is not true, so she asked a local college statistic class to conduct a survey. The class randomly selected 100 patrons and found that 82 borrowed books. Did the class demonstrate that the percentage was higher in Owensboro, KY? Use $\alpha = 0.01$ level of significance. What is the possible proportion of patrons who do borrow books from the Owensboro Library?

111. The Weather Underground reported that the mean amount of summer rainfall for the northeastern United States is at least 11.52 inches. Ten cities in the northeast are randomly selected and the mean rainfall amount is calculated to be 7.42 inches with a standard deviation of 1.3 inches. At the $\alpha = 0.05$ level, can it be concluded that the mean rainfall was below the reported average? What if $\alpha = 0.01$? Assume the amount of summer rainfall follows a normal distribution.

112. A survey in the *New York Times Almanac* finds the mean commute time (one way) is 25.4 minutes for the 15 largest US cities. The Austin, TX, chamber of commerce feels that Austin's commute time is less and wants to publicize this fact. The mean for 25 randomly selected commuters is 22.1 minutes with a standard deviation of 5.3 minutes. At the $\alpha = 0.10$ level, is the Austin, TX, commute significantly less than the mean commute time for the 15 largest U.S. cities?

113. A report by the Gallup Poll found that a woman visits her doctor, on average, at most 5.8 times each year. A random sample of 20 women results in these yearly visit totals:
3; 2; 1; 3; 7; 2; 9; 4; 6; 6; 8; 0; 5; 6; 4; 2; 1; 3; 4; 1
At the $\alpha = 0.05$ level, can it be concluded that the sample mean is higher than 5.8 visits per year?

114. According to the *New York Times Almanac* the mean family size in the United States is 3.18. A sample of a college math class resulted in the following family sizes:
5; 4; 5; 4; 4; 3; 6; 4; 3; 3; 5; 5; 6; 3; 3; 2; 7; 4; 5; 2; 2; 2; 3; 2
At $\alpha = 0.05$, is the class's mean family size greater than the national average? Does the *Almanac* result remain valid? Why?

115. The student academic group on a college campus claims that freshman students study at least 2.5 hours per day, on average. One Introduction to Statistics class was skeptical. The class took a random sample of 30 freshman students and found a mean study time of 137 minutes with a standard deviation of 45 minutes. At $\alpha = 0.01$ level, is the student academic group's claim correct?

References

9.1 Null and Alternative Hypotheses

Centers for Disease Control and Prevention. (n.d.). *Physical activity facts*. Retrieved from http://www.cdc.gov/healthyschools/physicalactivity/facts.htm

National Institute of Mental Health. (n.d.). *Publications about depression*. Retrieved from http://www.nimh.nih.gov/publicat/depression.cfm

9.5 Additional Information and Full Hypothesis Test Examples

American Automobile Association. (n.d.). Retrieved from www.aaa.com

American Library Association. (n.d.). Retrieved from www.ala.org

Amit Schitai. (n.d.). *Data*.

Bureau of Labor Statistics. (n.d.). *Occupational employment statistics*. Retrieved from http://www.bls.gov/oes/current/oes291111.htm

Centers for Disease Control and Prevention. (n.d.). Retrieved from www.cdc.gov

De Anza College. (2006). *Foothill-De Anza Community College District*. Retrieved from http://research.fhda.edu/factbook/DAdemofs/Fact_sheet_da_2006w.pdf

Federal Bureau of Investigation. (n.d.). *Uniform crime reports and index of crime in Daviess in the state of Kentucky enforced by Daviess County from 1985 to 2005*. Retrieved from http://www.disastercenter.com/kentucky/crime/3868.htm

Gallup. (n.d.). Retrieved from www.gallup.com

Johansen, C., Boice, Jr. J., McLaughlin, J., & Olsen, J. (2001, Feb. 7). Cellular telephones and cancer—a nationwide cohort study in Denmark. *Journal of National Cancer Institute*, 93(3), 203–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11158188

La Leche League International. (n.d.). Retrieved from http://www.lalecheleague.org/Law/BAFeb01.html

Online Learning Consortium. (2005 Nov.). *Growing by degrees: Online education in the United States, 2005*. Newburyport, MA: Allen, I. E., & Seaman, J. Available at http://files.eric.ed.gov/fulltext/ED530062.pdf

Toastmasters International. (n.d.). Retrieved from http://toastmasters.org/artisan/detail.asp?CategoryID=1&SubCategoryID=10&ArticleID=429&Page=1

U.S. Census Bureau. (n.d.). *Language use*. Retrieved from https://www.census.gov/topics/population/language-use.html

U.S. Census Bureau. (n.d.). *QuickFacts*. Retrieved from https://www.census.gov/quickfacts/table/PST045216/00

U.S. Department of Energy. (n.d.). Retrieved from http://energy.gov

Weather Underground. (n.d.). Retrieved from www.wunderground.com

Solutions

1. The random variable is the mean Internet speed in megabits per second.

3. The random variable is the mean number of children an American family has.

5. The random variable is the proportion of people picked at random in Times Square visiting the city.

7. a. H_0: $p = 0.42$
b. H_a: $p < 0.42$

9. a. H_0: $\mu = 15$
b. H_a: $\mu \neq 15$

11. Type I: The mean price of mid-sized cars is \$32,000, but we conclude that it is not \$32,000.
Type II: The mean price of mid-sized cars is not \$32,000, but we conclude that it is \$32,000.

13. α = the probability that you think the bag cannot withstand –15 degrees F, when, in fact, it can.
β = the probability that you think the bag can withstand –15 degrees F, when, in fact, it cannot.

15. Type I: The procedure will go well, but the doctors think it will not.

Type II: The procedure will not go well, but the doctors think it will.

17. 0.019

19. 0.998

21. A normal distribution or a Student's *t*-distribution

23. Use a Student's *t*-distribution

25. a normal distribution for a single population mean

27. It must be approximately normally distributed.

29. They must both be greater than five.

31. binomial distribution

33. The outcome of winning is very unlikely.

35. H_0: $\mu >= 73$
H_a: $\mu < 73$
The *p*-value is almost zero, which means there is sufficient data to conclude that the mean height of high school students who play basketball on the school team is less than 73 inches at the 5 percent level. The data do support the claim.

37. The shaded region shows a low *p*-value.

39. Do not reject H_0.

41. means

43. the mean time spent on homework for 26 students

45.
a. 3
b. 1.5
c. 1.8
d. 26

47. $\overline{X} \sim N\left(2.5, \frac{1.5}{\sqrt{26}}\right)$

49. This is a left-tailed test.

51. This is a two-tailed test.

53.

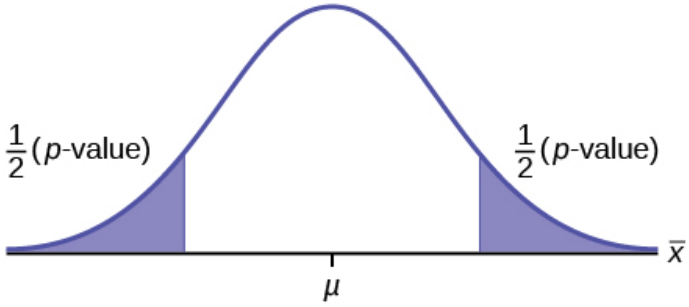


Figure 9.23

55. a right-tailed test

57. a left-tailed test

59. This is a left-tailed test.

61. This is a two-tailed test.

62.
a. H_0: $\mu = 34$; H_a: $\mu \neq 34$
b. H_0: $p \leq 0.60$; H_a: $p > 0.60$
c. H_0: $\mu \geq 100{,}000$; H_a: $\mu < 100{,}000$
d. H_0: $p = 0.29$; H_a: $p \neq 0.29$
e. H_0: $p = 0.05$; H_a: $p < 0.05$
f. H_0: $\mu \leq 10$; H_a: $\mu > 10$
g. H_0: $p = 0.50$; H_a: $p \neq 0.50$
h. H_0: $\mu = 6$; H_a: $\mu \neq 6$
i. H_0: $p \geq 0.11$; H_a: $p < 0.11$
j. H_0: $\mu \leq 20{,}000$; H_a: $\mu > 20{,}000$

64. c

66.
a. Type I error: We conclude that the mean is not 34 years, when it really is 34 years. Type II error: We conclude that the mean is 34 years, when in fact it really is not 34 years.
b. Type I error: We conclude that more than 60 percent of Americans vote in presidential elections, when the actual percentage is at most 60 percent.Type II error: We conclude that at most 60 percent of Americans vote in presidential elections when, in fact, more than 60 percent do.
c. Type I error: We conclude that the mean starting salary is less than $100,000, when it really is at least $100,000. Type II error: We conclude that the mean starting salary is at least $100,000 when, in fact, it is less than $100,000.
d. Type I error: We conclude that the proportion of high school seniors who take physical education daily is not 29%, when it really is 29%. Type II error: We conclude that the proportion of high school seniors who take physical education daily is 29% when, in fact, it is not 29%.
e. Type I error: We conclude that fewer than 5 percent of adults ride the bus to work in Los Angeles, when the percentage that do is really 29%. Type II error: We conclude that 29%. or more adults ride the bus to work in Los Angeles when, in fact, fewer that 29% do.
f. Type I error: We conclude that the mean number of cars a person owns in his or her lifetime is more than 10, when in reality it is not more than 10. Type II error: We conclude that the mean number of cars a person owns in his or her lifetime is not more than 10 when, in fact, it is more than 10.
g. Type I error: We conclude that the proportion of Americans who prefer to live away from cities is not about half, though the actual proportion is about half. Type II error: We conclude that the proportion of Americans who prefer to live away from cities is half when, in fact, it is not half.
h. Type I error: We conclude that the duration of paid vacations each year for Europeans is not six weeks, when in fact it is six weeks. Type II error: We conclude that the duration of paid vacations each year for Europeans is six

weeks when, in fact, it is not.

i. Type I error: We conclude that the proportion is less than 11 percent, when it is really at least 11 percent. Type II error: We conclude that the proportion of women who develop breast cancer is at least 11 percent, when in fact it is less than 11 percent.

j. Type I error: We conclude that the average tuition cost at private universities is more than $20,000, though in reality it is at most $20,000. Type II error: We conclude that the average tuition cost at private universities is at most $20,000 when, in fact, it is more than $20,000.

68. b

70. d

72. d

74.
a. H_0: $\mu \geq 50{,}000$
b. H_a: $\mu < 50{,}000$
c. Let $\overline{X}$ = the average lifespan of a brand of tires.
d. normal distribution
e. $z = -2.315$
f. p-value = 0.0103
g. Check student's solution.
h.
   i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: The p-value is less than 0.05.
   iv. Conclusion: There is sufficient evidence to conclude that the mean lifespan of the tires is less than 50,000 miles.
i. (43,537, 49,463)

75.
a. H_0: $\mu \geq 35.5$
b. H_a: $\mu < 35.5$
c. Let $\overline{x}$ = the average mpg for the sample of cars and trucks in the fleet
d. normal distribution
e. $z = -0.648$
f. p-value = 0.2578
g. Check student's solution.
h.
   i. Alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.05.
   iv. Conclusion: There is sufficient evidence to support the claim that the manufacturer's fleet meets the fuel economy standards in the 2016 policy.
i. (31.88 mpg, 37.32 mpg)

76.
a. H_0: $\mu = \$1.00$
b. H_a: $\mu \neq \$1.00$
c. Let $\overline{x}$ = the average cost of a daily newspaper.
d. normal distribution
e. $z = -0.866$
f. p-value = 0.3865
g. Check student's solution.
h.
   i. Alpha: 0.01
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.01.
   iv. Conclusion: There is sufficient evidence to support the claim that the mean cost of daily papers is $1. The mean cost could be $1.
i. ($0.84, $1.06)

78. a. H_0: $\mu = 10$
b. H_a: $\mu \neq 10$
c. Let $\overline{X}$ = the mean number of sick days an employee takes per year.
d. Student's *t*-distribution
e. $t = -1.12$
f. *p*-value = 0.300
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: The *p*-value is greater than 0.05.
iv. Conclusion: At the 5 percent significance level, there is insufficient evidence to conclude that the mean number of sick days is not 10.

i. (4.9443, 11.806)

80. a. H_0: $p \geq 0.6$
b. H_a: $p < 0.6$
c. Let P' = the proportion of students who feel more enriched as a result of taking elementary statistics.
d. normal for a single proportion
e. 1.12
f. *p*-value = 0.1308
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: The *p*-value is greater than 0.05.
iv. Conclusion: There is insufficient evidence to conclude that less than 60 percent of her students feel more enriched.

i. Confidence interval: (0.409, 0.654)
The "plus-4s" confidence interval is (0.411, 0.648)

82. a. H_0: $\mu = 4$
b. H_a: $\mu \neq 4$
c. Let $\overline{X}$ the average I.Q. of a set of brown trout.
d. two-tailed Student's *t*-test
e. $t = 1.95$
f. *p*-value = 0.076
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Reject the null hypothesis.
iii. Reason for decision: The *p*-value is greater than 0.05
iv. Conclusion: There is insufficient evidence to conclude that the average IQ of brown trout is not four.

i. (3.8865, 5.9468)

84. a. H_0: $p \geq 0.13$
b. H_a: $p < 0.13$
c. Let P' = the proportion of Americans who have the disease
d. normal for a single proportion
e. –2.688
f. *p*-value = 0.0036
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Reject the null hypothesis.
iii. Reason for decision: The *p*-value is less than 0.05.
iv. Conclusion: There is sufficient evidence to conclude that the percentage of Americans who have been diagnosed with the disease is less than 13 percent.

i. (0, 0.0623).
The *plus-4s* confidence interval is (0.0022, 0.0978)

86.
a. H_0: $\mu \geq 129$
b. H_a: $\mu < 129$
c. Let $\overline{X}$ = the average time in seconds that Terri finishes Lap 4.
d. Student's *t*-distribution
e. $t = 1.209$
f. 0.8792
g. Check student's solution.
h.
   i. Alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The *p*-value is greater than 0.05.
   iv. Conclusion: There is insufficient evidence to conclude that Terri's mean lap time is less than 129 seconds.

i. (128.63, 130.37)

88.
a. H_0: $p = 0.60$
b. H_a: $p < 0.60$
c. Let P' = the proportion of family members who shed tears at a reunion.
d. normal for a single proportion
e. –1.71
f. 0.0438
g. Check student's solution.
h.
   i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: *p*-value < alpha
   iv. Conclusion: At the 5 percent significance level, there is sufficient evidence to conclude that the proportion of family members who shed tears at a reunion is less than 0.60. However, the test is weak because the *p*-value and alpha are quite close, so other tests should be done.

i. We are 95 percent confident that between 38.29 percent and 61.71 percent of family members will shed tears at a family reunion. (0.3829, 0.6171). The *plus-4s* confidence interval (see chapter 8) is (0.3861, 0.6139)

Note that here the *large-sample* 1 – PropZTest provides the approximate *p*-value of 0.0438. Whenever a *p*-value based on a normal approximation is close to the level of significance, the exact *p*-value based on binomial probabilities should be calculated whenever possible. This is beyond the scope of this course.

89.
a. H_0: $\mu \geq 22$
b. H_a: $\mu < 22$
c. Let $\overline{X}$ = the mean number of bubbles per blow.
d. Student's *t*-distribution
e. –2.667
f. *p*-value = 0.00486
g. Check student's solution.
h.
   i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: The *p*-value is less than 0.05.
   iv. Conclusion: There is sufficient evidence to conclude that the mean number of bubbles per blow is less than 22.

i. (18.501, 21.499)

91.
a. H_0: $\mu \leq 1$
b. H_a: $\mu > 1$
c. Let $\overline{X}$ = the mean cost in dollars of macaroni and cheese in a certain town.
d. Student's *t*-distribution
e. $t = 0.340$

f. *p*-value = 0.36756
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: The *p*-value is greater than 0.05
iv. Conclusion: The mean cost could be $1, or less. At the 5 percent significance level, there is insufficient evidence to conclude that the mean price of a box of macaroni and cheese is more than $1.

i. (0.8291, 1.241)

93. a. H_0: $p = 0.01$
b. H_a: $p > 0.01$
c. Let P' = the proportion of errors generated
d. Normal for a single proportion
e. 2.13
f. 0.0165
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Reject the null hypothesis.
iii. Reason for decision: The *p*-value is less than 0.05.
iv. Conclusion: At the 5 percent significance level, there is sufficient evidence to conclude that the proportion of errors generated is more than 0.01.

i. Confidence interval: (0, 0.094).
The *plus-4s* confidence interval is (0.004, 0.144).

95. a. H_0: $p = 0.50$
b. H_a: $p < 0.50$
c. Let P' = the proportion of friends that has a pierced ear.
d. normal for a single proportion
e. –1.70
f. *p*-value = 0.0448
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Reject the null hypothesis.
iii. Reason for decision: The *p*-value is less than 0.05. (However, they are very close.)
iv. Conclusion: There is sufficient evidence to support the claim that less than 50 percent of his friends have pierced ears.

i. Confidence interval: (0.245, 0.515): The *plus-4s* confidence interval is (0.259, 0.519).

97. a. H_0: $p = 0.40$
b. H_a: $p < 0.40$
c. Let P' = the proportion of schoolmates who fear public speaking.
d. normal for a single proportion
e. –1.01
f. *p*-value = 0.1563
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: The *p*-value is greater than 0.05.
iv. Conclusion: There is insufficient evidence to support the claim that less than 40 percent of students at the school fear public speaking.

i. Confidence interval: (0.3241, 0.4240): The *plus-4s* confidence interval is (0.3257, 0.4250).

99. a. H_0: $p = 0.14$
b. H_a: $p < 0.14$

c. Let P' = the proportion of nursing home residents that have the disease.
d. normal for a single proportion
e. –0.2756
f. *p*-value = 0.3914
g. Check student's solution.
h. i. Alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The *p*-value is greater than 0.05.
   iv. At the 5 percent significance level, there is insufficient evidence to conclude that the proportion of nursing home residents that have the disease is less than 0.14.
i. Confidence interval: (0.0502, 0.2070): The *plus-4s* confidence interval (see chapter 8) is (0.0676, 0.2297).

101. a. H_0: $\mu = 69{,}110$
b. H_a: $\mu > 69{,}110$
c. Let $\overline{X}$ = the mean salary in dollars for California registered nurses.
d. Student's *t*-distribution
e. $t = 1.719$
f. *p*-value: 0.0466
g. Check student's solution.
h. i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: The *p*-value is less than 0.05.
   iv. Conclusion: At the 5 percent significance level, there is sufficient evidence to conclude that the mean salary of California registered nurses exceeds $69,110.
i. ($68,757, $73,485)

103. a. H_0: $p \geq 0.14$, H_a: $p < 0.14$
b. *p*-value < 0.0002
c. Alpha: 0.05
d. Reject the null hypothesis.
e. At the 5 percent significance level, there is sufficient evidence to conclude that the proportion of Harleys stolen is significantly less than their share of all motorcycles. (conclusion a)

105. c

107. a. H_0: $p = 0.488$ H_a: $p \neq 0.488$
b. *p*-value = 0.0114
c. alpha = 0.05
d. Reject the null hypothesis.
e. At the 5 percent level of significance, there is enough evidence to conclude that 48.8 percent of families own stocks.
f. The survey does not appear to be accurate.

109. a. H_0: $p = 0.517$ H_a: $p \neq 0.517$
b. *p*-value = 0.9203.
c. alpha = 0.05.
d. Do not reject the null hypothesis.
e. At the 5 percent significance level, there is not enough evidence to conclude that the proportion of homes in Kentucky that are heated by natural gas is 0.517.
f. However, we cannot generalize this result to the entire nation. First, the sample's population is only the state of Kentucky. Second, it is reasonable to assume that homes in the extreme north and south will have extreme high usage and low usage, respectively. We would need to expand our sample base to include these possibilities if we wanted to generalize this claim to the entire nation.

111. a. H_0: $\mu \geq 11.52$ H_a: $\mu < 11.52$
b. p-value = 0.000002 which is almost 0.
c. alpha = 0.05.
d. Reject the null hypothesis.
e. At the 5 percent significance level, there is enough evidence to conclude that the mean amount of summer rain in the northeaster US is less than 11.52 inches, on average.
f. We would make the same conclusion if alpha was 1 percent because the p-value is almost 0.

113. a. H_0: $\mu \leq 5.8$ H_a: $\mu > 5.8$
b. p-value = 0.9987
c. alpha = 0.05
d. Do not reject the null hypothesis.
e. At the 5 percent level of significance, there is not enough evidence to conclude that a woman visits her doctor, on average, more than 5.8 times a year.

115. a. H_0: $\mu \geq 150$ H_a: $\mu < 150$
b. p-value = 0.0622
c. alpha = 0.01
d. Do not reject the null hypothesis.
e. At the 1 percent significance level, there is not enough evidence to conclude that freshmen students study less than 2.5 hours per day, on average.
f. The student academic group's claim appears to be correct.

10

HYPOTHESIS TESTING WITH TWO SAMPLES

Figure 10.1 If you want to test a claim that involves two groups (the types of breakfasts eaten east and west of the Mississippi River), you can use a slightly different technique when conducting a hypothesis test. (credit: Chloe Lim)

Chapter Objectives

By the end of this chapter, the student should be able to do the following:

- Classify hypothesis tests by type
- Conduct and interpret hypothesis tests for two population means, population standard deviations known
- Conduct and interpret hypothesis tests for two population means, population standard deviations unknown
- Conduct and interpret hypothesis tests for two population proportions
- Conduct and interpret hypothesis tests for matched or paired samples

Introduction

Studies often compare two groups. For example, researchers are interested in the effect aspirin has in preventing heart attacks. Over the last few years, newspapers and magazines have reported various aspirin studies involving two groups. Typically, one group is given aspirin and the other group is given a placebo. Then, the heart attack rate is studied over several years.

There are other situations that deal with the comparison of two groups. For example, studies compare various diet and exercise programs. Politicians compare the proportion of individuals from different income brackets who might vote for them. Students are interested in whether the SAT or GRE preparatory courses really help raise their scores.

You have learned to conduct hypothesis tests on single means and single proportions. You will expand upon that in this chapter. You will compare two means or two proportions to each other. The general procedure is the same, just expanded.

To compare two means or two proportions, you work with two groups. The groups are classified as *independent groups* or *matched pairs*. *Independent groups* consist of two samples that are independent, that is, sample values selected from

one population are not related in any way to sample values selected from the other population. *Matched pairs* consist of two samples that are dependent. The parameter tested using matched pairs is the population mean. The parameters tested using independent groups are either population means or population proportions.

NOTE
This chapter relies on either a calculator or a computer to calculate the degrees of freedom, the test statistics, and *p* values. TI-83+ and TI-84 instructions are included, as well as the test statistic formulas. When using a TI-83+ or TI-84 calculator, we do not need to separate two population means, independent groups, or population variances unknown into large and small sample sizes. However, most statistical computer software has the ability to differentiate these tests.

This chapter deals with the following hypothesis tests:

- Independent groups (samples are independent)
  - Test of two population means
  - Test of two population proportions
- Matched or paired samples (samples are dependent)
  - Test of the two population proportions by testing one population mean of differences

10.1 Two Population Means with Unknown Standard Deviations

1. The two independent samples are simple random samples from two distinct populations.
2. For the two distinct populations
   - if the sample sizes are small, the distributions are important (should be normal), and
   - if the sample sizes are large, the distributions are not important (need not be normal)

The test comparing two independent population means with unknown and possibly unequal population standard deviations is called the Aspin-Welch *t*-test. The degrees of freedom formula was developed by Aspin-Welch.

The comparison of two population means is very common. A difference between the two samples depends on both the means and the standard deviations. Very different means can occur by chance if there is great variation among the individual samples. To account for the variation, we take the difference of the sample means, $\overline{X}_1 - \overline{X}_2$, and divide by the standard error to standardize the difference. The result is a *t*-score test statistic.

Because we do not know the population standard deviations, we estimate them using the two sample standard deviations from our independent samples. For the hypothesis test, we calculate the estimated **standard deviation**, or **standard error**, of the difference in sample means, $\overline{X}_1 - \overline{X}_2$.

The standard error is calculated as follows:

$$\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}} \quad \text{10.1}$$

The test statistic (*t*-score) is calculated as follows:

$$\frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

where

- s_1 and s_2, the sample standard deviations, are estimates of σ_1 and σ_2, respectively,
- σ_1 and σ_1 are the unknown population standard deviations,
- $\overline{x}_1$ and $\overline{x}_2$ are the sample means, and
- μ_1 and μ_2 are the population means.

The number of **degrees of freedom (*df*)** requires a somewhat complicated calculation. However, a computer or calculator calculates it easily. The *df* are not always a whole number. The test statistic calculated previously is approximated by the Student's *t*-distribution with *df* as follows:

Degrees of freedom

$$df = \frac{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{(s_1)^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{(s_2)^2}{n_2}\right)^2}$$

When both sample sizes n_1 and n_2 are five or larger, the Student's t approximation is very good. Notice that the sample variances $(s_1)^2$ and $(s_2)^2$ are not pooled. (If the question comes up, do not pool the variances.)

It is not necessary to compute this by hand. A calculator or computer easily computes it.

EXAMPLE 10.1

Independent groups

The average amount of time boys and girls aged 7 to 11 spend playing sports each day is believed to be the same. A study is done and data are collected, resulting in the data in Table 10.1. Each populations has a normal distribution.

	Sample Size	Average Number of Hours Playing Sports per Day	Sample Standard Deviation
Girls	9	2	0.866
Boys	16	3.2	1.00

Table 10.1

Problem

Is there a difference in the mean amount of time boys and girls aged 7 to 11 play sports each day? Test at the 5 percent level of significance.

Solution

The population standard deviations are not known. Let g be the subscript for girls and b be the subscript for boys. Then, μ_g is the population mean for girls and μ_b is the population mean for boys. This is a test of two **independent groups**, two population **means**.

Random variable: $\overline{X}_g - \overline{X}_b$ = difference in the sample mean amount of time girls and boys play sports each day.

H_0: $\mu_g = \mu_b$ H_0: $\mu_g - \mu_b = 0$

H_a: $\mu_g \neq \mu_b$ H_a: $\mu_g - \mu_b \neq 0$

The words *the same* tell you H_0 has an "=". Since there are no other words to indicate H_a, assume it says *is different.* This is a two-tailed test.

Distribution for the test: Use t_{df} where df is calculated using the df formula for independent groups, two population means. Using a calculator, df is approximately 18.8462. *Do not pool the variances.*

Calculate the *p*-value using a Student's *t*-distribution: p-value = 0.0054

Graph:

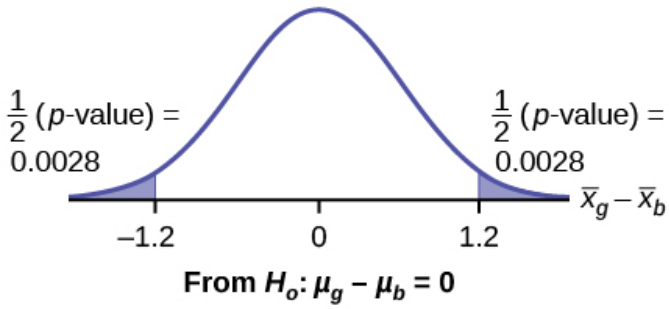


Figure 10.2

$s_g = 0.866$

$s_b = 1$
So, $\overline{x}_g - \overline{x}_b = 2 - 3.2 = -1.2$
Half the *p*-value is below –1.2, and half is above 1.2.

Make a decision: Since $\alpha > p$-value, reject H_0. This means you reject $\mu_g = \mu_b$. The means are different.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Press STAT. Arrow over to TESTS and press `4:2-SampTTest`. Arrow over to `Stats` and press ENTER. Arrow down and enter 2 for the first sample mean, `0.866` for Sx1, 9 for n1, `3.2` for the second sample mean, 1 for Sx2, and 16 for n2. Arrow down to `μ1:` and arrow to `does not equal` μ2. Press ENTER. Arrow down to `Pooled:` and `No`. Press ENTER. Arrow down to `Calculate` and press ENTER. The *p*-value is $p = 0.0054$, the *df*s are approximately 18.8462, and the test statistic is –3.14. Do the procedure again, but instead of `Calculate` do `Draw`.

Conclusion—: At the 5 percent level of significance, the sample data show there is sufficient evidence to conclude that the mean number of hours that girls and boys aged 7 to 11 play sports per day is different (mean number of hours boys aged 7 to 11 play sports per day is greater than the mean number of hours played by girls OR the mean number of hours girls aged 7 to 11 play sports per day is greater than the mean number of hours played by boys).

TRY IT 10.1

Two samples are shown in Table 10.2. Both have normal distributions. The means for the two populations are thought to be the same. Is there a difference in the means? Test at the 5 percent level of significance.

	Sample Size	Sample Mean	Sample Standard Deviation
Population A	25	5	1
Population B	16	4.7	1.2

Table 10.2

NOTE

When the sum of the sample sizes is larger than 30 ($n_1 + n_2 > 30$), you can use the normal distribution to approximate the Student's *t*.

EXAMPLE 10.2

A study is done by a community group in two neighboring colleges to determine which one graduates students with more math classes. College A samples 11 graduates. Their average is 4 math classes with a standard deviation of 1.5 math classes. College B samples nine graduates. Their average is 3.5 math classes with a standard deviation of 1 math class. The community group believes that a student who graduates from College A *has taken more math classes,* on average. Both populations have a normal distribution. Test at a 1 percent significance level. Answer the following questions:

Problem

a. Is this a test of two means or two proportions?

Solution
a. two means

Problem
b. Are the populations standard deviations known or unknown?

Solution
b. unknown

Problem
c. Which distribution do you use to perform the test?

Solution
c. Student's *t*

Problem
d. What is the random variable?

Solution
d. $\overline{X}_A - \overline{X}_B$

Problem
e. What are the null and alternate hypotheses? Write the null and alternate hypotheses in symbols.

Solution
e. $H_o : \mu_A \leq \mu_B$
$H_a : \mu_A > \mu_B$

Problem
f. Is this test right-, left-, or two-tailed?

Solution
f.

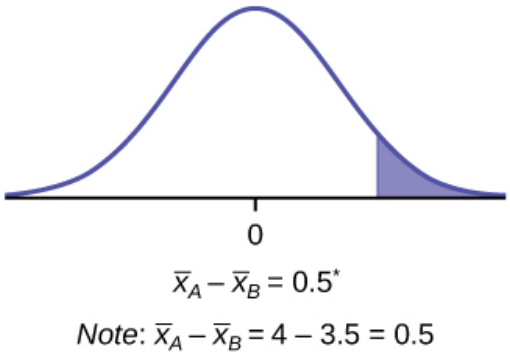


Figure 10.3

right

Problem
g. What is the *p*-value?

Solution
g. 0.1928

Problem

h. Do you reject or not reject the null hypothesis?

Solution

h. do not reject

Problem

i. **Conclusion:**

Solution

i. At the 1 percent level of significance, from the sample data, there is not sufficient evidence to conclude that a student who graduates from College A has taken more math classes, on average, than a student who graduates from College B.

TRY IT 10.2

A study is done to determine if Company A retains its workers longer than Company B. Company A samples 15 workers, and their average time with the company is 5 years with a standard deviation of 1.2. Company B samples 20 workers, and their average time with the company is 4.5 years with a standard deviation of 0.8. The populations are normally distributed.

a. Are the population standard deviations known?
b. Conduct an appropriate hypothesis test. At the 5 percent significance level, what is your conclusion?

EXAMPLE 10.3

A professor at a large community college wanted to determine whether there is a difference in the means of final exam scores between students who took his statistics course online and the students who took his face-to-face statistics class. He believed that the mean of the final exam scores for the online class would be lower than that of the face-to-face class. Was the professor correct? The randomly selected 30 final exam scores from each group are listed in Table 10.3 and Table 10.4.

67.6	41.2	85.3	55.9	82.4	91.2	73.5	94.1	64.7	64.7
70.6	38.2	61.8	88.2	70.6	58.8	91.2	73.5	82.4	35.5
94.1	88.2	64.7	55.9	88.2	97.1	85.3	61.8	79.4	79.4

Table 10.3 Online Class

77.9	95.3	81.2	74.1	98.8	88.2	85.9	92.9	87.1	88.2
69.4	57.6	69.4	67.1	97.6	85.9	88.2	91.8	78.8	71.8
98.8	61.2	92.9	90.6	97.6	100	95.3	83.5	92.9	89.4

Table 10.4 Face-to-Face Class

Problem

Is the mean of the final exam scores of the online class lower than the mean of the final exam scores of the face-to-face class? Test at a 5 percent significance level. Answer the following questions:

a. Is this a test of two means or two proportions?
b. Are the population standard deviations known or unknown?
c. Which distribution do you use to perform the test?

d. What is the random variable?
e. What are the null and alternative hypotheses? Write the null and alternative hypotheses in words and in symbols.
f. Is this test right-, left-, or two-tailed?
g. What is the *p*-value?
h. Do you reject or not reject the null hypothesis?
i. At the _____ level of significance, from the sample data, there _____ (is/is not) sufficient evidence to conclude that _____.

(See the conclusion in Example 10.2, and write yours in a similar fashion.)

USING THE TI-83, 83+, 84, 84+ CALCULATOR

First put the data for each group into two lists (such as L1 and L2). Press STAT. Arrow over to TESTS and press `4:2SampTTest`. Make sure `Data` is highlighted and press ENTER. Arrow down and enter L1 for the first list and L2 for the second list. Arrow down to μ_1`:` and arrow to < μ_2 (less than). Press ENTER. Arrow down to `Pooled: No`. Press ENTER. Arrow down to `Calculate` and press ENTER.

Note

Be careful not to mix up the information for Group 1 and Group 2!

Solution

a. two means
b. unknown
c. Student's *t*
d. $\overline{X}_1 - \overline{X}_2$
e. 1. H_0: $\mu_1 = \mu_2$ Null hypothesis: The means of the final exam scores are equal for the online and face-to-face statistics classes.
   2. H_a: $\mu_1 < \mu_2$ Alternative hypothesis: The mean of the final exam scores of the online class is less than the mean of the final exam scores of the face-to-face class.
f. left-tailed
g. *p*-value = 0.0011

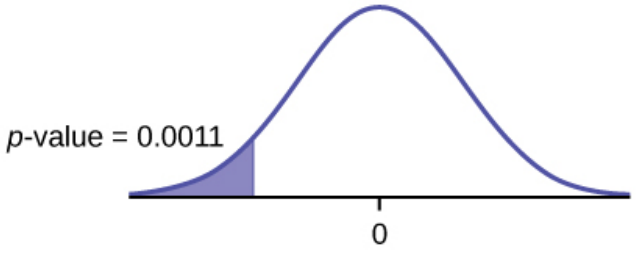


Figure 10.4

h. Reject the null hypothesis.
i. The professor was correct. The evidence shows that the mean of the final exam scores for the online class is lower than that of the face-to-face class.
   At the 5 percent level of significance, from the sample data, there is (is/is not) sufficient evidence to conclude that the mean of the final exam scores for the online class is less than the mean of final exam scores of the face-to-face class.

Cohen's Standards for Small, Medium, and Large Effect Sizes

Cohen's *d* is a measure of effect size based on the differences between two means. Cohen's *d*, named for U.S. statistician Jacob Cohen, measures the relative strength of the differences between the means of two populations based on sample

data. The calculated value of effect size is then compared to Cohen's standards of small, medium, and large effect sizes.

Size of Effect	*d*
Small	0.2
medium	0.5
Large	0.8

Table 10.5 Cohen's Standard Effect Sizes

Cohen's *d* is the measure of the difference between two means divided by the pooled standard deviation: $d = \frac{\overline{x}_1 - \overline{x}_2}{s_{pooled}}$

where $s_{pooled} = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$.

EXAMPLE 10.4

Problem

Calculate Cohen's *d* for Example 10.2. Is the size of the effect small, medium, or large? Explain what the size of the effect means for this problem.

Solution

$\mu_1 = 4$ $s_1 = 1.5$ $n_1 = 11$
$\mu_2 = 3.5$ $s_2 = 1$ $n_2 = 9$
$d = 0.384$
The effect is small because 0.384 is between Cohen's value of 0.2 for small effect size and 0.5 for medium effect size. The size of the differences of the means for the two colleges is small, indicating that there is not a significant difference between them.

EXAMPLE 10.5

Problem

Calculate Cohen's *d* for Example 10.3. Is the size of the effect small, medium, or large? Explain what the size of the effect means for this problem.

Solution

$d = 0.834$; large, because 0.834 is greater than Cohen's 0.8 for a large effect size. The size of the differences between the means of the final exam scores of online students and students in a face-to-face class is large, indicating a significant difference.

TRY IT 10.5

Weighted alpha is a measure of risk-adjusted performance of stocks over a period of a year. A high positive weighted alpha signifies a stock whose price has risen, while a small positive weighted alpha indicates an unchanged stock price during the time period. Weighted alpha is used to identify companies with strong upward or downward trends. The weighted alpha for the top 30 stocks of banks in the Northeast and in the West as identified by Nasdaq on May 24, 2013 are listed in Table 10.6 and Table 10.7, respectively.

94.2	75.2	69.6	52.0	48.0	41.9	36.4	33.4	31.5	27.6
77.3	71.9	67.5	50.6	46.2	38.4	35.2	33.0	28.7	26.5
76.3	71.7	56.3	48.7	43.2	37.6	33.7	31.8	28.5	26.0

Table 10.6 Northeast

126.0	70.6	65.2	51.4	45.5	37.0	33.0	29.6	23.7	22.6
116.1	70.6	58.2	51.2	43.2	36.0	31.4	28.7	23.5	21.6
78.2	68.2	55.6	50.3	39.0	34.1	31.0	25.3	23.4	21.5

Table 10.7 West

Is there a difference in the weighted alpha of the top 30 stocks of banks in the Northeast and in the West? Test at a 5 percent significance level. Answer the following questions:

a. Is this a test of two means or two proportions?
b. Are the population standard deviations known or unknown?
c. Which distribution do you use to perform the test?
d. What is the random variable?
e. What are the null and alternative hypotheses? Write the null and alternative hypotheses in words and in symbols.
f. Is this test right-, left-, or two-tailed?
g. What is the *p*-value?
h. Do you reject or not reject the null hypothesis?
i. At the _____ level of significance, from the sample data, there _____ (is/is not) sufficient evidence to conclude that _____.
j. Calculate Cohen's *d* and interpret it.

10.2 Two Population Means with Known Standard Deviations

Even though this situation is not likely (knowing the population standard deviations), the following example illustrates hypothesis testing for independent means, known population standard deviations. The sampling distribution for the difference between the means is normal, and both populations must be normal. The random variable is $\overline{X}_1 - \overline{X}_2$. The normal distribution has the following format:
Normal distribution is

$$\overline{X}_1 - \overline{X}_2 \sim N\left[\mu_1 - \mu_2, \sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}\right].$$

The standard deviation is $\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}$. 10.2

The test statistic (*z*-score) is $z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}}$.

EXAMPLE 10.6

Independent groups, population standard deviations known: The mean lasting time of two competing floor waxes is to be compared. Twenty floors are randomly assigned to test each wax. Both populations have a normal distribution.

The data are recorded in Table 10.8.

Wax	Sample Mean Number of Months Floor Wax Lasts	Population Standard Deviation
1	3	0.33
2	2.9	0.36

Table 10.8

Problem

Does the data indicate that Wax 1 is more effective than Wax 2? Test at a 5 percent level of significance.

Solution

This is a test of two independent groups, two population means, population standard deviations known.

Random Variable: $\overline{X}_1 - \overline{X}_2$ = difference in the mean number of months the competing floor waxes last.

H_0: $\mu_1 \leq \mu_2$

H_a: $\mu_1 > \mu_2$

The words *is more effective* says that **Wax 1 lasts longer than Wax 2**, on average. *Longer* is a > symbol and goes into H_a. Therefore, this is a right-tailed test.

Distribution for the test: The population standard deviations are known, so the distribution is normal. Using the formula, the distribution is

$$\overline{X}_1 - \overline{X}_2 \sim N\left(0, \sqrt{\frac{0.33^2}{20} + \frac{0.36^2}{20}}\right).$$

Since $\mu_1 \leq \mu_2$, then $\mu_1 - \mu_2 \leq 0$ and the mean for the normal distribution is zero.

Calculate the *p* value using the normal distribution: *p* value = 0.1799

Graph:

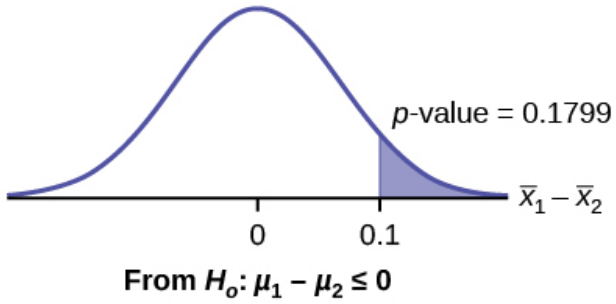


Figure 10.5

$\overline{X}_1 - \overline{X}_2 = 3 - 2.9 = 0.1$

Compare *α* and the *p* value: $\alpha = 0.05$ and *p* value = 0.1799. Therefore, $\alpha < p$ value.

Make a decision: Since $\alpha < p$ value, do not reject H_0.

Conclusion: At the 5 percent level of significance, from the sample data, there is not sufficient evidence to conclude that the mean time Wax 1 lasts is longer (Wax 1 is more effective) than the mean time Wax 2 lasts.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Press STAT. Arrow over to TESTS and press `3:2-SampZTest`. Arrow over to `Stats` and press ENTER. Arrow down and enter `.33` for sigma1, `.36` for sigma2, 3 for the first sample mean, `20` for n1, `2.9` for the second sample mean, and `20` for n2. Arrow down to μ`1:` and arrow to > μ_2. Press ENTER. Arrow down to `Calculate` and press ENTER. The p value is $p = 0.1799$, and the test statistic is 0.9157. Do the procedure again, but instead of `Calculate` do `Draw`.

TRY IT 10.6

The means of the number of revolutions per minute of two competing engines are to be compared. Thirty engines are randomly assigned to be tested. Both populations have normal distributions. Table 10.9 shows the result. Do the data indicate that Engine 2 has higher RPM than Engine 1? Test at a 5 percent level of significance.

Engine	Sample Mean Number of RPM	Population Standard Deviation
1	1,500	50
2	1,600	60

Table 10.9

EXAMPLE 10.7

An interested citizen wanted to know if Democratic U.S. senators are older than Republican U.S. senators, on average. On May 26, 2013, the mean age of 30 randomly selected Republican senators was 61 years 247 days (61.675 years) with a standard deviation of 10.17 years. The mean age of 30 randomly selected Democratic senators was 61 years 257 days (61.704 years) with a standard deviation of 9.55 years.

Problem

Do the data indicate that Democratic senators are older than Republican senators, on average? Test at a 5 percent level of significance.

Solution

This is a test of two independent groups, two population means. The population standard deviations are unknown, but the sum of the sample sizes is 30 + 30 = 60, which is greater than 30, so we can use the normal approximation to the Student's-t distribution.
Subscripts: 1: Democratic senators; 2: Republican senators

Random variable: $\overline{X}_1 - \overline{X}_2$ = difference in the mean age of Democratic and Republican U.S. senators.

H_0: $\mu_1 \leq \mu_2$ H_0: $\mu_1 - \mu_2 \leq 0$

H_a: $\mu_1 > \mu_2$ H_a: $\mu_1 - \mu_2 > 0$

The words *older than* translates as a > symbol and goes into H_a. Therefore, this is a right-tailed test.

Distribution for the test: The distribution is the normal approximation to the Student's t for means, independent groups. Using the formula, the distribution is

$$\overline{X}_1 - \overline{X}_2 \sim N[0, \sqrt{\frac{(9.55)^2}{30} + \frac{(10.17)^2}{30}}]$$

Since $\mu_1 \leq \mu_2$, $\mu_1 - \mu_2 \leq 0$ and the mean for the normal distribution is zero.

Calculating the p value using the normal distribution gives p value = 0.4040.

Graph:

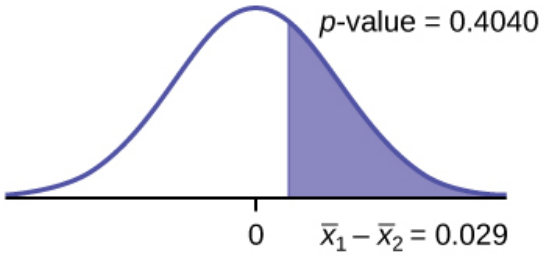


Figure 10.6

Compare α and the p value: $\alpha = 0.05$ and p value = 0.4040. Therefore, $\alpha < p$ value.

Make a decision: Since $\alpha < p$ value, do not reject H_0.

Conclusion: At the 5 percent level of significance, from the sample data, there is not sufficient evidence to conclude that the mean age of Democratic senators is greater than the mean age of the Republican senators.

10.3 Comparing Two Independent Population Proportions

When conducting a hypothesis test that compares two independent population proportions, the following characteristics should be present:

1. The two independent samples are simple random samples that are independent.
2. The number of successes is at least five, and the number of failures is at least five, for each of the samples.
3. Growing literature states that the population must be at least 10 or 20 times the size of the sample. This keeps each population from being over-sampled and causing incorrect results.

Comparing two proportions, like comparing two means, is common. If two estimated proportions are different, it may be due to a difference in the populations or it may be due to chance. A hypothesis test can help determine if a difference in the estimated proportions reflects a difference in the population proportions.

The difference of two proportions follows an approximate normal distribution. Generally, the null hypothesis states that the two proportions are the same. That is, H_0: $p_A = p_B$. To conduct the test, we use a **pooled proportion**, p_c.

The pooled proportion is calculated as follows:

$$p_c = \frac{x_A + x_B}{n_A + n_B}. \quad \textbf{10.3}$$

The distribution for the differences is

$$P'_A - P'_B \sim N[0, \sqrt{p_c(1-p_c)(\frac{1}{n_A} + \frac{1}{n_B})}].$$

The test statistic (z-score) is

$$z = \frac{(p'_A - p'_B) - (p_A - p_B)}{\sqrt{p_c(1-p_c)(\frac{1}{n_A} + \frac{1}{n_B})}}.$$

EXAMPLE 10.8

Problem

Two types of medication for hives are being tested to determine if there is a difference in the proportions of adult patient reactions. Twenty out of a random sample of 200 adults given Medication A still had hives 30 minutes after taking the medication. Twelve out of another random sample of 200 adults given Medication B still had hives 30 minutes after taking the medication. Test at a 1 percent level of significance.

Solution

The problem asks for a difference in proportions, making it a test of two proportions.

Let *A* and *B* be the subscripts for Medication A and Medication B, respectively. Then, p_A and p_B are the desired population proportions.

Random Variable:

$P'_A - P'_B$ = difference in the proportions of adult patients who did not react after 30 minutes to Medication A and to Medication B.

H_0: $p_A = p_B$

$p_A - p_B = 0$

H_a: $p_A \neq p_B$

$p_A - p_B \neq 0$

The words *is a difference* tell you the test is two-tailed.

Distribution for the test: Since this is a test of two binomial population proportions, the distribution is normal:

$$p_c = \frac{x_A + x_B}{n_A + n_B} = \frac{20+12}{200+200} = 0.08 \quad 1 - p_c = 0.92$$

$$P'_A - P'_B \sim N\left[0, \sqrt{(0.08)(0.92)\left(\frac{1}{200} + \frac{1}{200}\right)}\right]$$

$P'_A - P'_B$ follows an approximate normal distribution.

Calculate the *p*-value using the normal distribution: *p*-value = 0.1404.

Estimated proportion for group A: $p'_A = \frac{x_A}{n_A} = \frac{20}{200} = 0.1$

Estimated proportion for group B: $p'_B = \frac{x_B}{n_B} = \frac{12}{200} = 0.06$

Graph:

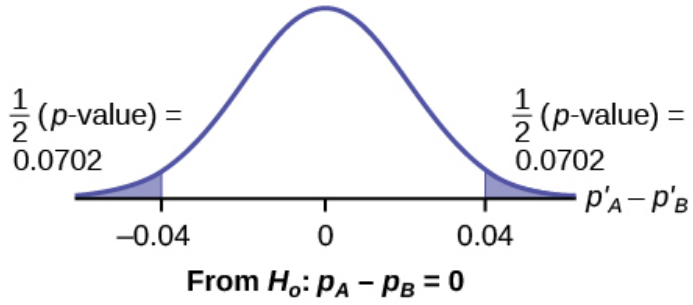


Figure 10.7

$P'_A - P'_B = 0.1 - 0.06 = 0.04$.

Half the *p*-value is below –0.04, and half is above 0.04.

Compare α and the *p*-value: $\alpha = 0.01$ and the *p*-value = 0.1404. $\alpha <$ *p*-value.

Make a decision: Since $\alpha <$ *p*-value, do not reject H_0.

Conclusion: At a 1 percent level of significance, from the sample data, there is not sufficient evidence to conclude that there is a difference in the proportions of adult patients who did not react after 30 minutes to Medication A and Medication B.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Press STAT. Arrow over to TESTS and press `6:2-PropZTest`. Arrow down and enter `20` for x1, `200` for n1, `12` for x2, and `200` for n2. Arrow down to `p1:` and arrow to `not equal p2`. Press ENTER. Arrow down to `Calculate` and press ENTER. The *p*-value is $p = 0.1404$, and the test statistic is 1.47. Do the procedure again, but instead of `Calculate` do `Draw`.

TRY IT 10.8

Two types of valves are being tested to determine if there is a difference in pressure tolerances. Fifteen out of a random sample of 100 of Valve A cracked under 4,500 psi. Six out of a random sample of 100 of Valve B cracked under 4,500 psi. Test at a 5 percent level of significance.

EXAMPLE 10.9

Problem

A research study was conducted about gender differences in texting. The researcher believed that the proportion of girls involved in texting is less than the proportion of boys involved. The data collected in spring 2010 among a random sample of middle and high school students in a large school district in the southern United States is summarized in Table 10.10. Is the proportion of girls sending texts less than the proportion of boys texting? Test at a 1 percent level of significance.

	Males	Females
Sent texts	183	156
Total number surveyed	2231	2169

Table 10.10

Solution

This is a test of two population proportions. Let M and F be the subscripts for males and females. Then, p_M and p_F are the desired population proportions.

Random variable:

$p'_F - p'_M$ = difference in the proportions of males and females who sent texts.

H_0: $p_F = p_M$ H_0: $p_F - p_M = 0$

H_a: $p_F < p_M$ H_a: $p_F - p_M < 0$

The words *less than* tell you the test is left-tailed.

Distribution for the test: Since this is a test of two population proportions, the distribution is normal:

$$p_c = \frac{x_F + x_M}{n_F + n_M} = \frac{156+183}{2169+2231} = 0.077$$

$$1 - p_c = 0.923$$

Therefore,

$$p'_F - p'_M \sim N\left(0, \sqrt{(0.077)(0.923)\left(\frac{1}{2169} + \frac{1}{2231}\right)}\right)$$

$p'_F - p'_M$ follows an approximate normal distribution.

Calculate the *p*-value using the normal distribution:

p-value = 0.1045
Estimated proportion for females: 0.0719
Estimated proportion for males: 0.082

Graph:

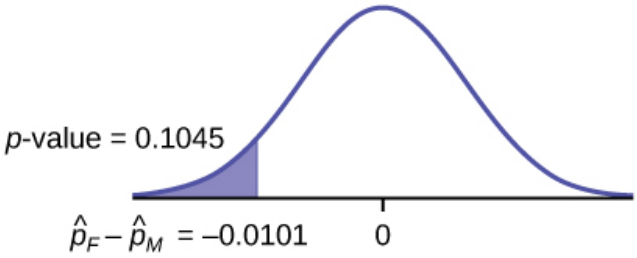


Figure 10.8

Decision: Since $\alpha < p$-value, do not reject H_0.

Conclusion: At the 1 percent level of significance, from the sample data, there is not sufficient evidence to conclude that the proportion of girls sending texts is less than the proportion of boys sending texts.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Press STAT. Arrow over to TESTS and press `6:2-PropZTest`. Arrow down and enter 156 for x1, 2169 for n1, 183 for x2, and 2231 for n2. Arrow down to `p1:` and arrow to `less than p2`. Press ENTER. Arrow down to `Calculate` and press ENTER. The *p*-value is $p = 0.1045$ and the test statistic is $z = -1.256$.

EXAMPLE 10.10

Problem

Researchers conducted a study of smartphone use (Phone A versus Phone B) among adults. A cell phone company claimed that Phone B smartphones are more popular with whites (non-Hispanic) than with African Americans. The results of the survey indicate that of the 232 African American cell phone owners randomly sampled, 5 percent own Phone B. Of the 1,343 white cell phone owners randomly sampled, 10 percent own Phone B. Test at the 5 percent level of significance. Is the proportion of white Phone B owners greater than the proportion of African American Phone B owners?

Solution

This is a test of two population proportions. Let *W* and *A* be the subscripts for the whites and African Americans. Then, p_W and p_A are the desired population proportions.

Random variable:

$p'_W - p'_A$ = difference in the proportions of Phone A and Phone B users.

H_0: $p_W = p_A$ H_0: $p_W - p_A = 0$

H_a: $p_W > p_A$ H_a: $p_W - p_A > 0$

The words *more popular* indicate that the test is right-tailed.

Distribution for the test: The distribution is approximately normal.

$$p_c = \frac{x_W + x_A}{n_W + n_A} = \frac{134+12}{1343+232} = 0.0927$$

$$1 - p_c = 0.9073$$

Therefore,

$p'_W - p'_A \backsim N\left(0, \sqrt{(0.0927)(0.9073)\left(\frac{1}{1343} + \frac{1}{232}\right)}\right)$

$p'_W - p'_A$ follows an approximate normal distribution.

Calculate the *p*-value using the normal distribution:

p-value = 0.0077
Estimated proportion for group A: 0.10
Estimated proportion for group B: 0.05

Graph:

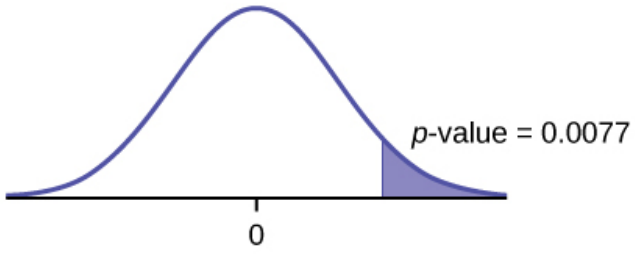


Figure 10.9

Decision: Since $\alpha > p$-value, reject the H_0.

Conclusion: At the 5 percent level of significance, from the sample data, there is sufficient evidence to conclude that a larger proportion of white cell phone owners use Phone B than African Americans.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

TI-83+ and TI-84: Press STAT. Arrow over to TESTS and press `6:2-PropZTest`. Arrow down and enter 135 for x1, 1343 for n1, 12 for x2, and 232 for n2. Arrow down to `p1:` and arrow to `greater than p2`. Press ENTER. Arrow down to `Calculate` and press ENTER. The *p*-value is $p = 0.0092$, and the test statistic is $z = 2.33$.

TRY IT 10.10

A group of citizens wanted to know if the proportion of homeowners in their small city was different in 2011 than in 2010. Their research showed that of the 113,231 available homes in their city in 2010, 7,622 of them were owned by the families who live there. In 2011, 7,439 of the 104,873 of the available homes were owned by city residents. Test at a 5 percent significance level. Answer the following questions:

a. Is this a test of two means or two proportions?

b. Which distribution do you use to perform the test?

c. What is the random variable?

d. What are the null and alternative hypotheses? Write the null and alternative hypotheses in symbols.

e. Is this test right-, left-, or two-tailed?

f. What is the *p*-value?

g. Do you reject or not reject the null hypothesis?

h. At the _____ level of significance, from the sample data, there _____ (is/is not) sufficient evidence to conclude that _____.

10.4 Matched or Paired Samples (Optional)

When using a hypothesis test for matched or paired samples, the following characteristics should be present:

1. Simple random sampling is used.
2. Sample sizes are often small.
3. Two measurements (samples) are drawn from the same pair of individuals or objects.
4. Differences are calculated from the matched or paired samples.
5. The differences form the sample that is used for the hypothesis test.
6. Either the matched pairs have differences that come from a population that is normal or the number of differences is sufficiently large so that distribution of the sample mean of differences is approximately normal.

In a hypothesis test for matched or paired samples, subjects are matched in pairs and differences are calculated. The differences are the data. The population mean for the differences, μ_d, is then tested using a Student's-*t* test for a single population mean with $n - 1$ degrees of freedom, where n is the number of differences.

The test statistic (*t*-score) is

$$t = \frac{\overline{x}_d - \mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)}. \quad \text{10.4}$$

EXAMPLE 10.11

Problem

A study was conducted to investigate the effectiveness of pain-reducing medication. Results for randomly selected subjects are shown in Table 10.11. A lower score indicates less pain. The *before* value is matched to an *after* value, and the differences are calculated. The differences have a normal distribution. Are the sensory measurements, on average, lower after the medication? Test at a 5 percent significance level.

Subject:	A	B	C	D	E	F	G	H
Before	6.6	6.5	9.0	10.3	11.3	8.1	6.3	11.6
After	6.8	2.4	7.4	8.5	8.1	6.1	3.4	2.0

Table 10.11

Solution

Corresponding *before* and *after* values form matched pairs. (Calculate *after* – *before*.)

After Data	Before Data	Difference
6.8	6.6	0.2
2.4	6.5	–4.1
7.4	9	–1.6
8.5	10.3	–1.8
8.1	11.3	–3.2
6.1	8.1	–2

Table 10.12

After Data	Before Data	Difference
3.4	6.3	–2.9
2	11.6	–9.6

Table 10.12

The data for the test are the differences: {0.2, –4.1, –1.6, –1.8, –3.2, –2, –2.9, –9.6}

The sample mean and sample standard deviation of the differences are: $x_d = -3.13$ and $s_d = 2.91$
Verify these values.

Let μ_d be the population mean for the differences. We use the subscript d to denote *differences*.

Random variable: $\overline{X}_d$ = the mean difference of the sensory measurements.

H_0: $\mu_d \geq 0$

The null hypothesis is zero or positive, meaning that there is the same or more pain felt after taking the medication. That means the subject shows no improvement. μ_d is the population mean of the differences.

H_a: $\mu_d < 0$

The alternative hypothesis is negative, meaning there is less pain felt after taking the medication. That means the subject shows improvement. The score should be lower after taking the medication, so the difference ought to be negative to indicate improvement.

Distribution for the test: The distribution is a Student's t with $df = n - 1 = 8 - 1 = 7$. Use t_7. Note —that the test is for a single population mean.

Calculate the *p*-value using the Student's-*t* distribution: p-value = 0.0095

Graph:

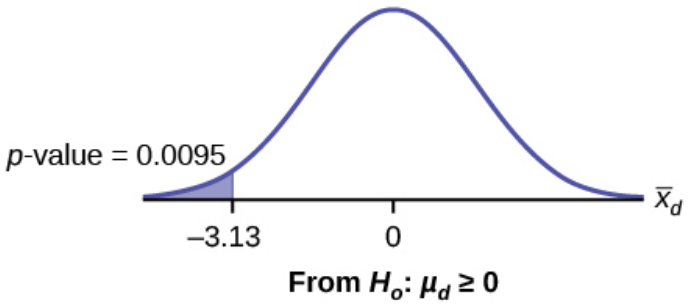


Figure 10.10

$\overline{X}_d$ is the random variable for the differences.

The sample mean and sample standard deviation of the differences are as follows:

$\overline{x}_d$ = –3.13

$\overline{s}_d$ = 2.91

Compare *α* and the *p*-value: $\alpha = 0.05$ and p-value = 0.0095. $\alpha > p$-value.

Make a decision: Since $\alpha > p$-value, reject H_0. This means that $\mu_d < 0$ and there is improvement.

Conclusion: At a 5 percent level of significance, from the sample data, there is sufficient evidence to conclude that the sensory measurements, on average, are lower after taking the medication. The medication appears to be effective in reducing pain.

Note
For the TI-83+ and TI-84 calculators, you can either calculate the differences ahead of time (*after* - *before*) and put the differences into a list or you can put the *after* data into a first list and the *before* data into a second list. Then, go to a third list and arrow up to the name. Enter `1st list name – 2nd list name`. The calculator will do the subtraction, and you will have the differences in the third list.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Use your list of differences as the data. Press STAT and arrow over to TESTS. Press `2:T-Test`. Arrow over to `Data` and press ENTER. Arrow down and enter `0` for μ_0, the name of the list where you put the data, and `1` for Freq:. Arrow down to μ: and arrow over to < μ_0. Press ENTER. Arrow down to `Calculate` and press ENTER. The *p*-value is 0.0094, and the test statistic is –3.04. Do these instructions again except, arrow to `Draw` instead of `Calculate`. Press ENTER.

TRY IT 10.11

A study was conducted to investigate how effective a new diet was in lowering cholesterol. Results for the randomly selected subjects are shown in the table. The differences have a normal distribution. Are the subjects' cholesterol levels lower on average after the diet? Test at the 5 percent level.

Subject	A	B	C	D	E	F	G	H	I
Before	209	210	205	198	216	217	238	240	222
After	199	207	189	209	217	202	211	223	201

Table 10.13

EXAMPLE 10.12

A college football coach was interested in whether the college's strength development class increased his players' maximum lift (in pounds) on the bench press exercise. He asked four of his players to participate in a study. The amount of weight they could each lift was recorded before they took the strength development class. After completing the class, the amount of weight they could each lift was again measured. The data are as follows:

Weight (in pounds)	Player 1	Player 2	Player 3	Player 4
Amount of weight lifted prior to the class	205	241	338	368
Amount of weight lifted after the class	295	252	330	360

Table 10.14

The coach wants to know if the strength development class makes his players stronger, on average.
Record the *differences* data. Calculate the differences by subtracting the amount of weight lifted prior to the class from the weight lifted after completing the class. The data for the differences are: {90, 11, -8, -8}. Assume the differences have a normal distribution.

Using the differences data, calculate the sample mean and the sample standard deviation.

$\overline{x}_d = 21.3$, $s_d = 46.7$

Note

The data given here would indicate that the distribution is right-skewed. The difference 90 may be an extreme outlier. It is pulling the sample mean to be 21.3 (positive). The means of the other three data values are negative.

Using the difference data, this becomes a test of a single _____.

Define the random variable: $\overline{X}_d$ is the mean difference in the maximum lift per player.

The distribution for the hypothesis test is t_3.

H_0: $\mu_d \leq 0$, H_a: $\mu_d > 0$

Graph:

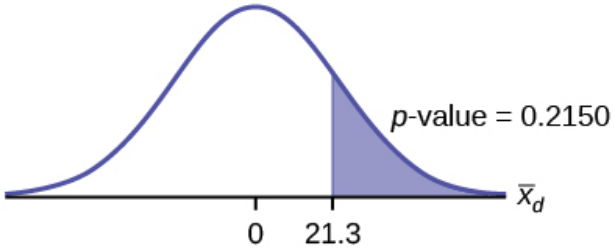


Figure 10.11

Calculate the *p*-value: The *p*-value is 0.2150.

Decision: If the level of significance is 5 percent, the decision is not to reject the null hypothesis, because $\alpha < p$-value.

What is the conclusion?

At a 5 percent level of significance, from the sample data, there is not sufficient evidence to conclude that the strength development class helped make the players stronger, on average.

TRY IT 10.12

A new prep class was designed to improve SAT test scores. Five students were selected at random. Their scores on two practice exams were recorded, one before the class and one after. The data are recorded in Table 10.15. Are the scores, on average, higher after the class? Test at a 5 percent level.

SAT Scores	Student 1	Student 2	Student 3	Student 4
Score before class	1840	1960	1920	2150
Score after class	1920	2160	2200	2100

Table 10.15

EXAMPLE 10.13

Seven eighth-graders at Kennedy Middle School measured how far they could push the shot put with their dominant (writing) hand and their weaker (nonwriting) hand. They thought that they could push equal distances with both hands. The data are collected and recorded in Table 10.16.

Distance (in feet) using	Student 1	Student 2	Student 3	Student 4	Student 5	Student 6	Student 7
Dominant Hand	30	26	34	17	19	26	20
Weaker Hand	28	14	27	18	17	26	16

Table 10.16

Conduct a hypothesis test to determine whether the mean difference in distances between the children's dominant versus weaker hands is significant.

Record the *differences* data. Calculate the differences by subtracting the distances with the weaker hand from the distances with the dominant hand. The data for the differences are: {2, 12, 7, –1, 2, 0, 4}. The differences have a normal distribution.

Using the differences data, calculate the sample mean and the sample standard deviation. $\overline{x}_d = 3.71$, $s_d = 4.5$.

Random variable: $\overline{X}_d$ = mean difference in the distances between the hands.

Distribution for the hypothesis test: t_6

H_0: $\mu_d = 0$ H_a: $\mu_d \neq 0$

Graph:

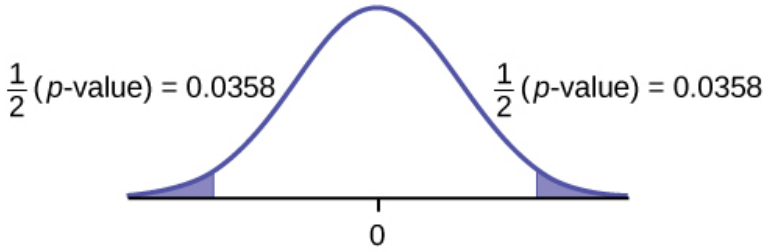


Figure 10.12

Calculate the *p*-value: The *p*-value is 0.0716 (using the data directly).

Test statistic = 2.18. *p*-value = 0.0719 using $(\overline{x}_d = 3.71,\ s_d = 4.5)$.

Decision: Assume $\alpha = 0.05$. Since $\alpha < p$-value, do not reject H_0.

Conclusion: At the 5 percent level of significance, from the sample data, there is not sufficient evidence to conclude that there is a difference in the children's weaker and dominant hands to push the shot put.

TRY IT 10.13

Five ball players think they can throw the same distance with their dominant hand (throwing) and off-hand (catching hand). The data were collected and recorded in Table 10.17. Conduct a hypothesis test to determine whether the mean 5 difference in distances between the dominant and off-hand is significant. Test at the 5 percent level.

	Player 1	Player 2	Player 3	Player 4	Player 5
Dominant Hand	120	111	135	140	125
Off-Hand	105	109	98	111	99

Table 10.17

10.5 Hypothesis Testing for Two Means and Two Proportions

Stats Lab

Hypothesis Testing for Two Means and Two Proportions

Student Learning Outcomes

- The student will select the appropriate distributions to use in each case.
- The student will conduct hypothesis tests and interpret the results.

Supplies:

- The business section from two consecutive days' newspapers
- Three small packages of multicolored chocolates
- Five small packages of peanut butter candies

Increasing Stocks Survey

Look at yesterday's newspaper business section. Conduct a hypothesis test to determine if the proportion of New York Stock Exchange (NYSE) stocks that increased is greater than the proportion of NASDAQ stocks that increased. As randomly as possible, choose 40 NYSE stocks and 32 NASDAQ stocks and complete the following statements.

1. H_0: _____
2. H_a: _____
3. In words, define the random variable.
4. The distribution to use for the test is _____.
5. Calculate the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.
   a. Graph:

Figure 10.13

   b. Calculate the *p* value.
7. Do you reject or not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Decreasing Stocks Survey

Randomly pick eight stocks from the newspaper. Using two consecutive days' business sections, test whether the stocks went down, on average, for the second day.

1. H_0: _____
2. H_a: _____
3. In words, define the random variable.
4. The distribution to use for the test is _____.
5. Calculate the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.

a. Graph

Figure 10.14

b. Calculate the *p* value:

7. Do you reject or not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Candy Survey

Buy three small packages of multicolored chocolates and five small packages of peanut butter candies (same net weight as the multicolored chocolates). Test whether the mean number of candy pieces per package is the same for the two brands.

1. H_0: _____
2. H_a: _____
3. In words, define the random variable.
4. What distribution should be used for this test?
5. Calculate the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.
   a. Graph

Figure 10.15

b. Calculate the *p* value.

7. Do you reject or not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Shoe Survey

Test whether women have, on average, more pairs of shoes than men. Include all forms of sneakers, shoes, sandals, and boots. Use your class as the sample.

1. H_0: _____
2. H_a: _____
3. In words, define the random variable.

4. The distribution to use for the test is _____.
5. Calculate the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.
   a. Graph

Figure 10.16

   b. Calculate the *p* value.
7. Do you reject or not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Key Terms

degrees of freedom (*df*) the number of objects in a sample that are free to vary
pooled proportion estimate of the common value of p_1 and p_2
standard deviation a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and σ for population standard deviation
variable (random variable) a characteristic of interest in a population being studied.
Common notation for variables are uppercase Latin letters X, Y, Z,... Common notation for a specific value from the domain (set of all possible values of a variable) are lowercase Latin letters x, y, z,.... For example, if X is the number of children in a family, then x represents a specific integer 0, 1, 2, 3, ... Variables in statistics differ from variables in intermediate algebra in two ways:

- The domain of the random variable (RV) is not necessarily a numerical set; the domain may be expressed in words; for example, if X = hair color, then the domain is {black, blond, gray, green, orange}.
- We can tell what specific value x of the random variable X takes only after performing the experiment.

Chapter Review

10.1 Two Population Means with Unknown Standard Deviations

Two population means from independent samples where the population standard deviations are not known

- Random variable: $\overline{X}_1 - \overline{X}_2$ = the difference of the sampling means
- Distribution: Student's t-distribution with degrees of freedom (variances not pooled)

10.2 Two Population Means with Known Standard Deviations

A hypothesis test of two population means from independent samples where the population standard deviations are known (typically approximated with the sample standard deviations) will have these characteristics:

- Random variable: $\overline{X}_1 - \overline{X}_2$ = the difference of the means
- Distribution: normal distribution

10.3 Comparing Two Independent Population Proportions

Test of two population proportions from independent samples

- Random variable: $\hat{p}_A - \hat{p}_B$ = difference between the two estimated proportions
- Distribution: normal distribution

10.4 Matched or Paired Samples (Optional)

A hypothesis test for matched or paired samples (t-test) has these characteristics:

- Test the differences by subtracting one measurement from the other measurement
- Random variable: $\overline{x}_d$ = mean of the differences.
- Distribution: Student's t distribution with n – 1 degrees of freedom.
- If the number of differences is small (less than 30), the differences must follow a normal distribution.
- Two samples are drawn from the same set of objects.
- Samples are dependent.

Formula Review

10.1 Two Population Means with Unknown Standard Deviations

Standard error: $SE = \sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}$

Test statistic (t-score): $t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$

Degrees of freedom:

$$df = \frac{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{(s_1)^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{(s_2)^2}{n_2}\right)^2}$$

where:

s_1 and s_2 are the sample standard deviations, and n_1 and n_2 are the sample sizes.

$\overline{x}_1$ and $\overline{x}_2$ are the sample means.

Cohen's d is the measure of effect size:

$$d = \frac{\overline{x}_1 - \overline{x}_2}{s_{pooled}}$$

where $s_{pooled} = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$.

10.2 Two Population Means with Known Standard Deviations

Normal distribution:

$$\overline{X}_1 - \overline{X}_2 \sim N\left[\mu_1 - \mu_2, \sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}\right].$$

Generally, $\boldsymbol{\mu_1 - \mu_2 = 0}$.

Test statistic (*z*-score):

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}}$$

Generally, $\boldsymbol{\mu_1 - \mu_2 = 0}$.

where
σ_1 and σ_2 are the known population standard deviations, n_1 and n_2 are the sample sizes, $\overline{x}_1$ and $\overline{x}_2$ are the sample means, and μ_1 and μ_2 are the population means.

10.3 Comparing Two Independent Population Proportions

Pooled proportion: $p_c = \frac{x_F + x_M}{n_F + n_M}$

Distribution for the differences:

$$p'_A - p'_B \sim N\left[0, \sqrt{p_c(1-p_c)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}\right]$$

where the null hypothesis is H_0: $p_A = p_B$ or H_0: $p_A - p_B = 0$

Test statistic (*z*-score): $z = \frac{(p'_A - p'_B)}{\sqrt{p_c(1-p_c)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$

where the null hypothesis is H_0: $p_A = p_B$ or H_0: $p_A - p_B = 0$

and where

p'_A and p'_B are the sample proportions, p_A and p_B are the population proportions,

P_c is the pooled proportion, and n_A and n_B are the sample sizes.

10.4 Matched or Paired Samples (Optional)

Test statistic (*t*-score): $t = \frac{\overline{x}_d - \mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)}$

where:

$\overline{x}_d$ is the mean of the sample differences, μ_d is the mean of the population differences, s_d is the sample standard deviation of the differences, and n is the sample size.

Practice

10.1 Two Population Means with Unknown Standard Deviations

Use the following information to answer the next 15 exercises. Indicate if the hypothesis test is for

a. independent group means, population standard deviations, and/or variances known,
b. independent group means, population standard deviations, and/or variances unknown,
c. matched or paired samples,
d. single mean,
e. two proportions, or
f. single proportion.

1. It is believed that 70 percent of males pass their drivers test in the first attempt, while 65 percent of females pass the test in the first attempt. Of interest is whether the proportions are equal.

2. A new laundry detergent is tested on consumers. Of interest is the proportion of consumers who prefer the new brand over the leading competitor. A study is done to test this.

3. A new windshield treatment claims to repel water more effectively. Ten windshields are tested by simulating rain without the new treatment. The same windshields are then treated, and the experiment is run again. A hypothesis test is conducted.

4. The known standard deviation in salary for all mid-level professionals in the financial industry is $11,000. Company A and Company B are in the financial industry. Suppose samples are taken of mid-level professionals from Company A and from Company B. The sample mean salary for mid-level professionals in Company A is $80,000. The sample mean salary for mid-level professionals in Company B is $96,000. Company A and Company B management want to know if their mid-level professionals are paid differently, on average.

5. The average worker in Germany gets eight weeks of paid vacation.

6. According to a television commercial, 80% of dentists agree that a brand of fluoridated toothpaste is the best on the market.

7. It is believed that the average grade on an English essay in a particular school system is higher for females than for males. A random sample of 31 females had a mean score of 82 with a standard deviation of 3, and a random sample of 25 males had a mean score of 76 with a standard deviation of 4.

8. The league mean batting average is 0.280 with a known standard deviation of 0.06. The Rattlers and the Vikings belong to the league. The mean batting average for a sample of eight Rattlers is 0.210, and the mean batting average for a sample of eight Vikings is 0.260. There are 24 players on the Rattlers and 19 players on the Vikings. Are the batting averages of the Rattlers and Vikings statistically different?

9. In a random sample of 100 forests in the United States, 56 were coniferous or contained conifers. In a random sample of 80 forests in Mexico, 40 were coniferous or contained conifers. Is the proportion of conifers in the United States statistically more than the proportion of conifers in Mexico?

10. A new medicine is said to help improve sleep. Eight subjects are picked at random and given the medicine. The mean hours slept for each person were recorded before starting the medication and after.

11. It is thought that teenagers sleep more than adults on average. A study is done to verify this. A sample of 16 teenagers has a mean of 8.9 hours slept and a standard deviation of 1.2. A sample of 12 adults has a mean of 6.9 hours slept and a standard deviation of 0.6.

12. Varsity athletes practice five times a week, on average.

13. A sample of 12 in-state graduate school programs at School A has a mean tuition of $64,000 with a standard deviation of $8,000. At School B, a sample of 16 in-state graduate programs has a mean tuition of $80,000 with a standard deviation of $6,000. On average, are the mean tuitions different?

14. A new WiFi range booster is being offered to consumers. A researcher tests the native range of 12 different routers under the same conditions. The ranges are recorded. Then, the researcher uses the new WiFi range booster and records the new ranges. Does the new WiFi range booster do a better job?

15. A high school principal claims that 30 percent of student athletes drive themselves to school, while 4 percent of nonathletes drive themselves to school. In a sample of 20 student athletes, 45 percent drive themselves to school. In a sample of 35 nonathlete students, 6 percent drive themselves to school. Is the percent of student athletes who drive themselves to school more than the percent of nonathletes?

Use the following information to answer the next three exercises: A study is done to determine which of two soft drinks has more sugar. There are 13 cans of Beverage A in a sample and six cans of Beverage B. The mean amount of sugar in Beverage A is 36 grams with a standard deviation of 0.6 grams. The mean amount of sugar in Beverage B is 38 grams with a standard deviation of 0.8 grams. The researchers believe that Beverage B has more sugar than Beverage A, on average. Both populations have normal distributions.

16. Are standard deviations known or unknown?

17. What is the random variable?

18. Is this a one-tailed or two-tailed test?

Use the following information to answer the next 12 exercises. The U.S. Centers for Disease Control reports that the mean life expectancy was 47.6 years for whites born in 1900 and 33.0 years for nonwhites. Suppose that you randomly survey death records for people born in 1900 in a certain county. Of the 124 whites, the mean life span was 45.3 years with a standard deviation of 12.7 years. Of the 82 nonwhites, the mean life span was 34.1 years with a standard deviation of 15.6 years. Conduct a hypothesis test to see if the mean life spans in the county were the same for whites and nonwhites.

19. Is this a test of means or proportions?

20. State the null and alternative hypotheses.
   a. H_0: _____
   b. H_a: _____

21. Is this a right-tailed, left-tailed, or two-tailed test?

22. In symbols, what is the random variable of interest for this test?

23. In words, define the random variable of interest for this test.

24. Which distribution (normal or Student's *t*) would you use for this hypothesis test?

25. Explain why you chose the distribution you did for Exercise 10.24.

26. Calculate the test statistic and *p*-value.

27. Sketch a graph of the situation. Label the horizontal axis. Mark the hypothesized difference and the sample difference. Shade the area corresponding to the *p*-value.

28. Find the *p*-value.

29. At a preconceived $\alpha = 0.05$, write the following:
   a. Your decision:
   b. The reason for your decision:
   c. Your conclusion (write out in a complete sentence):

30. Does it appear that the means are the same? Why or why not?

10.2 Two Population Means with Known Standard Deviations

Use the following information to answer the next five exercises. The mean speeds of fastball pitches from two different baseball pitchers are to be compared. A sample of 14 fastball pitches is measured from each pitcher. The populations have normal distributions. Table 10.18 shows the result. Scouters believe that Rodriguez pitches a speedier fastball.

Pitcher	Sample Mean Speed of Pitches (mph)	Population Standard Deviation
Wesley	86	3
Rodriguez	91	7

Table 10.18

31. What is the random variable?

32. State the null and alternative hypotheses.

33. What is the test statistic?

34. What is the *p* value?

35. At the 1 percent significance level, what is your conclusion?

Use the following information to answer the next five exercises. A researcher is testing the effects of plant food on plant growth. Nine plants have been given the plant food. Another nine plants have not been given the plant food. The heights of the plants are recorded after eight weeks. The populations have normal distributions. The following table is the result. The researcher thinks the food makes the plants grow taller.

Plant Group	Sample Mean Height of Plants (inches)	Population Standard Deviation
Food	16	2.5
No food	14	1.5

Table 10.19

36. Is the population standard deviation known or unknown?

37. State the null and alternative hypotheses.

38. What is the *p* value?

39. Draw the graph of the *p* value.

40. At the 1 percent significance level, what is your conclusion?

Use the following information to answer the next five exercises. Two metal alloys are being considered as material for ball bearings. The mean melting point of the two alloys is to be compared. Fifteen pieces of each metal are being tested. Both populations have normal distributions. The following table is the result. It is believed that Alloy Zeta has a different melting point.

	Sample Mean Melting Temperatures (°F)	Population Standard Deviation
Alloy Gamma	800	95
Alloy Zeta	900	105

Table 10.20

41. State the null and alternative hypotheses.

42. Is this a right-, left-, or two-tailed test?

43. What is the *p* value?

44. Draw the graph of the *p* value.

45. At the 1 percent significance level, what is your conclusion?

10.3 Comparing Two Independent Population Proportions

Use the following information for the next five exercises. Two types of phone operating system are being tested to determine if there is a difference in the proportions of system failures (crashes). Fifteen out of a random sample of 150 phones with OS_1 had system failures within the first eight hours of operation. Nine out of another random sample of 150 phones with OS_2 had system failures within the first eight hours of operation. OS_2 is believed to be more stable (have fewer crashes) than OS_1.

46. Is this a test of means or proportions?

47. What is the random variable?

48. State the null and alternative hypotheses.

49. What is the *p*-value?

50. What can you conclude about the two operating systems?

Use the following information to answer the next 12 exercises. In the recent U.S. Census, 3 percent of the U.S. population reported being of two or more races. However, the percent varies tremendously from state to state. Suppose that two random surveys are conducted. In the first random survey, out of 1,000 North Dakotans, only 9 people reported being of two or more races. In the second random survey, out of 500 Nevadans, 17 people reported being of two or more races. Conduct a hypothesis test to determine if the population percents are the same for the two states or if the percent for Nevada is statistically higher than for North Dakota.

51. Is this a test of means or proportions?

52. State the null and alternative hypotheses.
   a. H_0: _____
   b. H_a: _____

53. Is this a right-tailed, left-tailed, or two-tailed test? How do you know?

54. What is the random variable of interest for this test?

55. In words, define the random variable for this test.

56. Which distribution (normal or Student's *t*) would you use for this hypothesis test?

57. Explain why you chose the distribution you did for the Exercise 10.56.

58. Calculate the test statistic.

59. Sketch a graph of the situation. Mark the hypothesized difference and the sample difference. Shade the area corresponding to the *p*-value.

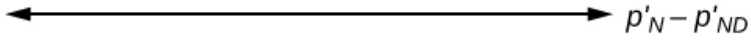


Figure 10.17

60. Find the *p*-value.

61. At a preconceived $\alpha = 0.05$, write the following:
a. Your decision:
b. The reason for your decision:
c. Your conclusion (write out in a complete sentence):

62. Does it appear that the proportion of Nevadans who are two or more races is higher than the proportion of North Dakotans? Why or why not?

10.4 Matched or Paired Samples (Optional)

Use the following information to answer the next five exercises. A study was conducted to test the effectiveness of a software patch in reducing system failures over a six-month period. Results for randomly selected installations are shown in Table 10.21. The *before* value is matched to an *after* value, and the differences are calculated. The differences have a normal distribution. Test at the 1 percent significance level.

Installation	A	B	C	D	E	F	G	H
Before	3	6	4	2	5	8	2	6
After	1	5	2	0	1	0	2	2

Table 10.21

63. What is the random variable?

64. State the null and alternative hypotheses.

65. What is the *p*-value?

66. Draw the graph of the *p*-value.

67. What conclusion can you draw about the software patch?

Use the following information to answer next five exercises. A study was conducted to test the effectiveness of a juggling class. Before the class started, six subjects juggled as many balls as they could at once. After the class, the same six subjects juggled as many balls as they could. The differences in the number of balls are calculated. The differences have a normal distribution. Test at the 1 percent significance level.

Subject	A	B	C	D	E	F
Before	3	4	3	2	4	5
After	4	5	6	4	5	7

Table 10.22

68. State the null and alternative hypotheses.

69. What is the *p*-value?

70. What is the sample mean difference?

71. Draw the graph of the *p*-value.

72. What conclusion can you draw about the juggling class?

Use the following information to answer the next five exercises. A doctor wants to know if a blood pressure medication is effective. Six subjects have their blood pressures recorded. After twelve weeks on the medication, the same six subjects have their blood pressure recorded again. For this test, only systolic pressure is of concern. Test at the 1 percent significance level.

Patient	A	B	C	D	E	F
Before	161	162	165	162	166	171
After	158	159	166	160	167	169

Table 10.23

73. State the null and alternative hypotheses.

74. What is the test statistic?

75. What is the *p*-value?

76. What is the sample mean difference?

77. What is the conclusion?

Homework

10.1 Two Population Means with Unknown Standard Deviations

DIRECTIONS: For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in Appendix E. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

NOTE

If you are using a Student's *t*-distribution for a homework problem in what follows, including for paired data, you may assume that the underlying population is normally distributed. (When using these tests in a real situation, you must first prove that assumption.)

78. The mean number of English courses taken in a two-year period by male and female college students is believed to be about the same. An experiment is conducted and data are collected from 29 males and 16 females. The males took an average of 3 English courses with a standard deviation of 0.8. The females took an average of 4 English courses with a standard deviation of 1.0. Are the means statistically the same?

79. A student at a four-year college claims that mean enrollment at four-year colleges is higher than at two-year colleges in the United States. Two surveys are conducted. Of the 35 two-year colleges surveyed, the mean enrollment was 5,068 with a standard deviation of 4,777. Of the 35 four-year colleges surveyed, the mean enrollment was 5,466 with a standard deviation of 8,191.

80. At Rachel's eleventh birthday party, eight girls were timed to see how long (in seconds) they could sit perfectly still in a relaxed position. After a two-minute rest, they timed themselves while jumping. The girls thought that the mean difference between their jumping and relaxed times would be zero. Test their hypothesis.

Relaxed time (seconds)	Jumping time (seconds)
26	21
47	40
30	28
22	21
23	25
45	43
37	35
29	32

Table 10.24

81. Mean entry-level salaries for college graduates with mechanical engineering degrees and electrical engineering degrees are believed to be approximately the same. A recruiting office thinks that the mean mechanical engineering salary is lower than the mean electrical engineering salary. The recruiting office randomly surveys 50 entry-level mechanical engineers and 60 entry-level electrical engineers. Their mean salaries were $46,100 and $46,700, respectively. Their standard deviations were $3,450 and $4,210, respectively. Conduct a hypothesis test to determine if you agree that the mean entry-level mechanical engineering salary is lower than the mean entry-level electrical engineering salary.

82. Marketing companies have collected data implying that teenage girls use more ringtones on their smartphones than teenage boys do. In one study of 40 randomly chosen teenage girls and boys (20 of each) with smartphones, the mean number of ringtones for the girls was 3.2 with a standard deviation of 1.5. The mean for the boys was 1.7 with a standard deviation of 0.8. Conduct a hypothesis test to determine if the means are approximately the same or if the girls' mean is higher than the boys' mean.

Use the information from Appendix C to answer the next four exercises.

83. Using the data from Lap 1 only, conduct a hypothesis test to determine if the mean time for completing a lap in races is the same as it is in practices.

84. Repeat the test in Exercise 10.83, but use Lap 5 data this time.

85. Repeat the test in Exercise 10.83, but this time combine the data from Laps 1 and 5.

86. In two to three complete sentences, explain in detail how you might use Terri Vogel's data to answer the following question: Does Terri Vogel drive faster in races than she does in practices?

Use the following information to answer the next two exercises. The Eastern and Western Major League Soccer conferences have a new Reserve Division that allows new players to develop their skills. Data for a randomly picked date showed the following annual goals.

Western	Eastern
Los Angeles 9	D United 9
FC Dallas 3	Chicago 8

Table 10.25

Western	Eastern
Chivas USA 4	Columbus 7
Real Salt Lake 3	New England 6
Colorado 4	MetroStars 5
San Jose 4	Kansas City 3

Table 10.25

Conduct a hypothesis test to answer the next two exercises.

87. The **exact** distribution for the hypothesis test is
   a. the normal distribution
   b. the Student's *t*-distribution
   c. the uniform distribution
   d. the exponential distribution

88. If the level of significance is 0.05, the conclusion is:
   a. There is sufficient evidence to conclude that the **W** Division teams score fewer goals, on average, than the **E** teams.
   b. There is insufficient evidence to conclude that the **W** Division teams score more goals, on average, than the **E** teams.
   c. There is insufficient evidence to conclude that the **W** teams score fewer goals, on average, than the **E** teams.
   d. There is not sufficient evidence to determine a conclusion.

89. Suppose a statistics instructor believes that there is no significant difference between the mean class scores of statistics day students on Exam 2 and statistics night students on Exam 2. She takes random samples from each of the populations. The mean and standard deviation for 35 statistics day students were 75.86 and 16.91. The mean and standard deviation for 37 statistics night students were 75.41 and 19.73. The *day* subscript refers to the statistics day students. The *night* subscript refers to the statistics night students. Which of the following is a concluding statement:
   a. There is sufficient evidence to conclude that statistics night students' mean on Exam 2 is better than the statistics day students' mean on Exam 2.
   b. There is insufficient evidence to conclude that the statistics day students' mean on Exam 2 is better than the statistics night students' mean on Exam 2.
   c. There is insufficient evidence to conclude that there is a significant difference between the means of the statistics day students and night students on Exam 2.
   d. There is sufficient evidence to conclude that there is a significant difference between the means of the statistics day students and night students on Exam 2.

90. Elijah wants to know whether textbook costs are different for different courses of study. He selects a random sample of 33 sociology textbooks offered on a popular online site. The mean price of his sample is $74.64 with a standard deviation of $49.36. He then selects a random sample of 33 math and science textbooks from the same site. The mean price of this sample is $111.56 with a standard deviation of $66.90. Is the mean price of a sociology textbook lower than the mean price of a math or science textbook? Test at a 1% significance level.

91. A powder diet is tested on 49 people, and a liquid diet is tested on 36 different people. Of interest is whether the liquid diet yields a higher mean weight loss than the powder diet. The powder diet group had a mean weight loss of 42 pounds with a standard deviation of 12 pounds. The liquid diet group had a mean weight loss of 45 pounds with a standard deviation of 14 pounds.

92. Suppose a statistics instructor believes that there is no significant difference between the mean class scores of statistics day students on Exam 2 and statistics night students on Exam 2. She takes random samples from each of the populations. The mean and standard deviation for 35 statistics day students were 75.86 and 16.91, respectively. The mean and standard deviation for 37 statistics night students were 75.41 and 19.73. The *day* subscript refers to the statistics day students. The *night* subscript refers to the statistics night students. An

appropriate alternative hypothesis for the hypothesis test is

a. $\mu_{day} > \mu_{night}$
b. $\mu_{day} < \mu_{night}$
c. $\mu_{day} = \mu_{night}$
d. $\mu_{day} \neq \mu_{night}$

10.2 Two Population Means with Known Standard Deviations

DIRECTIONS: For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in Appendix E. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

Note

If you are using a Student's *t*-distribution for one of the following homework problems, including for paired data, you may assume that the underlying population is normally distributed. (When using these tests in a real situation, you must first prove that assumption.)

93. A study is done to determine if students in the California state university system take longer to graduate, on average, than students enrolled in private universities. One hundred students from both the California state university system and private universities are surveyed. Suppose that from years of research, it is known that the population standard deviations are 1.5811 years and 1 year, respectively. The following data are collected. The California state university system students took on average 4.5 years with a standard deviation of 0.8. The private university students took on average 4.1 years with a standard deviation of 0.3.

94. Parents of teenage boys often complain that auto insurance costs more, on average, for teenage boys than for teenage girls. A group of concerned parents examines a random sample of insurance bills. The mean annual cost for 36 teenage boys was \$679. For 23 teenage girls, it was \$559. From past years, it is known that the population standard deviation for each group is \$180. Determine whether you believe that the mean cost for auto insurance for teenage boys is greater than that for teenage girls.

95. A group of transfer-bound students wondered if they will spend the same mean amount on texts and supplies each year at their four-year university as they have at their community college. They conducted a random survey of 54 students at their community college and 66 students at their local four-year university. The sample means were \$947 and \$1,011, respectively. The population standard deviations are known to be \$254 and \$87, respectively. Conduct a hypothesis test to determine if the means are statistically the same.

96. Some manufacturers claim that nonhybrid sedan cars have a lower mean miles per gallon (mpg) than hybrid ones. Suppose that consumers test 21 hybrid sedans and get a mean of 31 mpg with a standard deviation of 7 mpg. Thirty-one nonhybrid sedans get a mean of 22 mpg with a standard deviation of 4 mpg. Suppose that the population standard deviations are known to be 6 and 3, respectively. Conduct a hypothesis test to evaluate the manufacturers' claim.

97. A baseball fan wanted to know if there is a difference between the number of games played in a World Series when the American League won the series versus when the National League won the series. From 1922 to 2012, the population standard deviation of games won by the American League was 1.14, and the population standard deviation of games won by the National League was 1.11. Of 19 randomly selected World Series games won by the American League, the mean number of games won was 5.76. The mean number of 17 randomly selected games won by the National League was 5.42. Conduct a hypothesis test.

98. One of the questions in a study of marital satisfaction of dual-career couples was to rate the statement "I'm pleased with the way we divide the responsibilities for childcare." The ratings went from 1 (strongly agree) to 5 (strongly disagree). Table 10.26 contains 10 of the paired responses for husbands and wives. Conduct a hypothesis test to see if the mean difference in the husband's versus the wife's satisfaction level is negative (meaning that, within the partnership, the husband is happier than the wife).

Wife's Score	2	2	3	3	4	2	1	1	2	4
Husband's Score	2	2	1	3	2	1	1	1	2	4

Table 10.26

10.3 Comparing Two Independent Population Proportions

DIRECTIONS: For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in Appendix E. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

Note

If you are using a Student's *t*-distribution for one of the following homework problems, including for paired data, you may assume that the underlying population is normally distributed. (In general, you must first prove that assumption.)

99. A recent drug survey showed an increase in the use of prescription medication among local senior citizens as compared to the national percent. Suppose that a survey of 100 local seniors and 100 national seniors is conducted to see if the proportion of prescription medication use is higher locally or nationally. Locally, 65 senior citizens reported taking prescription medication within the past month, while 60 national seniors reported using them.

100. Elizabeth Mjelde, an art history professor, was interested in whether the value from the Golden Ratio formula, $\left(\frac{\text{larger + smaller dimension}}{\text{larger dimension}}\right)$, was the same in the Whitney Exhibit for works from 1900 to 1919 as for works from 1920 to 1942. Thirty-seven early works were sampled, averaging 1.74 with a standard deviation of 0.11. Sixty-five of the later works were sampled, averaging 1.746 with a standard deviation of 0.1064. Do you think that there is a significant difference in the Golden Ratio calculation?

101. A year was randomly picked from 1985 to the present. In that year, there were 2,051 Hispanic students at Cabrillo College out of a total of 12,328 students. At Lake Tahoe College, there were 321 Hispanic students out of a total of 2,441 students. In general, do you think that the percent of Hispanic students at the two colleges is basically the same or different?

Use the following information to answer the next three exercises. Neuroinvasive West Nile virus is a severe disease that affects a person's nervous system. It is spread by the *Culex* species of mosquito. In the United States in 2010, there were 629 reported cases of neuroinvasive West Nile virus out of a total of 1,021 reported cases, and there were 486 neuroinvasive reported cases out of a total of 712 cases reported in 2011. Is the 2011 proportion of neuroinvasive West Nile virus cases more than the 2010 proportion of neuroinvasive West Nile virus cases? Using a 1 percent level of significance, conduct an appropriate hypothesis test.

- *2011* subscript: 2011 group.
- *2010* subscript: 2010 group

102. This is
  a. a test of two proportions
  b. a test of two independent means
  c. a test of a single mean
  d. a test of matched pairs.

103. An appropriate null hypothesis is
  a. $p_{2011} \leq p_{2010}$
  b. $p_{2011} \geq p_{2010}$
  c. $\mu_{2011} \leq \mu_{2010}$
  d. $p_{2011} > p_{2010}$

104. The *p*-value is 0.0022. At a 1 percent level of significance, what is the appropriate conclusion?

a. There is sufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile virus is less than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile virus.
b. There is insufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile virus is more than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile virus.
c. There is insufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile virus is less than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile virus.
d. There is sufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile virus is more than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile virus.

105. Researchers conducted a study to find out if there is a difference in the use of e-readers by different age groups. Randomly selected participants were divided into two age groups. In the 16- to 29-year-old group, 7 percent of the 628 surveyed use e-readers, while 11 percent of the 2,309 participants 30 years old and older use e-readers.

106. Adults aged 18 years and older were randomly selected for a survey about a specific disease. The researchers wanted to determine if the proportion of women who have the disease is less than the proportion of southern men who do. The results are shown in Table 10.27. Test at the 1 percent level of significance.

	Number diagnosed with disease	Sample size
Men	42,769	155,525
Women	67,169	248,775

Table 10.27

107. Two computer users were discussing tablet computers. A higher proportion of people ages 16 to 29 use tablets than of people age 30 and older. Table 10.28 details the number of tablet owners for each age group. Test at the 1 percent level of significance.

	16–29 year olds	30 years and older
Own a Tablet	69	231
Sample Size	628	2,309

Table 10.28

108. A group of friends debated whether more men use smartphones than women. They consulted a research study of smartphone use among adults. The results of the survey indicate that of the 973 men randomly sampled, 379 use smartphones. For women, 404 of the 1,304 who were randomly sampled use smartphones. Test at the 5 percent level of significance.

109. While her husband spent 2.5 hours picking out new speakers, a statistician decided to determine whether the percent of men who enjoy shopping for electronic equipment is higher than the percent of women who do. The population was Saturday afternoon shoppers. Out of 67 men, 24 said they enjoyed the activity. Eight of the 24 women surveyed claimed to enjoy the activity. Interpret the results of the survey.

110. We are interested in whether children's educational computer software costs less, on average, than children's entertainment software. Thirty-six educational software titles were randomly picked from a catalog. The mean cost was $31.14 with a standard deviation of $4.69. Thirty-five entertainment software titles were randomly picked from the same catalog. The mean cost was $33.86 with a standard deviation of $10.87. Decide whether children's educational software costs less, on average, than children's entertainment software.

111. A researcher recently claimed that the proportion of college-age males who wear at least one piece of jewelery is as high as the proportion of college-age females. She conducted a survey in her classes. Out of 107 males, 20 wear at least one piece of jewelery. Out of 92 females, 47 wear at least one piece of jewelery. Do you believe that the proportion of males has reached the proportion of females?

112. Use the data sets found in Appendix C to answer this exercise. Is the proportion of race laps Terri completes slower than 130 seconds less than the proportion of practice laps she completes slower than 135 seconds?

113. *To Breakfast or Not to Breakfast?* by Richard Ayore
In the American society, birthdays are one of those days that everyone looks forward to. People of different ages and peer groups gather to mark the 18th, 20th, ..., birthdays. During this time, one looks back to see what he or she has achieved for the past year and also focuses ahead for more to come.
If, by any chance, I am invited to one of these parties, my experience is always different. Instead of dancing around with my friends while the music is booming, I get carried away by memories of my family back home in Kenya. I remember the good times I had with my brothers and sister while we did our daily routine.
Every morning, I remember we went to the shamba (garden) to weed our crops. I remember one day arguing with my brother as to why he always remained behind just to join us an hour later. In his defense, he said that he preferred waiting for breakfast before he came to weed. He said, "This is why I always work more hours than you guys!"
And so, to prove him wrong or right, we decided to give it a try. One day we went to work as usual without breakfast, and recorded the time we could work before getting tired and stopping. On the next day, we all ate breakfast before going to work. We recorded how long we worked again before getting tired and stopping. Of interest was our mean increase in work time. Though not sure, my brother insisted that it was more than two hours. Using the data in Table 10.29, solve our problem.

Work hours with breakfast	Work hours without breakfast
8	6
7	5
9	5
5	4
9	7
8	7
10	7
7	5
6	6
9	5

Table 10.29

10.4 Matched or Paired Samples (Optional)

DIRECTIONS: For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in Appendix E. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

Note

If you are using a Student's *t*-distribution for the homework problems, including for paired data, you may assume that the underlying population is normally distributed. (When using these tests in a real situation, you must first prove that assumption.)

114. Ten individuals went on a low-fat diet for 12 weeks to lower their cholesterol. The data are recorded in Table 10.30. Do you think that their cholesterol levels were significantly lowered?

Starting cholesterol level	Ending cholesterol level
140	140
220	230
110	120
240	220
200	190
180	150
190	200
360	300
280	300
260	240

Table 10.30

Use the following information to answer the next two exercises. A new preventative medication was tried on a group of 224 patients who had the same risk factors for a disease. 45 patients developed the disease after four years. In a control group of 224 patients, 68 developed the disease after four years. We want to test whether the method of treatment reduces the proportion of patients who develop the disease after four years.

Let the subscript t = treated patient and ut = untreated patient.

115. The appropriate hypotheses are
a. H_0: $p_t < p_{ut}$ and H_a: $p_t \geq p_{ut}$
b. H_0: $p_t \leq p_{ut}$ and H_a: $p_t > p_{ut}$
c. H_0: $p_t = p_{ut}$ and H_a: $p_t \neq p_{ut}$
d. H_0: $p_t = p_{ut}$ and H_a: $p_t < p_{ut}$

116. If the p-value is 0.0062, what is the conclusion? Use $\alpha = 0.05$.
a. The method has no effect.
b. There is sufficient evidence to conclude that the method reduces the proportion of patients who develop the disease after four years.

c. There is sufficient evidence to conclude that the method increases the proportion of patients who develop the disease after four years.
d. There is insufficient evidence to conclude that the method reduces the proportion of patients who develop the disease after four years.

Use the following information to answer the next two exercises. An experiment is conducted to show that blood pressure can be consciously reduced in people trained in a biofeedback exercise program. Six subjects were randomly selected, and blood pressure measurements were recorded before and after the training. The difference between blood pressures was calculated (after – before), producing the following results: $\overline{x}_d = -10.2$ $s_d = 8.4$. Using the data, test the hypothesis that the blood pressure has decreased after the training.

117. The distribution for the test is
a. t_5
b. t_6
c. $N(-10.2, 8.4)$
d. $N(-10.2, \frac{8.4}{\sqrt{6}})$

118. If $\alpha = 0.05$, the p-value and the conclusion are
a. 0.0014; There is sufficient evidence to conclude that the blood pressure decreased after the training.
b. 0.0014; There is sufficient evidence to conclude that the blood pressure increased after the training.
c. 0.0155; There is sufficient evidence to conclude that the blood pressure decreased after the training.
d. 0.0155; There is sufficient evidence to conclude that the blood pressure increased after the training.

119. A golf instructor is interested in determining if her new technique for improving players' golf scores is effective. She takes four new students. She records their 18-hole scores before learning the technique and then after having taken her class. She conducts a hypothesis test. The data are as follows.

	Player 1	Player 2	Player 3	Player 4
Mean score before class	83	78	93	87
Mean score after class	80	80	86	86

Table 10.31

The correct decision is
a. reject H_0.
b. do not reject H_0.

120. A local research group is studying a chronic disease. They believe the number of cases of the disease is higher in 2013 than in 2012 in the southern United States. The group compared the estimates of new cases by southern state in 2012 and 2013. The results are in Table 10.32.

Southern States	2012	2013
Alabama	3,450	3,720
Arkansas	2,150	2,280
Florida	15,540	15,710
Georgia	6,970	7,310
Kentucky	3,160	3,300

Table 10.32

Southern States	2012	2013
Louisiana	3,320	3,630
Mississippi	1,990	2,080
North Carolina	7,090	7,430
Oklahoma	2,630	2,690
South Carolina	3,570	3,580
Tennessee	4,680	5,070
Texas	15,050	14,980
Virginia	6,190	6,280

Table 10.32

121. A traveler wanted to know if the prices of hotels are different in the 10 cities that he visits the most often. The list of the cities with the corresponding prices for his two favorite hotel chains is in Table 10.33. Test at the 1 percent level of significance.

Cities	Hyatt Regency prices in dollars	Hilton prices in dollars
Atlanta	107	169
Boston	358	289
Chicago	209	299
Dallas	209	198
Denver	167	169
Indianapolis	179	214
Los Angeles	179	169
New York City	625	459
Philadelphia	179	159
Washington, DC	245	239

Table 10.33

122. A politician asked his staff to determine whether the underemployment rate in the Northeast decreased from 2011 to 2012. The results are in Table 10.34.

Northeastern States	2011	2012
Connecticut	17.3	16.4
Delaware	17.4	13.7
Maine	19.3	16.1
Maryland	16.0	15.5
Massachusetts	17.6	18.2
New Hampshire	15.4	13.5
New Jersey	19.2	18.7
New York	18.5	18.7
Ohio	18.2	18.8
Pennsylvania	16.5	16.9
Rhode Island	20.7	22.4
Vermont	14.7	12.3
West Virginia	15.5	17.3

Table 10.34

Bringing It Together: Homework

Use the following information to answer the next 10 exercises. Indicate which of the following choices best identifies the hypothesis test.

A. Independent group means, population standard deviations and/or variances known
B. Independent group means, population standard deviations and/or variances unknown
C. Matched or paired samples
D. Single mean
E. Two proportions
F. Single proportion

123. A powder diet is tested on 49 people, and a liquid diet is tested on 36 different people. The population standard deviations are two pounds and three pounds, respectively. Of interest is whether the liquid diet yields a higher mean weight loss than the powder diet.

124. A new chocolate bar is taste-tested on consumers. Of interest is whether the proportion of children who like the new chocolate bar is greater than the proportion of adults who like it.

125. The mean number of English courses taken in a two-year time period by male and female college students is believed to be about the same. An experiment is conducted and data are collected from 9 males and 16 females.

126. A football league reported that the mean number of touchdowns per game was five. A study is done to determine if the mean number of touchdowns has decreased.

127. A study is done to determine if students in the California state university system take longer to graduate than students enrolled in private universities. One hundred students from both the California state university system

and private universities are surveyed. From years of research, it is known that the population standard deviations are 1.5811 years and 1 year, respectively.

128. According to a doctor's magazine, 75 percent of senior citizens think that yearly checkups are very important. A study is done to verify this.

129. According to a recent study, U.S. companies have a mean maternity leave of six weeks.

130. A recent survey showed an increase in use of prescription medication among local senior citizens as compared to the national percent. Suppose that a survey of 100 local senior citizens and 100 national senior citizens is conducted to see if the proportion of prescription medication use is higher locally than nationally.

131. A new SAT study course is tested on 12 individuals. Pre-course and post-course scores are recorded. Of interest is the mean increase in SAT scores. The following data are collected:

Pre-course score	Post-course score
1	300
960	920
1010	1100
840	880
1100	1070
1250	1320
860	860
1330	1370
790	770
990	1040
1110	1200
740	850

Table 10.35

132. According to a statistics college professor, 68 percent of his students pass the final exam. A graduate researcher designs a study to determine if this claim is true.

133. Lesley E. Tan investigated the relationship between left-handedness versus right-handedness and motor competence in preschool children. Random samples of 41 left-handed preschool children and 41 right-handed preschool children were given several tests of motor skills to determine if there is evidence of a difference between the children based on this experiment. The experiment produced the means and standard deviations shown in Table 10.36. Determine the appropriate test and best distribution to use for that test.

	Left-handed	Right-handed
Sample size	41	41

Table 10.36

Sample mean	97.5	98.1
Sample standard deviation	17.5	19.2

Table 10.36

a. Two independent means, normal distribution
b. Two independent means, Student's *t*-distribution
c. Matched or paired samples, Student's *t*-distribution
d. Two population proportions, normal distribution

134. A golf instructor is interested in determining if her new technique for improving players' golf scores is effective. She takes four new students. She records their 18-hole scores before learning the technique and after having taken her class. She conducts a hypothesis test. The data are shown in Table 10.37.

	Player 1	Player 2	Player 3	Player 4
Mean score before class	83	78	93	87
Mean score after class	80	80	86	86

Table 10.37

This is
a. a test of two independent means.
b. a test of two proportions.
c. a test of a single mean.
d. a test of a single proportion.

References

10.1 Two Population Means with Unknown Standard Deviations

Baseball-Almanac. (2013). *World series history*. Retrieved from http://www.baseball-almanac.com/ws/wsmenu.shtml

Graduating Engineer + Computer Careers. (n.d.). Retrieved from http://www.graduatingengineer.com

Microsoft Bookshelf. (n.d.).

Nasdaq. (n.d.). *Sectoring by industry groups*. Retrieved from http://www.nasdaq.com/markets/barchart-sectors.aspx

U.S. Senate. (n.d.). Retrieved from www.senate.gov

Wikipedia. (n.d.). *List of current United States Senators by age*. Retrieved from http://en.wikipedia.org/wiki/List_of_current_United_States_Senators_by_age

10.2 Two Population Means with Known Standard Deviations

Centers for Disease Control and Prevention. (2008, July 18). State-specific prevalence of obesity among adults—United States, 2007. *MMWR*, 57(28), 765–768. Retrieved from http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5728a1.htm

Federal Bureau of Investigation. (n.d.). *Texas Crime Rates 1960–1012*. Available at http://www.disastercenter.com/crime/txcrime.htm

Hinduja, S. (2013). Sexting Research and Gender Differences. *Cyberbulling Research Center*. Retrieved from http://cyberbullying.us/blog/sexting-research-and-gender-differences/

Humes, K. R., Jones, N. A., & Ramirez, R. R. (2011 March). *Overview of race and Hispanic origin: 2010* (2010 Census Briefs). Washington, DC: U.S. Census Bureau. Available online at http://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf

Smith, A. (2011, July 11). 35% of American adults own a smartphone. Pew Internet. Available online at http://www.pewinternet.org/~/media/Files/Reports/2011/PIP_Smartphones.pdf

Visually. (2013). *Smart phone users, by the numbers*. Retrieved from http://visual.ly/smart-phone-users-numbers

10.3 Comparing Two Independent Population Proportions

American Cancer Society. (n.d.). Retrieved from http://www.cancer.org/index

Centers for Disease Control and Prevention. (n.d.). *West Nile virus*. Retrieved from http://www.cdc.gov/ncidod/dvbid/westnile/index.htm

Chancellor's Office, California Community Colleges. (1994, Nov.).

Educational Resources. (n.d.).

Gallup. (2013). *State of the states*. Retrieved from http://www.gallup.com/poll/125066/State-States.aspx?ref=interactive

Hilton Hotels. (n.d.). Retrieved from http://www.hilton.com

Hyatt Hotels. Retrieved from http://hyatt.com

San Jose Museum of Art. (n.d.). Whitney exhibit (on loan).

U.S. Department of Health and Human Services. (n.d). *Statistics*. Retrieved from https://www.hhs.gov/

Solutions

1. two proportions

3. matched or paired samples

5. single mean

7. independent group means, population standard deviations and/or variances unknown

9. two proportions

11. independent group means, population standard deviations and/or variances unknown

13. independent group means, population standard deviations and/or variances unknown

15. two proportions

17. The random variable is the difference between the mean amounts of sugar in the two soft drinks.

19. means

21. two-tailed

23. the difference between the mean life spans of whites and nonwhites

25. This is a comparison of two population means with unknown population standard deviations.

27. Check student's solution.

29.
a. Reject the null hypothesis.
b. p-value < 0.05
c. There is not enough evidence at the 5 percent level of significance to support the claim that life expectancy in the 1900s is different between whites and nonwhites.

31. the difference in mean speeds of the fastball pitches of the two pitchers

33. –2.46

35. At the 1 percent significance level, we can reject the null hypothesis. There is sufficient data to conclude that the mean speed of Rodriguez's fastball is faster than Wesley's.

37. Subscripts: 1 = Food, 2 = No Food
H_0: $\mu_1 \leq \mu_2$
H_a: $\mu_1 > \mu_2$

39.

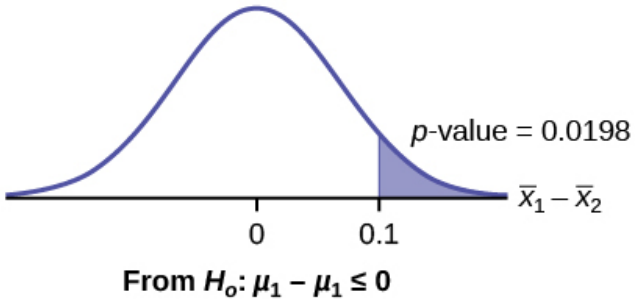


Figure 10.18

41. Subscripts: 1 = Gamma, 2 = Zeta
H_0: $\mu_1 = \mu_2$
H_a: $\mu_1 \neq \mu_2$

43. 0.0062

45. There is sufficient evidence to reject the null hypothesis. The data support that the melting point for Alloy Zeta is different from the melting point of Alloy Gamma.

47. $P'_{OS1} - P'_{OS2}$ = difference in the proportions of phones that had system failures within the first eight hours of operation with OS_1 and OS_2.

49. 0.1018

51. proportions

53. right-tailed

55. The random variable is the difference in proportions (percents) of the populations that are of two or more races in Nevada and North Dakota.

57. Our sample sizes are much greater than five each, so we use the normal for two proportions distribution for this hypothesis test.

59. Check student's solution.

61.
a. Reject the null hypothesis.
b. *p*-value < alpha
c. At the 5 percent significance level, there is sufficient evidence to conclude that the proportion (percent) of the population that is of two or more races in Nevada is statistically higher than that in North Dakota.

63. the mean difference of the system failures

65. 0.0067

67. With a *p*-value 0.0067, we can reject the null hypothesis. There is enough evidence to support that the software patch is effective in reducing the number of system failures.

69. 0.0021

71.

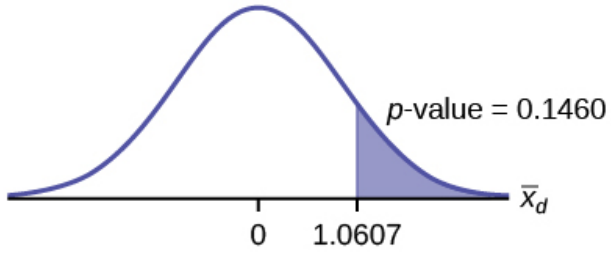


Figure 10.19

73. H_0: $\mu_d \geq 0$
H_a: $\mu_d < 0$

75. 0.0699

77. We decline to reject the null hypothesis. There is not sufficient evidence to support that the medication is effective.

79. Subscripts: 1: two-year colleges, 2: four-year colleges
a. H_0: $\mu_1 \geq \mu_2$
b. H_a: $\mu_1 < \mu_2$
c. $\overline{X}_1 - \overline{X}_2$ is the difference between the mean enrollments of the two-year colleges and the four-year colleges.
d. Student's t
e. test statistic: -0.2480
f. p-value: 0.4019
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Do not reject.
iii. Reason for Decision: p-value > alpha
iv. Conclusion: At the 5 percent significance level, there is sufficient evidence to conclude that the mean enrollment at four-year colleges is higher than at two-year colleges.

81. Subscripts: 1: mechanical engineering, 2: electrical engineering
a. H_0: $\mu_1 \geq \mu_2$
b. H_a: $\mu_1 < \mu_2$
c. $\overline{X}_1 - \overline{X}_2$ is the difference between the mean entry-level salaries of mechanical engineers and electrical engineers.
d. t_{108}
e. test statistic: $t = -0.82$
f. p-value: 0.2061
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Do not reject the null hypothesis.
iii. Reason for Decision: p-value > alpha
iv. Conclusion: At the 5 percent significance level, there is insufficient evidence to conclude that the mean entry-level salaries of mechanical engineers is lower than that of electrical engineers.

83. a. H_0: $\mu_1 = \mu_2$
b. H_a: $\mu_1 \neq \mu_2$
c. $\overline{X}_1 - \overline{X}_2$ is the difference between the mean times for completing a lap in races and in practices.
d. $t_{20.32}$
e. test statistic: –4.70
f. p-value: 0.0001
g. Check student's solution.
h. i. Alpha: 0.05

ii. Decision: Reject the null hypothesis.
iii. Reason for Decision: *p*-value < alpha
iv. Conclusion: At the 5 percent significance level, there is sufficient evidence to conclude that the mean time for completing a lap in races is different from that in practices.

85. a. H_0: $\mu_1 = \mu_2$
b. H_a: $\mu_1 \neq \mu_2$
c. is the difference between the mean times for completing a lap in races and in practices.
d. $t_{40.94}$
e. test statistic: –5.08
f. *p*-value: zero
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Reject the null hypothesis.
iii. Reason for Decision: *p*-value < alpha
iv. Conclusion: At the 5 percent significance level, there is sufficient evidence to conclude that the mean time for completing a lap in races is different from that in practices.

88. c

90. Test: two independent sample means, population standard deviations unknown.

- μ_1 = the mean price of a sociology text on the selected site
- μ_2 = the mean price of a math/science text on the selected site

Random variable: $\overline{X}_1 - \overline{X}_2$ = the difference in the sample mean textbook price between sociology texts and math/science texts.
Hypotheses: $H_0: \mu_1 - \mu_2 = 0$, $H_a: \mu_1 - \mu_2 < 0$, which can be expressed as $H_0: \mu_1 = \mu_2$, $H_a: \mu_1 < \mu_2$.
Distribution for the test: Use t_{df}, because each sample has more than 30 observations,
$df = n_1 + n_2 - 2 = 33 + 33 - 2 = 64$.
Estimate the critical value on the *t*-table using the nearest available degrees of freedom, 60. The critical value, 2.660, is found in the .0005 column.
Calculate the test statistic: .
Using a calculator with t_c = –2.55 and *df* = 64, the left-tailed *p*-value: 0.0066. Decision: Reject H_0. Conclusion: At the 1% level of significance, from the sample data, there is sufficient evidence to conclude that the mean price of sociology textbooks is less than the mean price of textbooks for math/science.
Distribution: H_0: $\mu_1 = \mu_2$, H_a: $\mu_1 < \mu_2$
The mean age of entering the industry in Canada is lower than the mean age in the United States.

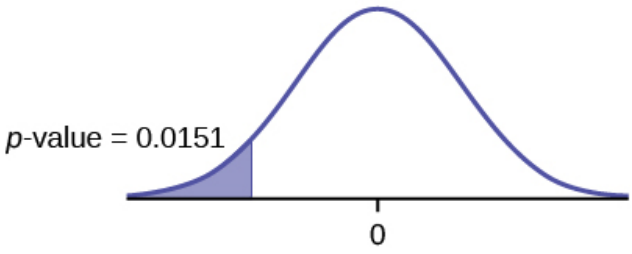


Figure 10.20

Graph: left-tailed
p-value : 0.0151
Decision: Do not reject H_0.
Conclusion: At the 1 percent level of significance, from the sample data, there is not sufficient evidence to conclude that the mean age of entering the industry in Canada is lower than the mean age in the United States.

92. d

94. Subscripts: 1 = boys, 2 = girls

a. H_0: $\mu_1 \leq \mu_2$
b. H_a: $\mu_1 > \mu_2$
c. The random variable is the difference in the mean auto insurance costs for boys and girls.
d. normal
e. test statistic: $z = 2.50$
f. p value: 0.0062
g. Check student's solution.
h. i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for Decision: p value < alpha
   iv. Conclusion: At the 5 percent significance level, there is sufficient evidence to conclude that the mean cost of auto insurance for teenage boys is greater than that for girls.

96. Subscripts: 1 = non-hybrid sedans, 2 = hybrid sedans
a. H_0: $\mu_1 \geq \mu_2$
b. H_a: $\mu_1 < \mu_2$
c. The random variable is the difference in the mean miles per gallon of nonhybrid sedans and hybrid sedans.
d. normal
e. test statistic: 6.36
f. p-value: 0
g. Check student's solution.
h. i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: p value < alpha
   iv. Conclusion: At the 5 percent significance level, there is sufficient evidence to conclude that the mean miles per gallon of non-hybrid sedans is less than that of hybrid sedans.

98. a. H_0: $\mu_d = 0$
b. H_a: $\mu_d < 0$
c. The random variable X_d is the average difference between husband's and wife's satisfaction level.
d. t_9
e. test statistic: $t = -1.86$
f. p value: 0.0479
g. Check student's solution
h. i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis, but run another test.
   iii. Reason for Decision: p value < alpha
   iv. Conclusion: This is a weak test because alpha and the p value are close. However, there is insufficient evidence to conclude that the mean difference is negative.

101. Subscripts: 1 = Cabrillo College, 2 = Lake Tahoe College
a. H_0: $p_1 = p_2$
b. H_a: $p_1 \neq p_2$
c. The random variable is the difference between the proportions of Hispanic students at Cabrillo College and Lake Tahoe College.
d. normal for two proportions
e. test statistic: 4.29
f. p-value: 0.00002
g. Check student's solution.
h. i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: p-value < alpha
   iv. Conclusion: There is sufficient evidence to conclude that the proportions of Hispanic students at Cabrillo College and Lake Tahoe College are different.

103. a

105. Test: two independent sample proportions.
Random variable: p'_1 - p'_2
Distribution:
H_0: $p_1 = p_2$
H_a: $p_1 \neq p_2$
The proportion of e-reader users is different for the 16- to 29-year-old users from that of the 30 and older users.
Graph: two-tailed

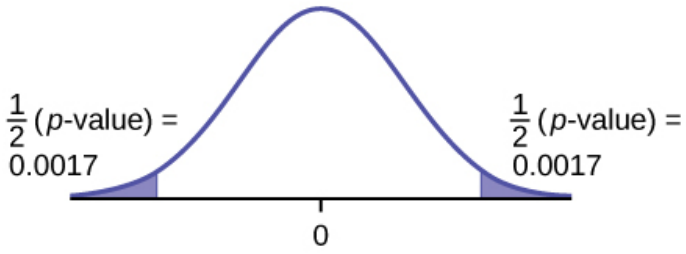


Figure 10.21

p-value : 0.0033
Decision: Reject the null hypothesis.
Conclusion: At the 5 percent level of significance, from the sample data, there is sufficient evidence to conclude that the proportion of e-reader users 16 to 29 years old is different from the proportion of e-reader users 30 and older.

107. Test: two independent sample means, population standard deviations unknown.
μ_1 = the the mean price of a sociology text on the selected site.
μ_2 = the mean price of a math/science text on the selected site.
Random variable: $\overline{X_1} - \overline{X_1}$ = the difference in the sample mean textbook price between sociology texts and math/science texts.
Hypotheses: $H_0 : \mu_1 - \mu_2 = 0$, $H_a : \mu_1 - \mu_2 < \mu_2$ which can be expressed as
H0s : μ1 − μ2, Ha μ1 < μ2.
Distribution for the test: Use t_{df}; because each sample has more than 30 observations,
$df = n_1 + n_2 - 2 = 33 + 33 - 2 = 64$.
Estimate the critical value on the *t*-table using the nearest available degrees of freedom, 60. The critical value, 2.660, is found in the .0005 column.
Calculate the test statistic: $t_c = \frac{(\overline{X}_1 - \overline{X}_2) - 0}{\sqrt{\frac{{s_1}^2}{n_2} + \frac{{s_2}^2}{n_2}}} = \frac{(74.64 - 111.56) - 0}{\sqrt{\frac{49.36^2}{33} + \frac{66.90^2}{33}}} = -2.55.$
Using a calculator with $t_c = -2.55$ and $df = 64$, the left-tailed *p*-value: Decision: Reject H_0. Conclusion: At the 1% level of significance, from the sample data, there is sufficient evidence to conclude that the mean price of sociology textbooks is less than the mean price of textbooks for math/science.

109. Subscripts: 1: men; 2: women

a. H_0: $p_1 \leq p_2$
b. H_a: $p_1 > p_2$
c. $P'_1 - P'_2$ is the difference between the proportions of men and women who enjoy shopping for electronic equipment.
d. normal for two proportions
e. test statistic: 0.22
f. *p*-value: 0.4133
g. Check student's solution.
h.
   i. Alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for Decision: *p*-value > alpha
   iv. Conclusion: At the 5 percent significance level, there is insufficient evidence to conclude that the proportion of men who enjoy shopping for electronic equipment is more than the proportion of women.

111. a. H_0: $p_1 = p_2$
b. H_a: $p_1 \neq p_2$
c. $P'_1 - P'_2$ is the difference between the proportions of men and women that have at least one pierced ear.
d. normal for two proportions
e. test statistic: –4.82
f. *p*-value: zero
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Reject the null hypothesis.
iii. Reason for Decision: *p*-value < alpha
iv. Conclusion: At the 5 percent significance level, there is sufficient evidence to conclude that the proportions of males and females with at least one pierced ear is different.

113. a. H_0: $\mu_d = 0$
b. H_a: $\mu_d > 0$
c. The random variable X_d is the mean difference in work times on days when eating breakfast and on days when not eating breakfast.
d. t_9
e. test statistic: 4.8963
f. *p*-value: 0.0004
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Reject the null hypothesis.
iii. Reason for Decision: *p*-value < alpha
iv. Conclusion: At the 5 percent level of significance, there is sufficient evidence to conclude that the mean difference in work times on days when eating breakfast and on days when not eating breakfast has increased.

114. *p*-value = 0.1494
At the 5 percent significance level, there is insufficient evidence to conclude that the medication lowered cholesterol levels after 12 weeks.

116. b

118. c

120. Test: two matched pairs or paired samples (*t*-test)
Random variable: $\overline{X}_d$
Distribution: t_{12}
H_0: $\mu_d = 0$ H_a: $\mu_d > 0$
The mean of the differences of new female breast cancer cases in the south between 2013 and 2012 is greater than zero. The estimate for new female breast cancer cases in the south is higher in 2013 than in 2012.
Graph: right-tailed
p-value: 0.0004

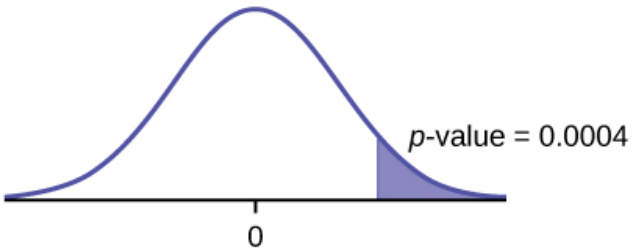


Figure 10.22

Decision: Reject H_0.
Conclusion: At the 5 percent level of significance, from the sample data, there is sufficient evidence to conclude

that there was a higher estimate of new female breast cancer cases in 2013 than in 2012.

122. Test: matched or paired samples (*t*-test)
Difference data: {–0.9, –3.7, –3.2, –0.5, 0.6, –1.9, –0.5, 0.2, 0.6, 0.4, 1.7, –2.4, 1.8}
Random Variable: $\overline{X}_d$
Distribution: H_0: $\mu_d = 0$ H_a: $\mu_d < 0$
The mean of the differences of the rate of underemployment in the northeastern states between 2012 and 2011 is less than zero. The underemployment rate went down from 2011 to 2012.
Graph: left-tailed.

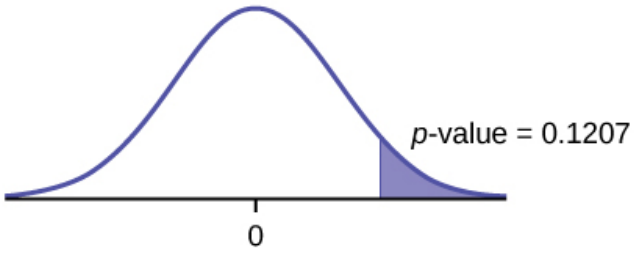


Figure 10.23

p-value: 0.1207
Decision: Do not reject H_0.
Conclusion: At the 5 percent level of significance, from the sample data, there is not sufficient evidence to conclude that there was a decrease in the underemployment rates of the northeastern states from 2011 to 2012.

124. e

126. d

128. f

130. e

132. f
The graduate researcher will be comparing a sample proportion to a population proportion or claim. Thus, the study includes the hypothesis test of a single proportion. A two proportion hypothesis test compares two sample proportions.

134. a

11 THE CHI-SQUARE DISTRIBUTION

Figure 11.1 The chi-square distribution can be used to find relationships between two things, like grocery prices at different stores. (credit: Pete/flickr)

Chapter Objectives

By the end of this chapter, the student should be able to do the following:

- Interpret the chi-square probability distribution as the sample size changes
- Conduct and interpret chi-square goodness-of-fit hypothesis tests
- Conduct and interpret chi-square test of independence hypothesis tests
- Conduct and interpret chi-square homogeneity hypothesis tests
- Conduct and interpret chi-square single variance hypothesis tests

Introduction

Have you ever wondered if lottery numbers were evenly distributed or if some numbers occurred with a greater frequency? How about if the types of movies people preferred were different across different age groups? What about if a coffee machine was dispensing approximately the same amount of coffee each time? You could answer these questions by conducting a hypothesis test.

You will now study a new distribution, one that is used to determine the answers to such questions. This distribution is called the chi-square distribution.

In this chapter, you will learn the three major applications of the chi-square distribution:

- The goodness-of-fit test, which determines if data fit a particular distribution, such as in the lottery example
- The test of independence, which determines if events are independent, such as in the movie example
- The test of a single variance, which tests variability, such as in the coffee example

NOTE
Though the chi-square distribution depends on calculators or computers for most of the calculations, there is a table available (see Appendix G Notes for the TI-83, 83+, 84, 84+ Calculators). TI-83+ and TI-84 calculator instructions are

included in the text.

COLLABORATIVE EXERCISE

Look in the sports section of a newspaper or on the internet for some sports data: baseball averages, basketball scores, golf tournament scores, football odds, swimming times, and the like. Plot a histogram and a boxplot using your data. See if you can determine a probability distribution that your data fits. Have a discussion with the class about your choice.

11.1 Facts About the Chi-Square Distribution

The notation for the **chi-square distribution** is

$$\chi \sim \chi^2_{df}$$

where df = degrees of freedom, which depends on how chi-square is being used. If you want to practice calculating chi-square probabilities then use $df = n - -1$. The degrees of freedom for the three major uses are calculated differently.

For the χ^2 distribution, the population mean is $\mu = df$, and the population standard deviation is $\sigma = \sqrt{2(df)}$.

The random variable is shown as χ^2, but it may be any uppercase letter.

The random variable for a chi-square distribution with k degrees of freedom is the sum of k independent, squared standard normal variables is

$\chi^2 = (Z_1)^2 + (Z_2)^2 + \ldots + (Z_k)^2$, where the following are true:

- The curve is nonsymmetrical and skewed to the right.
- There is a different chi-square curve for each df.

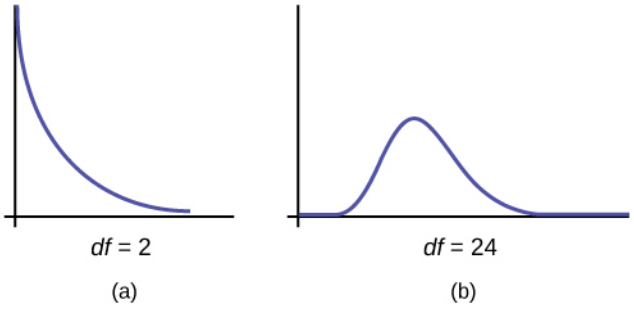


Figure 11.2

- The test statistic for any test is always greater than or equal to zero.
- When $df > 90$, the chi-square curve approximates the normal distribution. For $X \sim \chi^2_{1,000}$, the mean, $\mu = df = 1{,}000$ and the standard deviation, $\sigma = \sqrt{2(1{,}000)} = 44.7$. Therefore, $X \sim N(1{,}000, 44.7)$, approximately.
- The mean, μ, is located just to the right of the peak.

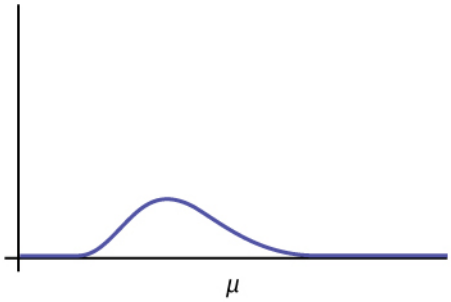


Figure 11.3

11.2 Goodness-of-Fit Test

In this type of hypothesis test, you determine whether the data fit a particular distribution. For example, you may suspect your unknown data fit a binomial distribution. You use a chi-square test, meaning the distribution for the hypothesis test is chi-square, to determine if there is a fit. The null and the alternative hypotheses for this test may be written in sentences or may be stated as equations or inequalities.

The test statistic for a goodness-of-fit test is:

$$\sum_k \frac{(O-E)^2}{E}$$

where

- O = **observed values** (data),
- E = **expected values** (from theory), and
- k = the number of different data cells or categories.

The observed values are the data values, and the expected values are the values you would expect to get if the null hypothesis were true. There are n terms of the form $\frac{(O-E)^2}{E}$.

The number of degrees of freedom is df = (number of categories – 1).

The goodness-of-fit test is almost always right-tailed. If the observed values and the corresponding expected values are not close to each other, then the test statistic can get very large and will be way out in the right tail of the chi-square curve.

Note

The expected value for each cell needs to be at least five for you to use this test.

EXAMPLE 11.1

Absenteeism of college students from math classes is a major concern to math instructors because missing class appears to increase the drop rate. Suppose that a study was done to determine if the actual student absenteeism rate follows faculty perception. The faculty expected that a group of 100 students would miss class according to Table 11.1.

Number of Absences per Term	Expected Number of Students
0–2	50
3–5	30
6–8	12
9–11	6
12+	2

Table 11.1

A random survey across all mathematics courses was then done to determine the number of observed absences in a course. Table 11.2 displays the results of that survey.

Number of Absences per Term	Actual Number of Students
0–2	35
3–5	40
6–8	20
9–11	1
12+	4

Table 11.2

Determine the null and alternative hypotheses needed to conduct a goodness-of-fit test.

H_0: Student absenteeism *fits* faculty perception.

The alternative hypothesis is the opposite of the null hypothesis.

H_a: Student absenteeism *does not fit* faculty perception.

Problem

a. Can you use the information as it appears in the charts to conduct the goodness-of-fit test?

Solution

a. **No.** Notice that the expected number of absences for the 12+ entry is less than five; it is two. Combine that group with the 9–11 group to create new tables where the number of students for each entry is at least five. The new results are in Table 11.3 and Table 11.4.

Number of Absences per Term	Expected Number of Students
0–2	50
3–5	30
6–8	12
9+	8

Table 11.3

Number of Absences per Term	Actual Number of Students
0–2	35
3–5	40
6–8	20
9+	5

Table 11.4

Problem

b. What is the number of degrees of freedom (*df*)?

Solution

b. There are four cells or categories in each of the new tables.

df = number of cells – 1 = 4 – 1 = 3.

TRY IT 11.1

A factory manager needs to understand how many products are defective versus how many are produced. The number of expected defects is listed in Table 11.5.

Number Produced	Number Defective
0–100	5
101–200	6
201–300	7
301–400	8
401–500	10

Table 11.5

A random sample was taken to determine the actual number of defects. Table 11.6 shows the results of the survey.

Number Produced	Number Defective
0–100	5
101–200	7
201–300	8
301–400	9
401–500	11

Table 11.6

State the null and alternative hypotheses needed to conduct a goodness-of-fit test, and state the degrees of freedom.

EXAMPLE 11.2

Problem

Employers want to know which days of the week employees are absent in a five-day work week. Most employers would like to believe that employees are absent equally during the week. Suppose a random sample of 60 managers were asked on which day of the week they had the highest number of employee absences. The results were distributed as in

Table 11.7. For the population of employees, do the days for the highest number of absences occur with equal frequencies during a five-day work week? Test at a 5 percent significance level.

	Monday	Tuesday	Wednesday	Thursday	Friday
Number of Absences	15	12	9	9	15

Table 11.7 Day of the Week Employees Were Most Absent

Solution

The null and alternative hypotheses are as follows:

- H_0: The absent days occur with equal frequencies; that is, they fit a uniform distribution.
- H_a: The absent days occur with unequal frequencies; that is, they do not fit a uniform distribution.

If the absent days occur with equal frequencies, then, out of 60 absent days (the total in the sample: 15 + 12 + 9 + 9 + 15 = 60) there would be 12 absences on Monday, 12 on Tuesday, 12 on Wednesday, 12 on Thursday, and 12 on Friday. These numbers are the *expected* (*E*) values. The values in the table are the *observed* (*O*) values or data.

This time, calculate the χ^2 test statistic by hand. Make a chart with the following headings and fill in the columns:

Expected (*E*) values (12, 12, 12, 12, 12)
Observed (*O*) values (15, 12, 9, 9, 15)
$(O - E)$
$(O - E)^2$
$\frac{(O - E)^2}{E}$

Now add (sum) the last column. The sum is three. This is the χ^2 test statistic.

To find the *p*-value, calculate $P(\chi^2 > 3)$. This test is right-tailed. Use a computer or calculator to find the *p*-value. You should get *p*-value = 0.5578.

The *df*s are the number of cells – 1 = 5 – 1 = 4.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Press 2nd DISTR. Arrow down to χ^2cdf. Press ENTER. Enter (3,10^99,4). Rounded to four decimal places, you should see .5578, which is the *p*-value.

Next, complete a graph like the following one with the proper labeling and shading. You should shade the right tail.

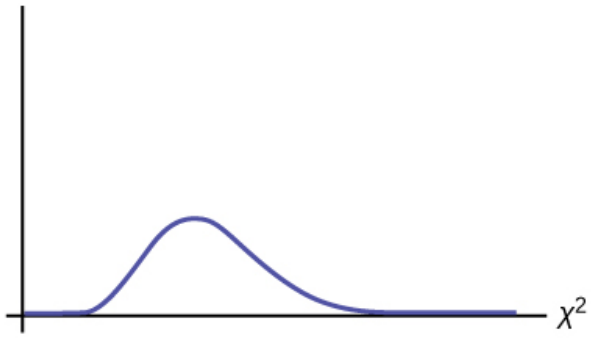


Figure 11.4

The decision is not to reject the null hypothesis.

Conclusion: At a 5 percent level of significance, from the sample data, there is not sufficient evidence to conclude that the absent days do not occur with equal frequencies.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

TI-83+ and some TI-84 calculators do not have a special program for the test statistic for the goodness-of-fit test. The next example, Example 11.3, has the calculator instructions. The newer TI-84 calculators have in STAT TESTS the test Chi2 GOF. To run the test, put the observed values—the data—into a first list and the expected values—the values you expect if the null hypothesis is true—into a second list. Press STAT TESTS and Chi2 GOF. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press Calculate or Draw. Make sure you clear any lists before you start. **To Clear Lists in the calculators:** Go into STAT EDIT and arrow up to the list name area of the particular list. Press CLEAR and then arrow down. The list will be cleared. Alternatively, you can press STAT and press 4 for ClrList. Enter the list name and press ENTER.

TRY IT 11.2

Teachers want to know which night each week their students are doing most of their homework. Most teachers think that students do homework equally throughout the week. Suppose a random sample of 56 students were asked on which night of the week they did the most homework. The results were distributed as in Table 11.8.

	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Number of Students	11	8	10	7	10	5	5

Table 11.8

From the population of students, do the nights for the highest number of students doing the majority of their homework occur with equal frequencies during a week? What type of hypothesis test should you use?

EXAMPLE 11.3

One study indicates that the number of televisions that American families have is distributed (this is the *given* distribution for the American population) as in Table 11.9.

Number of Televisions	Percent
0	10
1	16
2	55
3	11
4+	8

Table 11.9

The table contains expected (*E*) percents.

A random sample of 600 families in the far western U.S. resulted in the data in Table 11.10.

Number of Televisions	Frequency
0	66
1	119
2	340
3	60
4+	15
	Total = 600

Table 11.10

The table contains observed (*O*) frequency values.

Problem

At the 1 percent significance level, does it appear that the distribution *number of televisions* of far western U.S. families is different from the distribution for the American population as a whole?

Solution

This problem asks you to test whether the far western U.S. families distribution fits the distribution of the American families. This test is always right-tailed.

The first table contains expected percentages. To get expected (*E*) frequencies, multiply the percentage by 600. The expected frequencies are shown in Table 11.11.

Number of Televisions	Percent	Expected Frequency
0	10	(0.10)(600) = 60
1	16	(0.16)(600) = 96
2	55	(0.55)(600) = 330
3	11	(0.11)(600) = 66
more than 3	8	(0.08)(600) = 48

Table 11.11

Therefore, the expected frequencies are 60, 96, 330, 66, and 48. In the TI calculators, you can let the calculator do the math. For example, instead of 60, enter 0.10 * 600.

H_0: The *number of televisions* distribution of far western U.S. families is the same as the *number of televisions* distribution of the American population.

H_a: The *number of televisions* distribution of far western U.S. families is different from the *number of televisions* distribution of the American population.

Distribution for the test: χ^2_4 where *df* = (the number of cells) – 1 = 5 – 1 = 4.

Note

$df \neq 600 - 1$

Calculate the test statistic: $\chi^2 = 29.65$

Graph

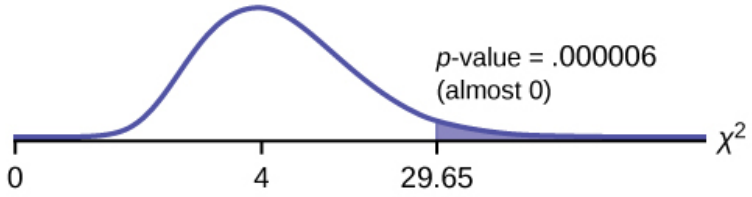


Figure 11.5

Probability statement: $p\text{-value} = P(\chi^2 > 29.65) = .000006$

Compare α and the p-value:

- $\alpha = .01$
- p-value = 0.000006

So, $\alpha > p$-value.

Make a decision: Since $\alpha > p$-value, reject H_o.

This means you reject the hypothesis that the distribution for the far western states is the same as that of the American population as a whole.

Conclusion: At the 1 percent significance level, from the data, there is sufficient evidence to conclude that the *number of televisions* distribution for the far western United States is different from the *number of televisions* distribution for the American population as a whole.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Press STAT and ENTER. Make sure to clear lists L1, L2, and L3 if they have data in them—see the note at the end of Example 11.2. Into L1, put the observed frequencies 66, 119, 349, 60, 15. Into L2, put the expected frequencies `.10*600, .16*600, .55*600, .11*600, .08*600`. Arrow over to list L3 and up to the name area L3. Enter `(L1-L2)^2/L2` and ENTER. Press `2nd QUIT`. Press `2nd LIST` and arrow over to `MATH`. Press 5. You should see `sum(Enter L3)`. Rounded to two decimal places, you should see `29.65`. Press `2nd DISTR`. Press 7 or Arrow down to `7:χ2cdf` and press ENTER. Enter `(29.65,1E99,4)`. Rounded to four places, you should see `5.77E-6 = .000006` (rounded to six decimal places), which is the p-value.

The newer TI-84 calculators have in `STAT TESTS` the test `Chi2 GOF`. To run the test, put the observed values (the data) into a first list and the expected values—the values you expect if the null hypothesis is true—into a second list. Press `STAT TESTS` and `Chi2 GOF`. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press `Calculate` or `Draw`. Make sure you clear any lists before you start.

TRY IT 11.3

The expected percentage of the number of pets students have in their homes is distributed (this is the given distribution for the student population of the United States) as in Table 11.12.

Number of Pets	Percent
0	18
1	25
2	30
3	18
4+	9

Table 11.12

A random sample of 1,000 students from the eastern United States resulted in the data in Table 11.13.

Number of Pets	Frequency
0	210
1	240
2	320
3	140
4+	90

Table 11.13

At the 1 percent significance level, does it appear that the distribution *number of pets* of students in the eastern United States is different from the distribution for the United States student population as a whole? What is the *p*-value?

EXAMPLE 11.4

Problem

Suppose you flip two coins 100 times. The results are 20 *HH*, 27 *HT*, 30 *TH*, and 23 *TT*. Are the coins fair? Test at a 5 percent significance level.

Solution

This problem can be set up as a goodness-of-fit problem. The sample space for flipping two fair coins is {*HH*, *HT*, *TH*, *TT*}. Out of 100 flips, you would expect 25 *HH*, 25 *HT*, 25 *TH*, and 25 *TT*. This is the expected distribution. The question, "Are the coins fair?" is the same as saying, "Does the distribution of the coins (20 *HH*, 27 *HT*, 30 *TH*, 23 *TT*) fit the expected distribution?"

Random variable: Let X = the number of heads in one flip of the two coins. X takes on the values 0, 1, 2. There are 0, 1, or 2 heads in the flip of two coins. Therefore, the number of cells is three. Since X = the number of heads, the observed frequencies are 20 for two heads, 57 for one head, and 23 for zero heads or both tails. The expected frequencies are 25 for two heads, 50 for one head, and 25 for zero heads or both tails. This test is right-tailed.

H_0: The coins are fair.

H_a: The coins are not fair.

Distribution for the test: χ^2_2 where $df = 3 - 1 = 2$.

Calculate the test statistic: $\chi^2 = 2.14$.

Graph

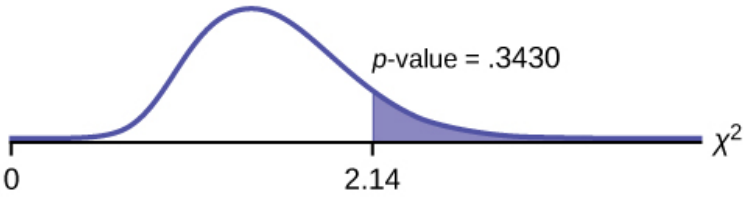


Figure 11.6

Probability statement: p-value = $P(\chi^2 > 2.14) = 0.3430$.

Compare α and the p-value:

- $\alpha = .05$
- p-value = 0.3430

$\alpha < p$-value.

Make a decision: Since $\alpha < p$-value, do not reject H_0.

Conclusion: There is insufficient evidence to conclude that the coins are not fair.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Press STAT and ENTER. Make sure you clear lists L1, L2, and L3 if they have data in them. Into L1, put the observed frequencies 20, 57, 23. Into L2, put the expected frequencies 25, 50, 25. Arrow over to list L3 and up to the name area L3. Enter `(L1-L2)^2/L2` and ENTER. Press 2nd QUIT. Press 2nd LIST and arrow over to MATH. Press 5. You should see `sum`. Enter L3. Rounded to two decimal places, you should see `2.14`. Press 2nd DISTR. Arrow down to `7:χ2cdf`—or press 7. Press ENTER. Enter `2.14,1E99,2)`. Rounded to four places, you should see `.3430`, which is the p-value.

The newer TI-84 calculators have in STAT TESTS the test `Chi2 GOF`. To run the test, put the observed values—the data—into a first list and the expected values—the values you expect if the null hypothesis is true—into a second list. Press STAT TESTS and `Chi2 GOF`. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press `Calculate` or `Draw`. Make sure you clear any lists before you start.

TRY IT 11.4

Students in a social studies class hypothesize that the literacy rates around the world for every region are 82 percent. Table 11.14 shows the actual literacy rates around the world broken down by region. What are the test statistic and the degrees of freedom?

MDG Region	Adult Literacy Rate (%)
Developed regions	99
Commonwealth of Independent States	99.5
Northern Africa	67.3
Sub-Saharan Africa	62.5
Latin America and the Caribbean	91
Eastern Asia	93.8
Southern Asia	61.9
Southeastern Asia	91.9
Western Asia	84.5
Oceania	66.4

Table 11.14

11.3 Test of Independence

Tests of independence involve using a **contingency table** of observed (data) values.

The test statistic for a **test of independence** is similar to that of a goodness-of-fit test

$$\sum_{(i \cdot j)} \frac{(O-E)^2}{E}$$

where

- O = observed values,
- E = expected values,
- i = the number of rows in the table, and
- j = the number of columns in the table.

There are $i \cdot j$ terms of the form $\frac{(O-E)^2}{E}$.

A test of independence determines whether two factors are independent. You first encountered the term *independence* in Probability Topics. As a review, consider the following example.

Note

The expected value for each cell needs to be at least five for you to use this test.

EXAMPLE 11.5

Suppose A = a speeding violation in the last year and B = a cell phone user while driving. If A and B are independent, then $P(A \text{ AND } B) = P(A)P(B)$. A AND B is the event that a driver received a speeding violation last year and also used a cell phone while driving. Suppose, in a study of drivers who received speeding violations in the last year, and who used cell

phones while driving, that 755 people were surveyed. Out of the 755, 70 had a speeding violation and 685 did not; 305 used cell phones while driving and 450 did not.

Let y = expected number of drivers who used a cell phone while driving and received speeding violations.

If A and B are independent, then $P(A \text{ AND } B) = P(A)P(B)$. By substitution,

$$\frac{y}{755} = \left(\frac{70}{755}\right)\left(\frac{305}{755}\right).$$

Solve for y: $y = \frac{(70)(305)}{755} = 28.3.$

About 28 people from the sample are expected to use cell phones while driving and to receive speeding violations.

In a test of independence, we state the null and alternative hypotheses in words. Since the contingency table consists of two factors, the null hypothesis states that the factors are independent and the alternative hypothesis states that they are not independent (dependent). If we do a test of independence using the example, then the null hypothesis is the following:

H_0: Being a cell phone user while driving and receiving a speeding violation are independent events.

If the null hypothesis were true, we would expect about 28 people to use cell phones while driving and to receive a speeding violation.

The test of independence is always right-tailed because of the calculation of the test statistic. If the expected and observed values are not close together, then the test statistic is very large and way out in the right tail of the chi-square curve, as it is in a goodness-of-fit.

The number of degrees of freedom for the test of independence is

df = (number of columns – 1)(number of rows – 1).

The following formula calculates the **expected number** (E):

$$E = \frac{(\text{row total})(\text{column total})}{\text{total number surveyed}}$$

TRY IT 11.5

A sample of 300 students is taken. Of the students surveyed, 50 were music students, while 250 were not. 97 were on the honor roll, while 203 were not. If we assume being a music student and being on the honor roll are independent events, what is the expected number of music students who are also on the honor roll?

EXAMPLE 11.6

In a volunteer group, adults 21 and older volunteer from one to nine hours each week to spend time with a disabled senior citizen. The program recruits among community college students, four-year college students, and non-students. In Table 11.15 is a sample of the adult volunteers and the number of hours they volunteer per week.

Type of Volunteer	1–3 Hours	4–6 Hours	7–9 Hours	Row Total
Community College Students	111	96	48	255
Four-year College Students	96	133	61	290

Table 11.15 Number of Hours Worked per Week by Volunteer Type (Observed) The table contains **observed (O)** values (data).

Type of Volunteer	1–3 Hours	4–6 Hours	7–9 Hours	Row Total
Non-students	91	150	53	294
Column Total	298	379	162	839

Table 11.15 Number of Hours Worked per Week by Volunteer Type (Observed) The table contains **observed (O)** values (data).

Problem

Is the number of hours volunteered **independent** of the type of volunteer?

Solution

The **observed values** and the question at the end of the problem, "Is the number of hours volunteered independent of the type of volunteer?" tell you this is a test of independence. The two factors are *number of hours volunteered* and *type of volunteer*. This test is always right-tailed.

H_0: The number of hours volunteered is *independent* of the type of volunteer.

H_a: The number of hours volunteered is *dependent* on the type of volunteer.

The expected result are in Table 11.16.

Type of Volunteer	1-3 Hours	4–6 Hours	7-9 Hours
Community College Students	90.57	115.19	49.24
Four-Year College Students	103	131	56
Nonstudents	104.42	132.81	56.77

Table 11.16 Number of Hours Worked per Week by Volunteer Type (Expected) The table contains **expected** (*E*) values (data).

For example, the calculation for the expected frequency for the top-left cell is

$$E = \frac{(\text{row total})(\text{column total})}{\text{total number surveyed}} = \frac{(255)\,(298)}{839} = 90.57.$$

Calculate the test statistic: $\chi^2 = 12.99$ (calculator or computer)

Distribution for the test: χ^2_4

df = (3 columns – 1)(3 rows – 1) = (2)(2) = 4

Graph

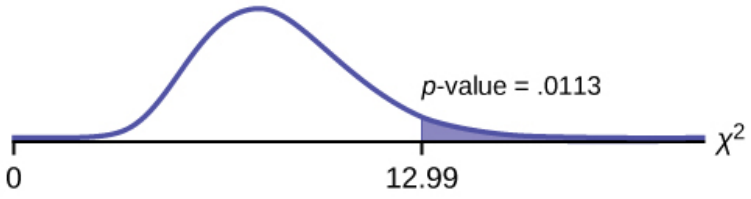


Figure 11.7

Probability statement: *p*-value = $P(\chi^2 > 12.99) = 0.0113$

Compare *α* and the *p*-value: Since no *α* is given, assume $\alpha = 0.05$. *p*-value = 0.0113. $\alpha >$ *p*-value.

Make a decision: Since $\alpha >$ *p*-value, reject H_0. This means that the factors are not independent.

Conclusion: At a 5 percent level of significance, from the data, there is sufficient evidence to conclude that the number of hours volunteered and the type of volunteer are dependent on each other.

For the example in Table 11.16, if there had been another type of volunteer, teenagers, what would the degrees of freedom be?

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Press the MATRX key and arrow over to EDIT. Press 1:[A]. Press 3 ENTER 3 ENTER. Enter the observed values for each group (but not the totals) by row from Table 11.15. Press ENTER after each. Press 2nd QUIT. Press the MATRX key and arrow over to EDIT. Press 2:[B]. Press 3 ENTER 3 ENTER. Enter the expected values by row from Table 11.16. Press ENTER after each. Press 2nd QUIT. Press STAT and arrow over to TESTS. Arrow down to C:χ2-TEST. Press ENTER. You should see Observed:[A] and Expected:[B]. Arrow down to Calculate. Press ENTER. The test statistic is 12.9909 and the *p*-value = .0113. Do the procedure a second time, but arrow down to Draw instead of Calculate.

TRY IT 11.6

The Bureau of Labor Statistics gathers data about employment in the United States. A sample is taken to calculate the number of U.S. citizens working in one of several industry sectors over time. Table 11.17 shows the results:

Industry Sector	2000	2010	2020	Total
Non-agriculture Wage and Salary	13,243	13,044	15,018	41,305
Goods-producing, Excluding Agriculture	2,457	1,771	1,950	6,178
Services-providing	10,786	11,273	13,068	35,127
Agriculture, Forestry, Fishing, and Hunting	240	214	201	655
Non-agriculture Self-employed and Unpaid Family Worker	931	894	972	2,797
Secondary Wage and Salary Jobs in Agriculture and Private Household Industries	14	11	11	36
Secondary Jobs as a Self-employed or Unpaid Family Worker	196	144	152	492
Total	27,867	27,351	31,372	86,590

Table 11.17

We want to know if the change in the number of jobs is independent of the change in years. State the null and alternative hypotheses and the degrees of freedom.

EXAMPLE 11.7

De Anza College is interested in the relationship between anxiety level and the need to succeed in school. A random sample of 400 students took a test that measured anxiety level and need to succeed in school. Table 11.18 shows the results. De Anza College wants to know if anxiety level and need to succeed in school are independent events.

Need to Succeed in School	High Anxiety	Med-High Anxiety	Medium Anxiety	Med-Low Anxiety	Low Anxiety	Row Total
High Need	35	42	53	15	10	155
Medium Need	18	48	63	33	31	193
Low Need	4	5	11	15	17	52
Column Total	57	95	127	63	58	400

Table 11.18 Need to Succeed in School vs. Anxiety Level

Problem

a. How many high anxiety level students are expected to have a high need to succeed in school?

Solution

a. The column total for a high anxiety level is 57. The row total for high need to succeed in school is 155. The sample size or total surveyed is 400.

$$E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}} = \frac{155 \cdot 57}{400} = 22.09$$

The expected number of students who have a high anxiety level and a high need to succeed in school is about 22.

Problem

b. If the two variables are independent, how many students do you expect to have a low need to succeed in school and a med-low level of anxiety?

Solution

b. The column total for a med-low anxiety level is 63. The row total for a low need to succeed in school is 52. The sample size or total surveyed is 400.

Problem

c. $E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}} =$ _____

Solution

c. $E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}} = 8.19$

Problem

d. The expected number of students who have a med-low anxiety level and a low need to succeed in school is about _____.

Solution

d. 8

TRY IT 11.7

Refer back to the information in Try It 11.6. How many services-providing jobs are there expected to be in 2020? How many nonagriculture wage and salary jobs are there expected to be in 2020?

11.4 Test for Homogeneity

The goodness-of-fit test can be used to decide whether a population fits a given distribution, but it will not suffice to decide whether two populations follow the same unknown distribution. A different test, called the **test for homogeneity**, can be used to draw a conclusion about whether two populations have the same distribution. To calculate

the test statistic for a test for homogeneity, follow the same procedure as with the test of independence.

Note

The expected value for each cell needs to be at least five for you to use this test.

Hypotheses

H_0: The distributions of the two populations are the same.

H_a: The distributions of the two populations are not the same.

Test Statistic
Use a χ^2 test statistic. It is computed in the same way as the test for independence.

Degrees of freedom (*df*)
df = (number of columns – 1)(number of rows – 1)

Requirements
All values in the table must be greater than or equal to five.

Common Uses
Comparing two populations. For example: men vs. women, before vs. after, east vs. west. The variable is categorical with more than two possible response values.

EXAMPLE 11.8

Problem
Do male and female college students have the same distribution of living arrangements? Use a level of significance of 0.05. Suppose that 250 randomly selected male college students and 300 randomly selected female college students were asked about their living arrangements: dormitory, apartment, with parents, other. The results are shown in Table 11.19. Do male and female college students have the same distribution of living arrangements?

	Dormitory	Apartment	With Parents	Other
Males	72	84	49	45
Females	91	86	88	35

Table 11.19 Distribution of Living Arragements for College Males and College Females

Solution
H_0: The distribution of living arrangements for male college students is the same as the distribution of living arrangements for female college students.

H_a: The distribution of living arrangements for male college students is not the same as the distribution of living arrangements for female college students.

Degrees of freedom (*df*):
df = (number of columns – 1)(number of rows – 1) = (4 – 1)(2– 1) = 3

Distribution for the test: χ^2_3

Calculate the test statistic: χ^2 = 10.1287 (calculator or computer)

Probability statement: *p*-value = $P(\chi^2 > 10.1287)$ = 0.0175

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Press the

`MATRX`

key and arrow over to

`EDIT`

. Press

`1:[A]`

. Press

`2 ENTER 4 ENTER`

. Enter the table values by row. Press

`ENTER`

after each. Press

`2nd QUIT`

. Press

`STAT`

and arrow over to

`TESTS`

. Arrow down to

`C:χ2-TEST`

. Press

`ENTER`

. You should see

`Observed:[A]`

and

`Expected:[B]`

. Arrow down to

`Calculate`

. Press

`ENTER`

. The test statistic is 10.1287 and the p-value = 0.0175. Do the procedure a second time but arrow down to

`Draw`

instead of

`Calculate`

.

Compare α and the p-value: Since no α is given, assume $\alpha = 0.05$. p-value = 0.0175. $\alpha > p$-value.

Make a decision: Since $\alpha > p$-value, reject H_0. This means that the distributions are not the same.

Conclusion: At a 5 percent level of significance, from the data, there is sufficient evidence to conclude that the distributions of living arrangements for male and female college students are not the same.

Notice that the conclusion is only that the distributions are not the same. We cannot use the test for homogeneity to draw any conclusions about how they differ.

TRY IT 11.8

Do families and singles have the same distribution of cars? Suppose that 100 randomly selected families and 200 randomly selected singles were asked what type of car they drove: sport, sedan, hatchback, truck, van/SUV. The results are shown in Table 11.20. Do families and singles have the same distribution of cars? Test at a level of significance of 0.05.

	Sport	Sedan	Hatchback	Truck	Van/SUV
Family	5	15	35	17	28
Single	45	65	37	46	7

Table 11.20

EXAMPLE 11.9

Problem

Both before and after a recent earthquake, surveys were conducted asking voters which of the three candidates they planned on voting for in the upcoming city council election. Has there been a change since the earthquake? Use a level of significance of 0.05. Table 11.21 shows the results of the survey. Has there been a change in the distribution of voter preferences since the earthquake?

	Perez	Chung	Stevens
Before	167	128	135
After	214	197	225

Table 11.21

Solution

H_0: The distribution of voter preferences was the same before and after the earthquake.

H_a: The distribution of voter preferences was not the same before and after the earthquake.

Degrees of freedom (*df*):
df = (number of columns – 1)(number of rows – 1) = (3 – 1)(2 – 1) = 2

Distribution for the test: χ^2_2

Calculate the test statistic: χ^2 = 3.2603 (calculator or computer)

Probability statement: *p*-value=$P(\chi^2 > 3.2603)$ = 0.1959

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Press the MATRX key and arrow over to EDIT. Press 1:[A]. Press 2 ENTER 3 ENTER. Enter the table values by row. Press ENTER after each. Press 2nd QUIT. Press STAT and arrow over to TESTS. Arrow down to C:χ2-TEST. Press ENTER. You should see Observed:[A] and Expected:[B]. Arrow down to Calculate. Press ENTER. The test statistic is 3.2603 and the *p*-value = 0.1959. Do the procedure a second time but arrow down to Draw instead of Calculate.

Compare *α* and the *p*-value: α = 0.05 and the *p*-value = 0.1959. α < *p*-value.

Make a decision: Since α < *p*-value, do not reject H_o.

Conclusion: At a 5 percent level of significance, from the data, there is insufficient evidence to conclude that the distribution of voter preferences was not the same before and after the earthquake.

TRY IT 11.9

Ivy League schools receive many applications, but only some can be accepted. At the schools listed in Table 11.22, two types of applications are accepted: regular and early decision.

Application Type Accepted	Brown	Columbia	Cornell	Dartmouth	Penn	Yale
Regular	2,115	1,792	5,306	1,734	2,685	1,245
Early Decision	577	627	1,228	444	1,195	761

Table 11.22

We want to know if the number of regular applications accepted follows the same distribution as the number of early applications accepted. State the null and alternative hypotheses, the degrees of freedom and the test statistic, sketch the graph of the *p*-value, and draw a conclusion about the test of homogeneity.

11.5 Comparison of the Chi-Square Tests

You have seen the χ^2 test statistic used in three different circumstances. The following bulleted list is a summary that will help you decide which χ^2 test is the appropriate one to use.

- **Goodness-of-Fit:** Use the goodness-of-fit test to decide whether a population with an unknown distribution *fits* a known distribution. In this case there will be a single qualitative survey question or a single outcome of an experiment from a single population. Goodness-of-fit is typically used to see if the population is uniform (all outcomes occur with equal frequency), the population is normal, or the population is the same as another population with a known distribution. The null and alternative hypotheses are as follows:
  H_0: The population fits the given distribution.
  H_a: The population does not fit the given distribution.
- **Independence:** Use the test for independence to decide whether two variables (factors) are independent or dependent. In this case there will be two qualitative survey questions or experiments and a contingency table will be constructed. The goal is to see if the two variables are unrelated/independent or related/dependent. The null and alternative hypotheses are as follows:
  H_0: The two variables (factors) are independent.
  H_a: The two variables (factors) are dependent.
- **Homogeneity:** Use the test for homogeneity to decide if two populations with unknown distributions have the same distribution. In this case there will be a single qualitative survey question or experiment given to two different populations. The null and alternative hypotheses are as follows:
  H_0: The two populations follow the same distribution.
  H_a: The two populations have different distributions.

11.6 Test of a Single Variance

A **test of a single variance** assumes that the underlying distribution is normal. The null and alternative hypotheses are stated in terms of the **population variance** or population standard deviation. The test statistic is

$$\frac{(n-1)s^2}{\sigma^2}$$

where

- n = the total number of data,
- s^2 = sample variance, and
- σ^2 = population variance.

You may think of s as the random variable in this test. The number of degrees of freedom is $df = n - 1$. A test of a single variance may be right-tailed, left-tailed, or two-tailed. Example 11.10 will show you how to set up the null and alternative hypotheses. The null and alternative hypotheses contain statements about the population variance.

EXAMPLE 11.10

Problem

Math instructors are not only interested in how their students do on exams, on average, but how the exam scores vary. To many instructors, the variance, or standard deviation, may be more important than the average.

Suppose a math instructor believes that the standard deviation for his final exam is five points. One of his best students thinks otherwise. The student claims that the standard deviation is more than five points. If the student were to conduct a hypothesis test, what would the null and alternative hypotheses be?

Solution

Even though we are given the population standard deviation, we can set up the test using the population variance as follows:

- H_0: $\sigma^2 = 5^2$
- H_a: $\sigma^2 > 5^2$

TRY IT 11.10

A scuba instructor wants to record the collective depths each of his students dives during their checkout. He is interested in how the depths vary, even though everyone should have been at the same depth. He believes the standard deviation is three feet. His assistant thinks the standard deviation is less than three feet. If the instructor were to conduct a test, what would the null and alternative hypotheses be?

EXAMPLE 11.11

Problem

With individual lines at its various windows, a post office finds that the standard deviation for normally distributed waiting times for customers on Friday afternoon is 7.2 minutes. The post office experiments with a single, main waiting line and finds that for a random sample of 25 customers, the waiting times for customers have a standard deviation of 3.5 minutes.

With a significance level of 5 percent, test the claim that a single line causes lower variation among waiting times (shorter waiting times) for customers.

Solution

Since the claim is that a single line causes less variation, this is a test of a single variance. The parameter is the population variance, σ^2, or the population standard deviation, σ.

Random variable: The sample standard deviation, s, is the random variable. Let s = standard deviation for the waiting times.

- H_0: $\sigma^2 = 7.2^2$
- H_a: $\sigma^2 < 7.2^2$

The word *less* tells you this is a left-tailed test.

Distribution for the test: χ^2_{24}, where

- n = the number of customers sampled, and
- $df = n - 1 = 25 - 1 = 24$.

Calculate the test statistic:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(25-1)(3.5)^2}{7.2^2} = 5.67$$

where $n = 25$, $s = 3.5$, and $\sigma = 7.2$.

Graph

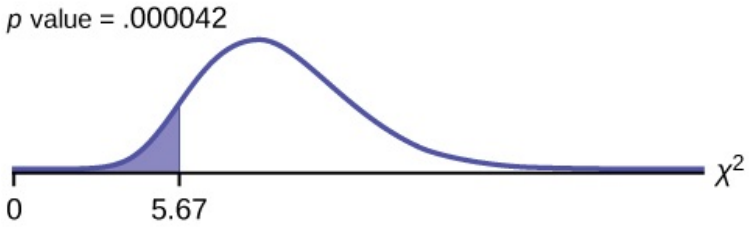


Figure 11.8

Probability statement: p-value = $P(\chi^2 < 5.67) = 0.000042$

Compare α and the p-value:
$\alpha = 0.05$
p-value = 0.000042
$\alpha > p$-value

Make a decision: Since $\alpha > p$-value, reject H_0. This means that you reject $\sigma^2 = 7.2^2$. In other words, you do not think the variation in waiting times is 7.2 minutes; you think the variation in waiting times is less.

Conclusion: At a 5 percent level of significance, from the data, there is sufficient evidence to conclude that a single line causes a lower variation among the waiting times *or* with a single line, the customer waiting times vary less than 7.2 minutes.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

In `2nd DISTR`, use `7:χ2cdf`. The syntax is `(lower, upper, df)` for the parameter list. For Example 11.11, `χ2cdf(-1E99,5.67,24)`. The p-value = 0.000042.

TRY IT 11.11

The FCC conducts broadband speed tests to measure how much data per second passes between a consumer's computer and the internet. As of August 2012, the standard deviation of internet speeds across internet service providers (ISPs) was 12.2 percent. Suppose a sample of 15 ISPs is taken, and the standard deviation is 13.2. An analyst claims that the standard deviation of speeds is more than what was reported. State the null and alternative hypotheses, compute the degrees of freedom, calculate the test statistic, sketch the graph of the p-value, and draw a conclusion. Test at the 1 percent significance level.

11.7 Lab 1: Chi-Square Goodness-of-Fit

Stats Lab

Lab 1: Chi-Square Goodness-of-Fit

Student Learning Outcome

- The student will evaluate data collected to determine if they fit either the uniform or exponential distributions.

Collect the Data

Go to your local supermarket. Ask 30 people as they leave for the total amount on their grocery receipts. Or, ask 3 cashiers for the last 10 amounts. Be sure to include the express lane, if it is open.

Note

You may need to combine two categories so that each cell has an expected value of at least five.

1. Record the values.

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

Table 11.23

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

Table 11.23

2. Construct a histogram of the data. Make five to six intervals. Sketch the graph using a ruler and pencil. Scale the axes.

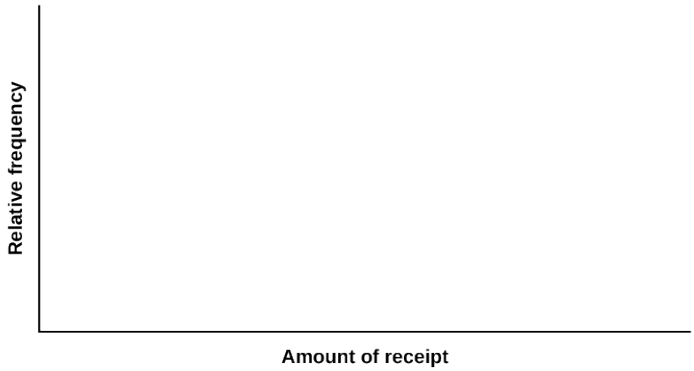


Figure 11.9

3. Calculate the following:
   a. $\overline{x}$ = _____
   b. s = _____
   c. s^2 = _____

Uniform Distribution

Test to see if grocery receipts follow the uniform distribution.

1. Using your lowest and highest values, $X \sim U$ (_____, _____).
2. Divide the distribution into fifths.
3. Calculate the following:
   a. lowest value = _____
   b. 20^{th} percentile = _____
   c. 40^{th} percentile = _____
   d. 60^{th} percentile = _____
   e. 80^{th} percentile = _____
   f. highest value = _____
4. For each fifth, count the observed number of receipts and record it. Then determine the expected number of receipts and record that.

Fifth	Observed	Expected
1^{st}		
2^{nd}		
3^{rd}		
4^{th}		
5^{th}		

Table 11.24

5. H_0: _____
6. H_a: _____
7. What distribution should you use for a hypothesis test?
8. Why did you choose this distribution?
9. Calculate the test statistic.
10. Find the *p*-value.
11. Sketch a graph of the situation. Label and scale the *x*-axis. Shade the area corresponding to the *p*-value.

Figure 11.10

12. State your decision.
13. State your conclusion in a complete sentence.

Exponential Distribution

Test to see if grocery receipts follow the exponential distribution with decay parameter $\frac{1}{\bar{x}}$.

1. Using $\frac{1}{\bar{x}}$ as the decay parameter, $X \sim Exp$(_____).
2. Calculate the following:
   a. lowest value = _____
   b. first quartile = _____
   c. 37^{th} percentile = _____
   d. median = _____
   e. 63^{rd} percentile = _____
   f. 3^{rd} quartile = _____
   g. highest value = _____
3. For each cell, count the observed number of receipts and record it. Then determine the expected number of receipts and record that.

Cell	Observed	Expected
1^{st}		
2^{nd}		
3^{rd}		
4^{th}		
5^{th}		
6^{th}		

Table 11.25

4. H_0: _____
5. H_a: _____
6. What distribution should you use for a hypothesis test?
7. Why did you choose this distribution?
8. Calculate the test statistic.
9. Find the *p*-value.
10. Sketch a graph of the situation. Label and scale the *x*-axis. Shade the area corresponding to the *p*-value.

Figure 11.11

11. State your decision.
12. State your conclusion in a complete sentence.

Discussion Questions

1. Did your data fit either distribution? If so, which?
2. In general, do you think it's likely that data could fit more than one distribution? In complete sentences, explain why or why not.

11.8 Lab 2: Chi-Square Test of Independence

Stats Lab

Lab 2: Chi-Square Test of Independence

Student Learning Outcome

- The student will evaluate if there is a significant relationship between favorite type of snack and gender.

Collect the Data

1. Using your class as a sample, complete the following chart. Ask one another what your favorite snack is, then total the results.

Note

You may need to combine two food categories so that each cell has an expected value of at least five.

	Sweets (candy & baked goods)	Ice Cream	Chips & Pretzels	Fruits & Vegetables	Total
Male					
Female					
Total					

Table 11.26 Favorite Type of Snack

2. Looking at Table 11.26, does it appear to you that there is a dependence between gender and favorite type of snack food? Why or why not?

Hypothesis Test

Conduct a hypothesis test to determine if the factors are independent:

1. H_0: _____
2. H_a: _____
3. What distribution should you use for a hypothesis test?
4. Why did you choose this distribution?
5. Calculate the test statistic.
6. Find the p value.
7. Sketch a graph of the situation. Label and scale the x axis. Shade the area corresponding to the p value.

Figure 11.12

8. State your decision.
9. State your conclusion in a complete sentence.

Discussion Questions

1. Is the conclusion of your study the same as or different from your answer to answer to Question 2 under Collect the Data?
2. Why do you think that occurred?

Key Terms

contingency table a table that displays sample values for two different factors that may be dependent or contingent on each other; facilitates determining conditional probabilities

Chapter Review

11.1 Facts About the Chi-Square Distribution

The chi-square distribution is a useful tool for assessment in a series of problem categories. These problem categories include primarily (i) whether a data set fits a particular distribution, (ii) whether the distributions of two populations are the same, (iii) whether two events might be independent, and (iv) whether there is a different variability than expected within a population.

An important parameter in a chi-square distribution is the degrees of freedom *df* in a given problem. The random variable in the chi-square distribution is the sum of squares of *df* standard normal variables, which must be independent. The key characteristics of the chi-square distribution also depend directly on the degrees of freedom.

The chi-square distribution curve is skewed to the right, and its shape depends on the degrees of freedom *df*. For $df > 90$, the curve approximates the normal distribution. Test statistics based on the chi-square distribution are always greater than or equal to zero. Such application tests are almost always right-tailed tests.

11.2 Goodness-of-Fit Test

To assess whether a data set fits a specific distribution, you can apply the goodness-of-fit hypothesis test that uses the chi-square distribution. The null hypothesis for this test states that the data come from the assumed distribution. The test compares observed values against the values you would expect to have if your data followed the assumed distribution. The test is almost always right-tailed. Each observation or cell category must have an expected value of at least five.

11.3 Test of Independence

To assess whether two factors are independent, you can apply the test of independence that uses the chi-square distribution. The null hypothesis for this test states that the two factors are independent. The test compares observed values to expected values. The test is right-tailed. Each observation or cell category must have an expected value of at least five.

11.4 Test for Homogeneity

To assess whether two data sets are derived from the same distribution, which need not be known, you can apply the test for homogeneity that uses the chi-square distribution. The null hypothesis for this test states that the populations of the two data sets come from the same distribution. The test compares the observed values against the expected values if the two populations followed the same distribution. The test is right-tailed. Each observation or cell category must have an expected value of at least five.

11.5 Comparison of the Chi-Square Tests

The goodness-of-fit test is typically used to determine if data fits a particular distribution. The test of independence makes use of a contingency table to determine the independence of two factors. The test for homogeneity determines whether two populations come from the same distribution, even if this distribution is unknown.

11.6 Test of a Single Variance

To test variability, use the chi-square test of a single variance. The test may be left-, right-, or two-tailed, and its hypotheses are always expressed in terms of the variance or standard deviation.

Formula Review

11.1 Facts About the Chi-Square Distribution

$\chi^2 = (Z_1)^2 + (Z_2)^2 + \ldots (Z_{df})^2$ chi-square distribution random variable

$\mu_{\chi^2} = df$ chi-square distribution population mean

$\sigma_{\chi^2} = \sqrt{2(df)}$ chi-square distribution population standard deviation

11.2 Goodness-of-Fit Test

$\sum_k \frac{(O-E)^2}{E}$ goodness-of-fit test statistic where

O: observed values
E: expected values

k: number of different data cells or categories

$df = k - 1$ degrees of freedom

11.3 Test of Independence

Test of Independence

- The number of degrees of freedom is equal to (number of columns–1)(number of rows–1).
- The test statistic is $\sum_{(i \cdot j)} \frac{(O-E)^2}{E}$ where O = observed values, E = expected values, i = the number of rows in the table, and j = the number of columns in the table.
- If the null hypothesis is true, the expected number $E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}}$.

11.4 Test for Homogeneity

$\sum_{i \cdot j} \frac{(O-E)^2}{E}$ Homogeneity test statistic where O = observed values
E = expected values
i = number of rows in data contingency table
j = number of columns in data contingency table

$df = (i - 1)(j - 1)$ degrees of freedom

11.6 Test of a Single Variance

$\chi^2 = \frac{(n-1) \cdot s^2}{\sigma^2}$ Test of a single variance statistic where
n: sample size
s: sample standard deviation
σ: population standard deviation

$df = n - 1$ degrees of freedom

Test of a Single Variance

- Use the test to determine variation.
- The degrees of freedom is the number of samples – 1.
- The test statistic is $\frac{(n-1) \cdot s^2}{\sigma^2}$, where n = the total number of data, s^2 = sample variance, and σ^2 = population variance.
- The test may be left-, right-, or two-tailed.

Practice

11.1 Facts About the Chi-Square Distribution

1. If the number of degrees of freedom for a chi-square distribution is 25, what is the population mean and standard deviation?
2. If $df > 90$, the distribution is _____. If $df = 15$, the distribution is _____.
3. When does the chi-square curve approximate a normal distribution?
4. Where is μ located on a chi-square curve?
5. Is it more likely the df is 90, 20, or 2 in the graph?

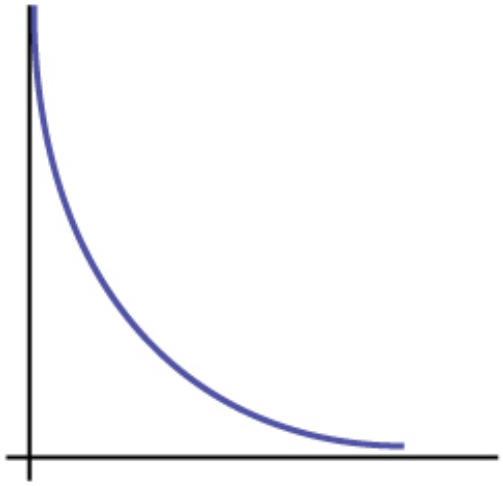


Figure 11.13

11.2 Goodness-of-Fit Test

Determine the appropriate test to be used in the next three exercises.

6. An archeologist is calculating the distribution of the frequency of the number of artifacts she finds in a dig site. Based on previous digs, the archeologist creates an expected distribution broken down by grid sections in the dig site. Once the site has been fully excavated, she compares the actual number of artifacts found in each grid

section to see if her expectation was accurate.

7. An economist is deriving a model to predict outcomes on the stock market. He creates a list of expected points on the stock market index for the next two weeks. At the close of each day's trading, he records the actual points on the index. He wants to see how well his model matched what actually happened.
8. A personal trainer is putting together a weight-lifting program for her clients. For a 90-day program, she expects each client to lift a specific maximum weight each week. As she goes along, she records the actual maximum weights her clients lifted. She wants to know how well her expectations met with what was observed.

Use the following information to answer the next five exercises. A teacher predicts the distribution of grades on the final exam. The predictions are shown in Table 11.27.

Grade	Proportion
A	0.25
B	0.30
C	0.35
D	0.10

Table 11.27

The actual distribution for a class of 20 is in Table 11.28.

Grade	Frequency
A	7
B	7
C	5
D	1

Table 11.28

9. $df =$ _____
10. State the null and alternative hypotheses.
11. χ^2 test statistic = _____
12. p-value = _____
13. At the 5 percent significance level, what can you conclude?

Use the following information to answer the next nine exercises. The cumulative number of cases of a chronic disease reported for Santa Clara County is broken down by ethnicity as in Table 11.29.

Ethnicity	Number of Cases
White	2,229
Hispanic	1,157

Table 11.29

Ethnicity	Number of Cases
Black/African American	457
Asian, Pacific Islander	232
	Total = 4,075

Table 11.29

The percentage of each ethnic group in Santa Clara County is as in Table 11.30.

Ethnicity	% of Total County Population	Number Expected (round to two decimal places)
White	42.9%	1,748.18
Hispanic	26.7%	
Black/African American	2.6%	
Asian, Pacific Islander	27.8%	
	Total = 100%	

Table 11.30

14. If the ethnicities of patients followed the ethnicities of the total county population, fill in the expected number of cases per ethnic group.
Perform a goodness-of-fit test to determine whether the occurrence of disease cases follows the ethnicities of the general population of Santa Clara County.

15. H_0: _____

16. H_a: _____

17. Is this a right-tailed, left-tailed, or two-tailed test?

18. degrees of freedom = _____

19. χ^2 test statistic = _____

20. *p*-value = _____

21. Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the *p*-value.

Figure 11.14

Let $\alpha = 0.05$.
Decision: _____
Reason for the decision: _____
Conclusion (write out in complete sentences): _____

22. Does it appear that the pattern of disease cases in Santa Clara County corresponds to the distribution of ethnic groups in this county? Why or why not?

11.3 Test of Independence

Determine the appropriate test to be used in the next three exercises.

23. A pharmaceutical company is interested in the relationship between age and presentation of symptoms for a common viral infection. A random sample is taken of 500 people with the infection across different age groups.

24. The owner of a baseball team is interested in the relationship between player salaries and team winning percentage. He takes a random sample of 100 players from different organizations.

25. A marathon runner is interested in the relationship between the brand of shoes runners wear and their run times. She takes a random sample of 50 runners and records their run times and the brand of shoes they were wearing.

Use the following information to answer the next seven exercises: Transit Railroads is interested in the relationship between travel distance and the ticket class purchased. A random sample of 200 passengers is taken. Table 11.31 shows the results. The railroad wants to know if a passenger's choice in ticket class is independent of the distance the passenger must travel.

Traveling Distance	Third Class	Second Class	First Class	Total
1–100 miles	21	14	6	41
101–200 miles	18	16	8	42
201–300 miles	16	17	15	48
301–400 miles	12	14	21	47
401–500 miles	6	6	10	22
Total	73	67	60	200

Table 11.31

26. State the hypotheses.
H_0: _____
H_a: _____

27. df = _____

28. How many passengers are expected to travel between 201 and 300 miles and purchase second-class tickets?

29. How many passengers are expected to travel between 401 and 500 miles and purchase first-class tickets?

30. What is the test statistic?

31. What is the p-value?

32. What can you conclude at the 5 percent level of significance?

Use the following information to answer the next ten exercises. An article in the *New England Journal of Medicine* discussed a study on people who used a certain product in California and Hawaii. In one part of the report, the self-reported ethnicity and product-use levels per day were given. Of the people using the product at most 10 times per day, there were 9,886 African Americans, 2,745 Native Hawaiians, 12,831 Latinos, 8,378 Japanese Americans, and 7,650 whites. Of the people using the product 11 to 20 times per day, there were 6,514 African Americans, 3,062 Native Hawaiians, 4,932 Latinos, 10,680 Japanese Americans, and 9,877 whites. Of the people using the product 21 to 30 times per day, there were 1,671 African Americans, 1,419 Native Hawaiians, 1,406 Latinos, 4,715 Japanese Americans, and 6,062 whites. Of the people using the product at least 31 times per day, there were 759 African Americans, 788 Native Hawaiians, 800 Latinos, 2,305 Japanese Americans, and 3,970 whites.

33. Complete the table.

Product use Per Day	African American	Native Hawaiian	Latino	Japanese American	White	TOTALS
1–10						
11–20						
21–30						
31+						
TOTALS						

Table 11.32

34. State the hypotheses.
H_0: _____
H_a: _____

35. Enter expected values in Table 11.32. Round to two decimal places.
Calculate the following values.

36. df = _____

37. χ^2 test statistic = _____

38. p-value = _____

39. Is this a right-tailed, left-tailed, or two-tailed test? Explain why.

40. Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the p-value.

Figure 11.15

State the decision and conclusion (in a complete sentence) for the following levels of α.

41. $\alpha = 0.05$
   a. Decision: _____
   b. Reason for the decision: _____
   c. Conclusion (write out in a complete sentence): _____

42. $\alpha = 0.01$
   a. Decision: _____
   b. Reason for the decision: _____
   c. Conclusion (write out in a complete sentence): _____

11.4 Test for Homogeneity

43. A math teacher wants to see if two of her classes have the same distribution of test scores. What test should she use?

44. What are the null and alternative hypotheses for Exercise 11.43?

45. A market researcher wants to see if two different stores have the same distribution of sales throughout the year. What type of test should he use?

46. A meteorologist wants to know if East and West Australia have the same distribution of storms. What type of test should she use?

47. What condition must be met to use the test for homogeneity?

Use the following information to answer the next five exercises. Do private practice doctors and hospital doctors have the same distribution of working hours? Suppose that a sample of 100 private practice doctors and 150 hospital doctors are selected at random and asked about the number of hours a week they work. The results are shown in Table 11.33.

	20–30	30–40	40–50	50–60
Private Practice	16	40	38	6
Hospital	8	44	59	39

Table 11.33

48. State the null and alternative hypotheses.

49. df = _____

50. What is the test statistic?

51. What is the p-value?

52. What can you conclude at the 5 percent significance level?

11.5 Comparison of the Chi-Square Tests

53. Which test do you use to decide whether an observed distribution is the same as an expected distribution?

54. What is the null hypothesis for the type of test from Exercise 11.53?

55. Which test would you use to decide whether two factors have a relationship?

56. Which test would you use to decide if two populations have the same distribution?

57. How are tests of independence similar to tests for homogeneity?

58. How are tests of independence different from tests for homogeneity?

11.6 Test of a Single Variance

Use the following information to answer the next three exercises. An archer's standard deviation for his hits is six, where the data are measured in distance from the center of the target. An observer claims the standard deviation is less than six.

59. What type of test should be used?

60. State the null and alternative hypotheses.

61. Is this a right-tailed, left-tailed, or two-tailed test?

Use the following information to answer the next three exercises. The standard deviation of heights for students in a school is 0.81. A random sample of 50 students is taken, and the standard deviation of heights of the sample is 0.96. A researcher in charge of the study believes the standard deviation of heights for the school is greater than 0.81.

62. What type of test should be used?

63. State the null and alternative hypotheses.

64. df = _____

Use the following information to answer the next four exercises: The average waiting time in a doctor's office varies. The standard deviation of waiting times in a doctor's office is 3.4 minutes. A random sample of 30 patients in the doctor's office has a standard deviation of waiting times of 4.1 minutes. One doctor believes the variance of waiting times is greater than originally thought.

65. What type of test should be used?

66. What is the test statistic?

67. What is the p-value?

68. What can you conclude at the 5 percent significance level?

Homework

11.1 Facts About the Chi-Square Distribution

Decide whether the following statements are true or false.

69. As the number of degrees of freedom increases, the graph of the chi-square distribution looks more and more symmetrical.

70. The standard deviation of the chi-square distribution is twice the mean.

71. The mean and the median of the chi-square distribution are the same if $df = 24$.

11.2 Goodness-of-Fit Test

For each problem, use a solution sheet to solve the hypothesis test problem. Go to Appendix E Solution Sheets for the chi-square solution sheet. Round expected frequency to two decimal places.

72. A six-sided die is rolled 120 times. Fill in the expected frequency column. Then, conduct a hypothesis test to determine if the die is fair. The data in Table 11.34 are the result of the 120 rolls.

Face Value	Frequency	Expected Frequency
1	15	
2	29	
3	16	

Table 11.34

Face Value	Frequency	Expected Frequency
4	15	
5	30	
6	15	

Table 11.34

73. The marital status distribution of the U.S. male population, ages 15 and older, is as shown in Table 11.35.

Marital Status	%	Expected Frequency
Never Married	31.3%	
Married	56.1%	
Widowed	2.5%	
Divorced/Separated	10.1%	

Table 11.35

Suppose that a random sample of 400 U.S. males, 18 to 24 years old, yielded the following frequency distribution. We are interested in whether this age group of males fits the distribution of the U.S. adult population. Calculate the frequency one would expect when surveying 400 people. Fill in Table 11.35, rounding to two decimal places.

Marital Status	Frequency
Never Married	140
Married	238
Widowed	2
Divorced/Separated	20

Table 11.36

Use the following information to answer the next two exercises. The columns in Table 11.37 contain the Race/Ethnicity of U.S. Public Schools for a recent year, the percentages for the Advanced Placement Examinee Population for that class, and the Overall Student Population. Suppose the right column contains the results of a survey of 1,000 local students from that year who took an AP exam.

Race/Ethnicity	AP Examinee Population	Overall Student Population	Survey Frequency
Asian, Asian American, or Pacific Islander	10.2%	5.4%	113
Black or African American	8.2%	14.5%	94

Table 11.37

Race/Ethnicity	AP Examinee Population	Overall Student Population	Survey Frequency
Hispanic or Latino	15.5%	15.9%	136
American Indian or Alaska Native	0.6%	1.2%	10
White	59.4%	61.6%	604
Not Reported/Other	6.1%	1.4%	43

Table 11.37

74. Perform a goodness-of-fit test to determine whether the local results follow the distribution of the U.S. overall student population based on ethnicity.

75. Perform a goodness-of-fit test to determine whether the local results follow the distribution of U.S. AP examinee population, based on ethnicity.

76. The city of South Lake Tahoe, California, has an Asian population of 1,419 out of a total population of 23,609. Suppose that a survey of 1,419 self-reported Asians in the borough of Manhattan in the New York City area yielded the data in Table 11.38. Conduct a goodness-of-fit test to determine if the self-reported subgroups of Asians in Manhattan fit that of the South Lake Tahoe area.

Race	South Lake Tahoe Frequency	Manhattan Frequency
Asian Indian	131	174
Chinese	118	557
Filipino	1,045	518
Japanese	80	54
Korean	12	29
Vietnamese	9	21
Other	24	66

Table 11.38

Use the following information to answer the next two exercises. UCLA conducted a survey of more than 263,000 college freshmen from 385 colleges in fall 2005. The results of students' expected majors by gender were reported in *The Chronicle of Higher Education (2/2/2006).* Suppose a survey of 5,000 graduating females and 5,000 graduating males was done as a follow-up last year to determine what their actual majors were. The results are shown in the tables for Exercise 11.77 and Exercise 11.78. The second column in each table does not add to 100 percent because of rounding.

77. Conduct a goodness-of-fit test to determine if the actual college majors of graduating females fit the distribution of their expected majors.

Major	Females—Expected Major	Females—Actual Major
Arts & Humanities	14%	670
Biological Sciences	8.4%	410
Business	13.1%	685
Education	13%	650
Engineering	2.6%	145
Physical Sciences	2.6%	125
Professional	18.9%	975
Social Sciences	13%	605
Technical	0.4%	15
Other	5.8%	300
Undecided	8%	420

Table 11.39

78. Conduct a goodness-of-fit test to determine if the actual college majors of graduating males fit the distribution of their expected majors.

Major	Males—Expected Major	Males—Actual Major
Arts & Humanities	11%	600
Biological Sciences	6.7%	330
Business	22.7%	1,130
Education	5.8%	305
Engineering	15.6%	800
Physical Sciences	3.6%	175
Professional	9.3%	460
Social Sciences	7.6%	370
Technical	1.8%	90

Table 11.40

Major	Males—Expected Major	Males—Actual Major
Other	8.2%	400
Undecided	6.6%	340

Table 11.40

Read the statement and decide whether it is true or false.

79. In a goodness-of-fit test, the expected values are the values we would expect if the null hypothesis were true.

80. In general, if the observed values and expected values of a goodness-of-fit test are not close together, then the test statistic can get very large and on a graph will be way out in the right tail.

81. Use a goodness-of-fit test to determine if high school principals believe that students are absent equally during the week.

82. The test to use to determine if a six-sided die is fair is a goodness-of-fit test.

83. In a goodness-of-fit test, if the *p*-value is 0.0113, in general, do not reject the null hypothesis.

84. A sample of 212 commercial businesses was surveyed for recycling one commodity; a commodity here means any one type of recyclable material such as plastic or aluminum. Table 11.41 shows the business categories in the survey, the sample size of each category, and the number of businesses in each category that recycle one commodity. Based on the study, on average half of the businesses were expected to be recycling one commodity. As a result, the last column shows the expected number of businesses in each category that recycle one commodity. At the 5 percent significance level, perform a hypothesis test to determine if the observed number of businesses that recycle one commodity follows the uniform distribution of the expected values.

Business Type	Number in Class	Observed Number that Recycle One Commodity	Expected Number that Recycle One Commodity
Office	35	19	17.5
Retail/ Wholesale	48	27	24
Food/ Restaurants	53	35	26.5
Manufacturing/ Medical	52	21	26
Hotel/Mixed	24	9	12

Table 11.41

85. Table 11.42 contains information from a survey of 499 participants classified according to their age groups. The second column shows the percentage of obese people per age class among the study participants. The last column comes from a different study at the national level that shows the corresponding percentages of obese people in the same age classes in the United States. Perform a hypothesis test at the 5 percent significance level to determine whether the survey participants are a representative sample of the USA obese population.

Age Class (years)	Obese (Percentage)	Expected USA Average (Percentage)
20–30	75	32.6
31–40	26.5	32.6
41–50	13.6	36.6
51–60	21.9	36.6
61–70	21	39.7

Table 11.42

11.3 Test of Independence

For each problem, use a solution sheet to solve the hypothesis test problem. Go to Appendix E for the chi-square solution sheet. Round expected frequency to two decimal places.

86. A recent debate about where in the U.S. skiers believe the skiing is best prompted the following survey. Test to see if the best ski area is independent of the level of the skier.

U.S. Ski Area	Beginner	Intermediate	Advanced
Tahoe	20	30	40
Utah	10	30	60
Colorado	10	40	50

Table 11.43

87. Car manufacturers are interested in whether there is a relationship between the size of car an individual drives and the number of people in the driver's family—that is, whether car size and family size are independent. To test this, suppose that 800 car owners were randomly surveyed with the results in Table 11.44. Conduct a test of independence.

Family Size	Sub & Compact	Mid-Size	Full-Size	Van & Truck
1	20	35	40	35
2	20	50	70	80
3–4	20	50	100	90
5+	20	30	70	70

Table 11.44

88. College students may be interested in whether their majors have any effect on starting salaries after graduation. Suppose that 300 recent graduates were surveyed as to their majors in college and their starting salaries after graduation. Table 11.45 shows the data. Conduct a test of independence.

Major	< $50,000	$50,000–$68,999	$69,000 +
English	5	20	5
Engineering	10	30	60
Nursing	10	15	15
Business	10	20	30
Psychology	20	30	20

Table 11.45

89. Some travel agents claim that honeymoon hotspots vary according to age of the bride. Suppose that 280 recent brides were interviewed as to where they spent their honeymoons. The information is given in Table 11.46. Conduct a test of independence.

Location	20–29	30–39	40–49	50+
Niagara Falls	15	25	25	20
Poconos	15	25	25	10
Europe	10	25	15	5
Virgin Islands	20	25	15	5

Table 11.46

90. A manager of a sports club keeps information concerning the main sport in which members participate and their ages. To test whether there is a relationship between the age of a member and his or her choice of sport, 643 members of the sports club are randomly selected. Conduct a test of independence.

Sport	18–25	26–30	31–40	41+
Racquetball	42	58	30	46
Tennis	58	76	38	65
Swimming	72	60	65	33

Table 11.47

91. A major food manufacturer is concerned that the sales for its skinny french fries have been decreasing. As a part of a feasibility study, the company conducts research into the types of fries sold across the country to determine if the type of fries sold is independent of the area of the country. The results of the study are shown in Table 11.48. Conduct a test of independence.

Type of Fries	Northeast	South	Central	West
Skinny Fries	70	50	20	25
Curly Fries	100	60	15	30
Steak Fries	20	40	10	10

Table 11.48

92. According to Dan Leonard, an independent insurance agent in the Buffalo, New York area, the following is a breakdown of the amount of life insurance purchased by males in the following age groups. He is interested in whether the age of the male and the amount of life insurance purchased are independent events. Conduct a test for independence.

Age of Males	None	< $200,000	$200,000–$400,000	$401,001–$1,000,000	$1,000,001+
20–29	40	15	40	0	5
30–39	35	5	20	20	10
40–49	20	0	30	0	30
50+	40	30	15	15	10

Table 11.49

93. Suppose that 600 thirty-year-olds were surveyed to determine whether there is a relationship between the level of education an individual has and salary. Conduct a test of independence.

Annual Salary	Not a High School Graduate	High School Graduate	College Graduate	Masters or Doctorate
< $30,000	15	25	10	5
$30,000–$40,000	20	40	70	30
$40,000–$50,000	10	20	40	55
$50,000–$60,000	5	10	20	60
$60,000+	0	5	10	150

Table 11.50

Read the statement and decide whether it is true or false.

94. The number of degrees of freedom for a test of independence is equal to the sample size minus one.

95. The test for independence uses tables of observed and expected data values.

96. The test to use when determining if the college or university a student chooses to attend is related to his or her socioeconomic status is a test for independence.

97. In a test of independence, the expected number is equal to the row total multiplied by the column total divided

by the total surveyed.

98. An ice cream maker performs a nationwide survey about favorite flavors of ice cream in different geographic areas of the United States. Based on Table 11.51, do the numbers suggest that geographic location is independent of favorite ice cream flavors? Test at the 5 percent significance level.

U.S. Region/ Flavor	Strawberry	Chocolate	Vanilla	Rocky Road	Mint Chocolate Chip	Pistachio	Row Total
West	12	21	22	19	15	8	97
Midwest	10	32	22	11	15	6	96
East	8	31	27	8	15	7	96
South	15	28	30	8	15	6	102
Column Total	45	112	101	46	60	27	391

Table 11.51

99. Table 11.52 provides results of a recent survey of the youngest online entrepreneurs whose net worth is estimated at one million dollars or more. Their ages range from 17 to 30. Each cell in the table illustrates the number of entrepreneurs who correspond to the specific age group and their net worth. Are the ages and net worth independent? Perform a test of independence at the 5 percent significance level.

Age Group/Net Worth Value (in millions of U.S. dollars)	1–5	6–24	≥25	Row Total
17–25	8	7	5	20
26–30	6	5	9	20
Column Total	14	12	14	40

Table 11.52

100. A 2013 poll in California surveyed people about a new tax. The results are presented in Table 11.53 and are classified by ethnic group and response type. Are the poll responses independent of the participants' ethnic group? Conduct a test of independence at the 5 percent significance level.

Opinion/Ethnicity	Asian American	White/Non-Hispanic	African American	Latino	Row Total
Against Tax	48	433	41	160	682
In Favor of Tax	54	234	24	147	459
No Opinion	16	43	16	19	94
Column Total	118	710	81	326	1,235

Table 11.53

11.4 Test for Homogeneity

For each word problem, use a solution sheet to solve the hypothesis test problem. Go to Appendix E Solution Sheets for the chi-square solution sheet. Round expected frequency to two decimal places.

101. A psychologist is interested in testing whether there is a difference in the distribution of personality types for business majors and social science majors. The results of the study are shown in Table 11.54. Conduct a test of homogeneity. Test at a 5 percent level of significance.

	Open	Conscientious	Extrovert	Agreeable	Neurotic
Business	41	52	46	61	58
Social Science	72	75	63	80	65

Table 11.54

102. Do men and women select different breakfasts? The breakfasts ordered by randomly selected men and women at a popular breakfast place are shown in Table 11.55. Conduct a test for homogeneity at a 5 percent level of significance.

	French Toast	Pancakes	Waffles	Omelettes
Men	47	35	28	53
Women	65	59	55	60

Table 11.55

103. A fisherman is interested in whether the distribution of fish caught in Green Valley Lake is the same as the distribution of fish caught in Echo Lake. Of the 191 randomly selected fish caught in Green Valley Lake, 105 were rainbow trout, 27 were other trout, 35 were bass, and 24 were catfish. Of the 293 randomly selected fish caught in Echo Lake, 115 were rainbow trout, 58 were other trout, 67 were bass, and 53 were catfish. Perform a test for homogeneity at a 5 percent level of significance.

104. In 2007, the United States had 1.5 million homeschooled students, according to the U.S. National Center for Education Statistics. In Table 11.56, you can see that parents decide to homeschool their children for different reasons, and some reasons are ranked by parents as more important than others. According to the survey results shown in the table, is the distribution of applicable reasons the same as the distribution of the most important reason? Provide your assessment at the 5 percent significance level. Did you expect the result you obtained?

Reasons for Homeschooling	Applicable Reason (in thousands of respondents)	Most Important Reason (in thousands of respondents)	Row Total
Concern About the Environment of Other Schools	1,321	309	1,630
Dissatisfaction with Academic Instruction at Other Schools	1,096	258	1,354
To Provide Religious or Moral Instruction	1,257	540	1,797

Table 11.56

Reasons for Homeschooling	Applicable Reason (in thousands of respondents)	Most Important Reason (in thousands of respondents)	Row Total
Child Has Special Needs, Other Than Physical or Mental	315	55	370
Nontraditional Approach to Child's Education	984	99	1,083
Other Reasons (e.g., finances, travel, family time, etc.)	485	216	701
Column Total	5,458	1,477	6,935

Table 11.56

105. When looking at energy consumption, we are often interested in detecting trends over time and how they correlate among different countries. The information in Table 11.57 shows the average energy use in units of kg of oil equivalent per capita in the United States and the joint European Union countries (EU) for the six-year period 2005 to 2010. Do the energy use values in these two areas come from the same distribution? Perform the analysis at the 5 percent significance level.

Year	European Union	United States	Row Total
2010	3,413	7,164	10,557
2009	3,302	7,057	10,359
2008	3,505	7,488	10,993
2007	3,537	7,758	11,295
2006	3,595	7,697	11,292
2005	3,613	7,847	11,460
Column Total	20,965	45,011	65,976

Table 11.57

106. The Insurance Institute for Highway Safety collects safety information about all types of cars every year and publishes a report of top safety picks among all cars, makes, and models. Table 11.58 presents the number of top safety picks in six car categories for the two years 2009 and 2013. Analyze the table data to conclude whether the distribution of cars that earned the top safety picks safety award has remained the same between 2009 and 2013. Derive your results at the 5 percent significance level.

Year/Car Type	Small	Mid-Size	Large	Small SUV	Mid-Size SUV	Large SUV	Row Total
2009	12	22	10	10	27	6	87
2013	31	30	19	11	29	4	124
Column Total	43	52	29	21	56	10	211

Table 11.58

11.5 Comparison of the Chi-Square Tests

For each word problem, use a solution sheet to solve the hypothesis test problem. Go to Appendix E Solution Sheets for the chi-square solution sheet. Round expected frequency to two decimal places.

107. Is there a difference between the distribution of community college statistics students and the distribution of university statistics students in what technology they use on their homework? Of some randomly selected community college students, 43 used a computer, 102 used a calculator with built-in statistics functions, and 65 used a table from the textbook. Of some randomly selected university students, 28 used a computer, 33 used a calculator with built-in statistics functions, and 40 used a table from the textbook. Conduct an appropriate hypothesis test using a 0.05 level of significance.

Read the statement and decide whether it is true or false.

108. If $df = 2$, the chi-square distribution has a shape that reminds us of the exponential.

11.6 Test of a Single Variance

Use the following information to answer the next 12 exercises. Suppose an airline claims that its flights are consistently on time with an average delay of at most 15 minutes. It claims that the average delay is so consistent that the variance is no more than 150 minutes. Doubting the consistency part of the claim, a disgruntled traveler calculates the delays for his next 25 flights. The average delay for those 25 flights is 22 minutes with a standard deviation of 15 minutes.

109. Is the traveler disputing the claim about the average or about the variance?

110. A sample standard deviation of 15 minutes is the same as a sample variance of _____ minutes.

111. Is this a right-tailed, left-tailed, or two-tailed test?

112. H_0: _____

113. df = _____

114. chi-square test statistic = _____

115. p-value = _____

116. Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade the p-value.

117. Let $\alpha = 0.05$
Decision: _____
Conclusion (write out in a complete sentence): _____

118. How did you know to test the variance instead of the mean?

119. If an additional test were done on the claim of the average delay, which distribution would you use?

120. If an additional test were done on the claim of the average delay, but 45 flights were surveyed, which distribution would you use?

For each word problem, use a solution sheet to solve the hypothesis test problem. Go to Appendix E Solution Sheets for the chi-square solution sheet. Round expected frequency to two decimal places.

121. A plant manager is concerned her equipment may need recalibrating. It seems that the actual weight of the 15-ounce cereal boxes it fills has been fluctuating. The standard deviation should be at most 0.5 ounces. To determine if the machine needs to be recalibrated, 84 randomly selected boxes of cereal from the next day's

production were weighed. The standard deviation of the 84 boxes was 0.54. Does the machine need to be recalibrated?

122. Consumers may be interested in whether the cost of a particular calculator varies from store to store. Based on surveying 43 stores, which yielded a sample mean of $84 and a sample standard deviation of $12, test the claim that the standard deviation is greater than $15.

123. Isabella, an accomplished Bay-to-Breakers runner, claims that the standard deviation for her time to run the 7.5 mile race is at most 3 minutes. To test her claim, Isabella looks up five of her race times. They are 55 minutes, 61 minutes, 58 minutes, 63 minutes, and 57 minutes.

124. Airline companies are interested in the consistency of the number of babies on each flight so that they have adequate safety equipment. They are also interested in the variation of the number of babies. Suppose that an airline executive believes the average number of babies on flights is six with a variance of nine at most. The airline conducts a survey. The results of the 18 flights surveyed give a sample average of 6.4 with a sample standard deviation of 3.9. Conduct a hypothesis test of the airline executive's belief.

125. The number of births per woman in China is 1.6, down from 5.91 in 1966. This fertility rate has been attributed to the law passed in 1979 restricting births to one per woman. Suppose that a group of students studied whether the standard deviation of births per woman was greater than 0.75. They asked 50 women across China the number of births they had. The results are shown in Table 11.59. Does the students' survey indicate that the standard deviation is greater than 0.75?

# of Births	Frequency
0	5
1	30
2	10
3	5

Table 11.59

126. According to an avid aquarist, the average number of fish in a 20-gallon tank is 10, with a standard deviation of two. His friend, also an aquarist, does not believe that the standard deviation is two. She counts the number of fish in 15 other 20-gallon tanks. Based on the results that follow, do you think that the standard deviation is different from two? Data: 11; 10; 9; 10; 10; 11; 11; 10; 12; 9; 7; 9; 11; 10; and 11.

127. The manager of Frenchies is concerned that patrons are not consistently receiving the same amount of French fries with each order. The chef claims that the standard deviation for a 10-ounce order of fries is at most 1.5 ounces, but the manager thinks that it may be higher. He randomly weighs 49 orders of fries, which yields a mean of 11 ounces and a standard deviation of 2 ounces.

128. You want to buy a specific computer. A sales representative of the manufacturer claims that retail stores sell this computer at an average price of $1,249 with a very narrow standard deviation of $25. You find a website that has a price comparison for the same computer at a series of stores as follows: $1,299; $1,229.99; $1,193.08; $1,279; $1,224.95; $1,229.99; $1,269.95; and $1,249. Can you argue that pricing has a larger standard deviation than claimed by the manufacturer? Use the 5 percent significance level. As a potential buyer, what would be the practical conclusion from your analysis?

129. A company packages apples by weight. One of the weight grades is Class A apples. Class A apples have a mean weight of 150 grams, and there is a maximum allowed weight tolerance of 5 percent above or below the mean for apples in the same consumer package. A batch of apples is selected to be included in a Class A apple package. Given the following apple weights of the batch, does the fruit comply with the Class A grade weight tolerance requirements? Conduct an appropriate hypothesis test.
(a) At the 5 percent significance level
(b) At the 1 percent significance level
Weights in selected apple batch (in grams): 158; 167; 149; 169; 164; 139; 154; 150; 157; 171; 152; 161; 141; 166;

and 172.

Bringing It Together: Homework

130. a. Explain why a goodness-of-fit test and a test of independence are generally right-tailed tests.
b. If you did a left-tailed test, what would you be testing?

References

11.1 Facts About the Chi-Square Distribution

Parade Magazine. (n.d.). Retrieved from https://parade.com/

Santa Clara County Public Health Department. (2011, May). *HIV/AIDS epidemiology Santa Clara County*. Retrieved from http://sccgov.iqm2.com/Citizens/FileOpen.aspx?Type=4&ID=32762

11.2 Goodness-of-Fit Test

College Board. (n.d.). Retrieved from http://www.collegeboard.com

Ma, Y., et al. (2003). Association between eating patterns and obesity in a free-living US adult population. *American Journal of Epidemiology 158*(1), 85–92.

Ogden, C. L., et al. (2012, January). *Prevalence of obesity in the United States, 2009–2010* (NCHS Data Brief No. 82). Hyattsville, MD: National Center for Health Statistics. Retrieved from http://www.cdc.gov/nchs/data/databriefs/db82.pdf

Stevens, B. J. (n.d.). *Multi-family and commercial solid waste and recycling survey*. Arlington County, VA. Retrieved from http://www.arlingtonva.us/departments/EnvironmentalServices/SW/file84429.pdf

U.S. Census Bureau. (n.d.). *Current population reports*. Retrieved from https://www.census.gov/main/www/cprs.html

U.S. Census Bureau. (n.d). Retrieved from https://www.census.gov/

11.3 Test of Independence

Harris Interactive. (n.d.). Retrieved from http://www.statisticbrain.com/favorite-flavor-of-ice-cream/

Statistics Brain. (2016, June 29). *Youngest online entrepreneurs list*. Retrieved from http://www.statisticbrain.com/youngest-online-entrepreneur-list

11.4 Test for Homogeneity

Bielick, S. (2008, December). 1.5 million homeschooled students in the United States in 2007 (NCES 2009030).

Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009030

Bielick, S. (2008, December). 1.5 million homeschooled students in the United States in 2007—supplemental tables (NCES 2009030). Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubs2009/2009030_sup.pdf

Insurance Institute for Highway Safety. (n.d.). Ratings. Retrieved from www.iihs.org/iihs/ratings

World Bank Group. (2014). *Energy use (kg of oil equivalent per capita*). Retrieved from http://data.worldbank.org/indicator/EG.USE.PCAP.KG.OE/countries

11.6 Test of a Single Variance

Apple Insider. (n.d.). Retrieved from http://appleinsider.com/mac_price_guide

World Bank. (n.d.). Retrieved from http://www.worldbank.org/

Solutions

1. mean = 25 and standard deviation = 7.0711

3. when the number of degrees of freedom is greater than 90

5. $df = 2$

7. a goodness-of-fit test

9. 3

11. 2.04

13. We decline to reject the null hypothesis. There is not enough evidence to suggest that the observed test scores are significantly different from the expected test scores.

15. H_0: the distribution of disease cases follows the ethnicities of the general population of Santa Clara County.

17. right-tailed

19. 2016.136

21. Graph: Check student's solution.
Decision: Reject the null hypothesis.
Reason for decision: *p*-value < alpha
Conclusion: The make-up of cases does not fit the ethnicities of the general population of Santa Clara County.

23. a test of independence

25. a test of independence

27. 8

29. 6.6

31. 0.0435

33.

Product-use Per Day	African American	Native Hawaiian	Latino	Japanese Americans	White	Totals
1–10	9,886	2,745	12,831	8,378	7,650	41,490
11–20	6,514	3,062	4,932	10,680	9,877	35,065
21–30	1,671	1,419	1,406	4,715	6,062	15,273
31+	759	788	800	2,305	3,970	8,622
Totals	18,830	8,014	19,969	26,078	27,559	10,0450

Table 11.60

35.

Product Use Per Day	African American	Native Hawaiian	Latino	Japanese Americans	White
1-10	7,777.57	3,310.11	8,248.02	10,771.29	11,383.01
11-20	6,573.16	2797.52	6970.76	9,103.29	9,620.27

Table 11.61

Product Use Per Day	African American	Native Hawaiian	Latino	Japanese Americans	White
21-30	2,863.02	1,218.49	3,036.20	3,965.05	4,190.23
31+	1,616.25	687.87	1,714.01	2,238.37	2,365.49

Table 11.61

37. 10,301.8

39. right-tailed

41.
a. Reject the null hypothesis.
b. *p*-value < alpha
c. There is sufficient evidence to conclude that product use is dependent on ethnic group.

43. test for homogeneity

45. test for homogeneity

47. All values in the table must be greater than or equal to five.

49. 3

51. 0.00005

53. a goodness-of-fit test

55. a test for independence

57. Answers will vary. Sample answer: Tests of independence and tests for homogeneity both calculate the test statistic the same way $\sum_{(i,j)} \frac{(O-E)^2}{E}$. In addition, all values must be greater than or equal to five.

59. a test of a single variance

61. a left-tailed test

63. H_0: $\sigma^2 = 0.81^2$;
H_a: $\sigma^2 > 0.81^2$

65. a test of a single variance

67. 0.0542

69. true

71. false

73.

Marital Status	%	Expected Frequency
Never Married	31.3%	125.2
Married	56.1%	224.4
Widowed	2.5%	10
Divorced/Separated	10.1%	40.4

Table 11.62

a. The data fit the distribution.
b. The data do not fit the distribution.
c. 3
d. chi-square distribution with $df = 3$
e. 19.27
f. 0.0002
g. Check student's solution.
h. i. Alpha = 0.05
   ii. Decision: Reject null hypothesis.
   iii. Reason for decision: p-value < alpha
   iv. Conclusion: Data do not fit the distribution.

75. a. H_0: The local results follow the distribution of the U.S. AP examinee population.
b. H_a: The local results do not follow the distribution of the U.S. AP examinee population.
c. $df = 5$
d. chi-square distribution with $df = 5$
e. chi-square test statistic = 13.4
f. p-value = 0.0199
g. Check student's solution.
h. i. Alpha = 0.05
   ii. Decision: Reject null when $a = 0.05$.
   iii. Reason for decision: p-value < alpha
   iv. Conclusion: Local data do not fit the AP examinee distribution.
   v. Decision: Do not reject null when $a = 0.01$
   vi. Conclusion: There is insufficient evidence to conclude that local data do not follow the distribution of the U.S. AP examinee distribution.

77. a. H_0: The actual college majors of graduating females fit the distribution of their expected majors.
b. H_a: The actual college majors of graduating females do not fit the distribution of their expected majors.
c. $df = 10$
d. chi-square distribution with $df = 10$
e. test statistic = 11.48
f. p-value = 0.3211
g. Check student's solution.
h. i. Alpha = 0.05
   ii. Decision: Do not reject null hypothesis when $a = 0.05$ and $a = 0.01$.
   iii. Reason for decision: p-value > alpha
   iv. Conclusion: There is insufficient evidence to conclude that the distribution of actual college majors of graduating females do not fit the distribution of their expected majors.

79. true

81. true

83. false

85. The hypotheses for the goodness-of-fit test are:
- H_0: Surveyed obese fit the distribution of expected obese.
- H_a: Surveyed obese do not fit the distribution of expected obese.

Use a chi-square distribution with $df = 4$ to evaluate the data.
- The test statistic is $\chi^2 = 9.85$
- The p-value = 0.0431
- At the 5% significance level, $\alpha = 0.05$. For this data, $p < \alpha$.
- At the 5% level of significance, from the data, there is sufficient evidence to conclude that the surveyed obese do not fit the distribution of expected obese.

87. a. H_0: Car size is independent of family size.
b. H_a: Car size is dependent on family size.
c. $df = 9$
d. chi-square distribution with $df = 9$
e. test statistic = 15.8284
f. p-value = 0.0706
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: p-value > alpha
iv. Conclusion: At the 5 percent significance level, there is insufficient evidence to conclude that car size and family size are dependent.

89. a. H_0: Honeymoon locations are independent of bride's age.
b. H_a: Honeymoon locations are dependent on bride's age.
c. $df = 9$
d. chi-square distribution with $df = 9$
e. test statistic = 15.7027
f. p-value = 0.0734
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: p-value > alpha
iv. Conclusion: At the 5 percent significance level, there is insufficient evidence to conclude that honeymoon location and bride age are dependent.

91. a. H_0: The types of fries sold are independent of the location.
b. H_a: The types of fries sold are dependent on the location.
c. $df = 6$
d. chi-square distribution with $df = 6$
e. test statistic =18.8369
f. p-value = 0.0044
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Reject the null hypothesis.
iii. Reason for decision: p-value < alpha
iv. Conclusion: At the 5 percent significance level, there is sufficient evidence that types of fries and location are dependent.

93. a. H_0: Salary is independent of level of education.
b. H_a: Salary is dependent on level of education.
c. $df = 12$
d. chi-square distribution with $df = 12$

e. test statistic = 255.7704
f. *p*-value = 0
g. Check student's solution.
h. Alpha: 0.05

Decision: Reject the null hypothesis.

Reason for decision: *p*-value < alpha

Conclusion: At the 5 percent significance level, there is sufficient evidence to conclude that salary and level of education are dependent.

95. true

97. true

99. a. H_0: Age is independent of the youngest online entrepreneurs' net worth.
b. H_a: Age is dependent on the net worth of the youngest online entrepreneurs.
c. $df = 2$
d. chi-square distribution with $df = 2$
e. test statistic = 1.76
f. *p*-value = 0.4144
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: *p*-value > alpha
iv. Conclusion: At the 5 percent significance level, there is insufficient evidence to conclude that age and net worth for the youngest online entrepreneurs are dependent.

101. a. H_0: The distribution for personality types is the same for both majors.
b. H_a: The distribution for personality types is not the same for both majors.
c. $df = 4$
d. chi-square with $df = 4$
e. test statistic = 3.01
f. *p*-value = 0.5568
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: *p*-value > alpha
iv. Conclusion: There is insufficient evidence to conclude that the distribution of personality types is different for business and social science majors.

103. a. H_0: The distribution for fish caught is the same in Green Valley Lake and in Echo Lake.
b. H_a: The distribution for fish caught is not the same in Green Valley Lake and in Echo Lake.
c. 3
d. chi-square with $df = 3$
e. 11.75
f. *p*-value = 0.0083
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Reject the null hypothesis.
iii. Reason for decision: *p*-value < alpha
iv. Conclusion: There is evidence to conclude that the distribution of fish caught is different in Green Valley Lake and in Echo Lake.

105. a. H_0: The distribution of average energy use in the United States is the same as in Europe between 2005 and 2010.

b. H_a: The distribution of average energy use in the United States is not the same as in Europe between 2005 and 2010.
c. $df = 4$
d. chi-square with $df = 4$
e. test statistic = 2.7434
f. p-value = 0.7395
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: p-value > alpha
iv. Conclusion: At the 5 percent significance level, there is insufficient evidence to conclude that the average energy use values in the United States and EU are not derived from different distributions for the period from 2005 to 2010.

107. a. H_0: The distribution for technology use is the same for community college students and university students.
b. H_a: The distribution for technology use is not the same for community college students and university students.
c. 2
d. chi-square with $df = 2$
e. 7.05
f. p value = 0.0294
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Reject the null hypothesis.
iii. Reason for decision: p value < alpha
iv. Conclusion: There is sufficient evidence to conclude that the distribution of technology use for statistics homework is not the same for statistics students at community colleges and at universities.

110. 225

112. H_0: $\sigma^2 \leq 150$

114. 36

116. Check student's solution.

118. The claim is that the variance is no more than 150 minutes.

120. a student's t or normal distribution

122. a. H_0: $\sigma = 15$
b. H_a: $\sigma > 15$
c. $df = 42$
d. chi-square with $df = 42$
e. test statistic = 26.88
f. p-value = 0.9663
g. Check student's solution.
h. i. Alpha = 0.05
ii. Decision: Do not reject null hypothesis.
iii. Reason for decision: p-value > alpha
iv. Conclusion: There is insufficient evidence to conclude that the standard deviation is greater than 15.

124. a. H_0: $\sigma \leq 3$
b. H_a: $\sigma > 3$
c. $df = 17$
d. chi-square distribution with $df = 17$

e. test statistic = 28.73
f. *p*-value = 0.0371
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Reject the null hypothesis.
iii. Reason for decision: *p*-value < alpha
iv. Conclusion: There is sufficient evidence to conclude that the standard deviation is greater than three.

126. a. H_0: $\sigma = 2$
b. H_a: $\sigma \neq 2$
c. $df = 14$
d. chi-square distiribution with $df = 14$
e. chi-square test statistic = 5.2094
f. *p*-value = 0.0346
g. Check student's solution.
h. i. Alpha = 0.05
ii. Decision: Reject the null hypothesis
iii. Reason for decision: *p*-value < alpha
iv. Conclusion: There is sufficient evidence to conclude that the standard deviation is different than two.

128. The sample standard deviation is $34.29.
$H_0: \sigma^2 = 25^2$
$H_a: \sigma^2 > 25^2$

$df = n - 1 = 7$

Test statistic: $x^2 = x_7^2 = \frac{(n-1)s^2}{25^2} = \frac{(8-1)(34.29)^2}{25^2} = 13.169$;

p-value: $P\left(x_7^2 > 13.169\right) = 1 - P\left(x_7^2 \leq 13.169\right) = .0681$

Alpha: 0.05

Decision: Do not reject the null hypothesis.

Reason for decision: *p*-value > alpha

Conclusion: At the 5 percent level, there is insufficient evidence to conclude that the variance is more than 625.

130. a. The test statistic is always positive and if the expected and observed values are not close together, the test statistic is large and the null hypothesis will be rejected.
b. Testing to see if the data fits the distribution too well or is too perfect.

12

LINEAR REGRESSION AND CORRELATION

Figure 12.1 Linear regression and correlation can help you determine whether an auto mechanic's salary is related to his work experience. (credit: Joshua Rothhaas)

Chapter Objectives

By the end of this chapter, the student should be able to do the following:

- Discuss basic ideas of linear regression and correlation
- Create and interpret a line of best fit
- Calculate and interpret the correlation coefficient
- Calculate and interpret outliers

Introduction

Professionals often want to know how two or more numeric variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is the relationship, and how strong is it?

In another example, your income may be determined by your education, your profession, your years of experience, and your ability. The amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee.

The type of data described in the examples is **bivariate** data—*bi*—for two variables. In reality, statisticians use **multivariate** data, meaning many variables.

In this chapter, you will study the simplest form of regression—*linear regression*—with one independent variable (x). This involves data that fit a line in two dimensions. You will also study correlation, which measures the strength of a relationship.

12.1 Linear Equations

Linear regression for two variables is based on a linear equation with one independent variable. The equation has the form

$$y = a + bx$$

where a and b are constant numbers.

The variable x is the *independent variable*; y is the *dependent variable*. Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

EXAMPLE 12.1

The following examples are linear equations.

$$y = 3 + 2x$$
$$y = -0.01 + 1.2x$$

TRY IT 12.1

Is the following an example of a linear equation?

$y = -0.125 - 3.5x$

The graph of a linear equation of the form $y = a + bx$ is a straight line. Any line that is not vertical can be described by this equation.

EXAMPLE 12.2

Graph the equation $y = -1 + 2x$.

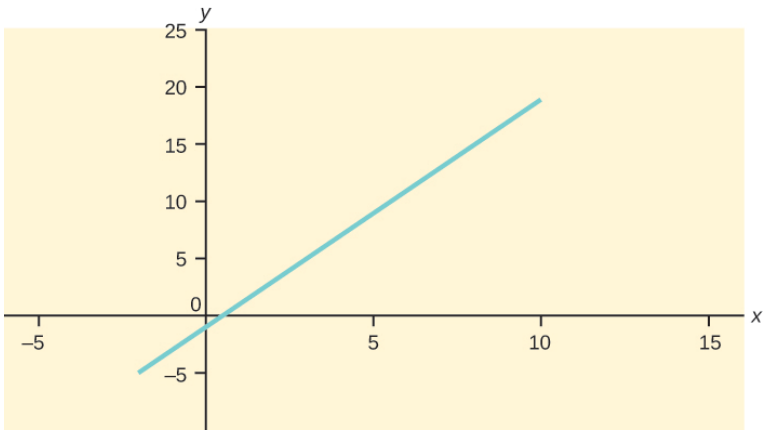


Figure 12.2

TRY IT 12.2

Is the following an example of a linear equation? Why or why not?

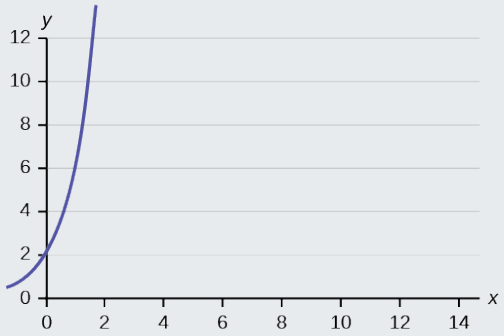


Figure 12.3

EXAMPLE 12.3

Aaron's Word Processing Service does word processing. The rate for services is $32 per hour plus a $31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to complete the job.

Problem

Find the equation that expresses the total cost in terms of the number of hours required to complete the job.

Solution

Let x = the number of hours it takes to get the job done.
Let y = the total cost to the customer.

The $31.50 is a fixed cost. If it takes x hours to complete the job, then (32)(x) is the cost of the word processing only. The total cost is $y = 31.50 + 32x$.

TRY IT 12.3

Emma's Extreme Sports hires hang-gliding instructors and pays them a fee of $50 per class, as well as $20 per student in the class. The total cost Emma pays depends on the number of students in a class. Find the equation that expresses the total cost in terms of the number of students in a class.

Slope and *y*-interceptof a Linear Equation

For the linear equation $y = a + bx$, b = slope and a = y-inttercept. From algebra, recall that the slope is a number that describes the steepness of a line; the y-intercept is the y-coordinate of the point (0, a), where the line crosses the y-axis.

Please note that in previous courses you learned $y = mx + b$ was the slope-intercept form of the equation, where m represented the slope and b represented the y-intercept. In this text, the form $y = a + bx$ is used, where a is the y-intercept and b is the slope. The key is remembering the coefficient of x is the slope, and the constant number is the y-intercept.

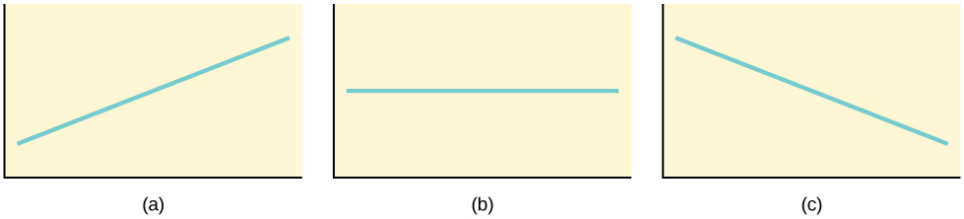


Figure 12.4 Three possible graphs of $y = a + bx$. (a) If $b > 0$, the line slopes upward to the right. (b) If $b = 0$, the line is horizontal. (c) If $b < 0$, the line slopes downward to the right.

EXAMPLE 12.4

Svetlana tutors to make extra money for college. For each tutoring session, she charges a one-time fee of $25 plus $15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is $y = 25 + 15x$.

Problem

What are the independent and dependent variables? What is the y-intercept, and what is the slope? Interpret them using complete sentences.

Solution

The independent variable (x) is the number of hours Svetlana tutors each session. The dependent variable (y) is the amount, in dollars, Svetlana earns for each session.

The y-intercept is 25 ($a = 25$). At the start of the tutoring session, Svetlana charges a one-time fee of $25 (this is when $x = 0$). The slope is 15 ($b = 15$). For each session, Svetlana earns $15 for each hour she tutors.

TRY IT 12.4

Ethan repairs household appliances such as dishwashers and refrigerators. For each visit, he charges $25 plus $20 per hour of work. A linear equation that expresses the total amount of money Ethan earns per visit is $y = 25 + 20x$.

What are the independent and dependent variables? What is the y-intercept, and what is the slope? Interpret them using complete sentences.

12.2 The Regression Equation

Data rarely fit a straight line exactly. Usually, you must be satisfied with rough predictions. Typically, you have a set of data with a scatter plot that appear to *fit* a straight line. This is called a *line of best fit or least-squares regression line*.

COLLABORATIVE EXERCISE

If you know a person's pinky (smallest) finger length, do you think you could predict that person's height? Collect data from your class (pinky finger length, in inches). The independent variable, x, is pinky finger length and the dependent variable, y, is height. For each set of data, plot the points on graph paper. Make your graph big enough and *use a ruler*. Then, *by eye*, draw a line that appears to *fit* the data. For your line, pick two convenient points and use them to find the slope of the line. Find the y-intercept of the line by extending your line so it crosses the y-axis. Using the slopes and the y-intercepts, write your equation of *best fit*. Do you think everyone will have the same equation? Why or why not? According to your equation, what is the predicted height for a pinky length of 2.5 inches?

EXAMPLE 12.5

A random sample of 11 statistics students produced the data in Table 12.1, where x is the third exam score out of 80 and y is the final exam score out of 200. Can you predict the final exam score of a random student if you know the third exam score?

x (third exam score)	y (final exam score)
65	175
67	133

Table 12.1

x (third exam score)	*y* (final exam score)
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

Table 12.1

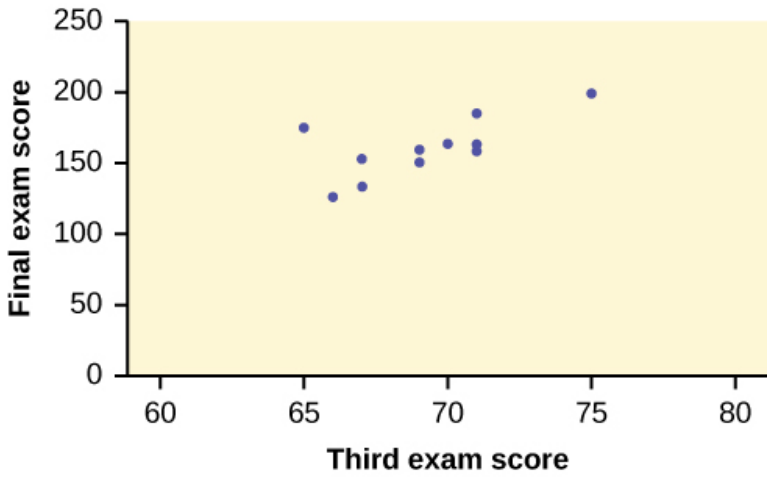


Figure 12.5 Using the *x*- and *y*-coordinates in the table, we plot the points on a graph to create the scatter plot showing the scores on the final exam based on scores from the third exam.

TRY IT 12.5

SCUBA divers have maximum dive times they cannot exceed when going to different depths. The data in Table 12.2 show different depths in feet, with the maximum dive times in minutes. Use your calculator to find the least squares regression line and predict the maximum dive time for 110 feet.

x (depth)	*y* (maximum dive time)
50	80
60	55
70	45
80	35
90	25
100	22

Table 12.2

The third exam score, *x*, is the independent variable, and the final exam score, *y*, is the dependent variable. We will plot a regression line that best *fits* the data. If each of you were to fit a line *by eye*, you would draw different lines. We can obtain a line of best fit using either the median-–median line approach or by calculating the least-squares regression line.

Let''s first find the line of best fit for the relationship between the third exam score and the final exam score using the median-median line approach. Remember that this is the data from Example 12.5 after the ordered pairs have been listed by ordering *x* values. If multiple data points have the same *x* values, then they are listed in order from least to greatest *x* (see data values where $x = 71$). We first divide our scores into three groups of approximately equal numbers of *x* values per group. The first and third groups have the same number of *x* values. We must remember first to put the *x* values in ascending order. The corresponding *y* values are then recorded. However, to find the median, we first must rearrange the *y* values in each group from the least value to the greatest value. Table 12.3 shows the correct ordering of the *x* values but does not show a reordering of the *y* values.

x (third exam score)	*y* (final exam score)
65	175
66	126
67	133
67	153
69	151
69	159
70	163
71	159
71	163

Table 12.3

x (third exam score)	*y* (final exam score)
71	185
75	198

Table 12.3

With this set of data, the first and last groups each have four *x* values and four corresponding *y* values. The second group has three *x* values and three corresponding *y* values. We need to organize the *x* and *y* values per group and find the median *x* and *y* values for each group. Let's now write out our *y* values for each group in ascending order. For group 1, the *y* values in order are 126, 133, 153, and 175. For group 2, the *y* values are already in order. For group 3, the *y* values are also already in order. We can represent these data as shown in Table 12.4, but notice that we have *broken* the ordered pairs; (65, 126) is not a data point in our original set:

Group	*x* (third exam score)	*y* (final exam score)	Median *x* value	Median *y* value
1	65 66 67 67	126 133 153 175	66.5	143
2	69 69 70	151 159 163	69	159
3	71 71 71 75	159 163 185 198	71	174

Table 12.4

When this is completed, we can write the ordered pairs for the median values. This allows us to find the slope and *y*-intercept of the –median-median line.

The ordered pairs are (66.5, 143), (69, 159), and (71, 174).

The slope can be calculated using the formula $m - \frac{y_2 - y_1}{x_2 - x_1}$. Substituting the median *x* and *y* values from the first and third groups gives $m = \frac{174-143}{71-66.5}$, which simplifies to $m \approx 6.9$.

The *y*-intercept may be found using the formula $b = \frac{\Sigma y - m\Sigma x}{3}$, which means the quantity of the sum of the median *y* values minus the slope times the sum of the median *x* values divided by three.

The sum of the median *x* values is 206.5, and the sum of the median *y* values is 476. Substituting these sums and the slope into the formula gives $b = \frac{476-6.9(206.5)}{3}$, which simplifies to $b \approx -316.3$.

The line of best fit is represented as $y = mx + b$.

Thus, the equation can be written as $y = 6.9x - 316.3$.

The median–median line may also be found using your graphing calculator. You can enter the *x* and *y* values into two separate lists; choose Stat, Calc, Med-Med, and press Enter. The slope, *a*, and *y*-intercept, *b*, will be provided. The calculator shows a slight deviation from the previous manual calculation as a result of rounding. Rounding to the nearest tenth, the calculator gives the –median-median line of $y = 6.9x - 315.5$. Each point of data is of the the form (*x*, *y*), and each point of the line of best fit using least-squares linear regression has the form (x, $\hat{y}$).

The $\hat{y}$ is read *y hat* and is the *estimated value of y*. It is the value of *y* obtained using the regression line. It is not generally equal to *y* from data, but it is still important because it can help make predictions for other values.

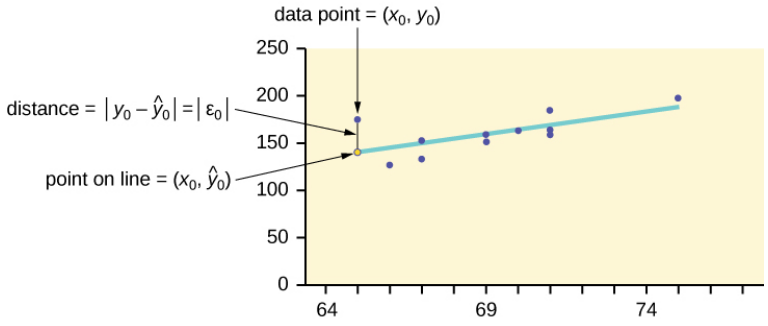


Figure 12.6

The term $y_0 - \hat{y}_0 = \varepsilon_0$ is called the *error* or *residual*. It is not an error in the sense of a mistake. The *absolute value of a residual* measures the vertical distance between the actual value of y and the estimated value of y. In other words, it measures the vertical distance between the actual data point and the predicted point on the line, or it measures how far the estimate is from the actual data value.

If the observed data point lies above the line, the residual is positive and the line underestimates the actual data value for y. If the observed data point lies below the line, the residual is negative and the line overestimates that actual data value for y.

In Figure 12.6, $y_0 - \hat{y}_0 = \varepsilon_0$ is the residual for the point shown. Here the point lies above the line and the residual is positive.

ε = the Greek letter epsilon

For each data point, you can calculate the residuals or errors, $y_i - \hat{y}_i = \varepsilon_i$ for $i = 1, 2, 3, \ldots, 11$.

Each $|\varepsilon|$ is a vertical distance.

For the example about the third exam scores and the final exam scores for the 11 statistics students, there are 11 data points. Therefore, there are 11 ε values. If you square each ε and add them, you get the sum of ε squared from $i = 1$ to $i = 11$, as shown below.

$$(\varepsilon_1)^2 + (\varepsilon_2)^2 + \ldots + (\varepsilon_{11})^2 = \sum_{i=1}^{11} \varepsilon^2.$$

This is called the **sum of squared errors (SSE)**.

Using calculus, you can determine the values of a and b that make the SSE a minimum. When you make the SSE a minimum, you have determined the points that are on the line of best fit. It turns out that the line of best fit has the equation

$$\hat{y} = a + bx$$

where

$a = \overline{y} - b\overline{x}$

and $b = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sum(x-\overline{x})^2}$.

The sample means of the x values and the y values are $\overline{x}$ and $\overline{y}$, respectively. The best-fit line always passes through the point $(\overline{x}, \overline{y})$.

The slope (b) can be written as $b = r\left(\frac{s_y}{s_x}\right)$ where s_y = the standard deviation of the y values and s_x = the standard deviation of the x values. r is the correlation coefficient, which shows the relationship between the x and y values. This will be discussed in more detail in the next section.

Least-Squares Criteria for Best Fit

The process of fitting the best-fit line is called **linear regression**. We assume that the data are scattered about a straight line. To find that line, we minimize the sum of the squared errors (SSE), or make it as small as possible. Any other line you might choose would have a higher SSE than the best-fit line. This best-fit line is called the *least-squares regression line*.

Note
Computer spreadsheets, statistical software, and many calculators can quickly calculate the best-fit line and create the graphs. The calculations tend to be tedious if done by hand. Instructions to use the TI-83, TI-83+, and TI-84+ calculators to find the best-fit line and create a scatter plot are shown at the end of this section.

Third Exam vs. Final Exam Example

The graph of the line of best fit for the third exam/final exam example is as follows:

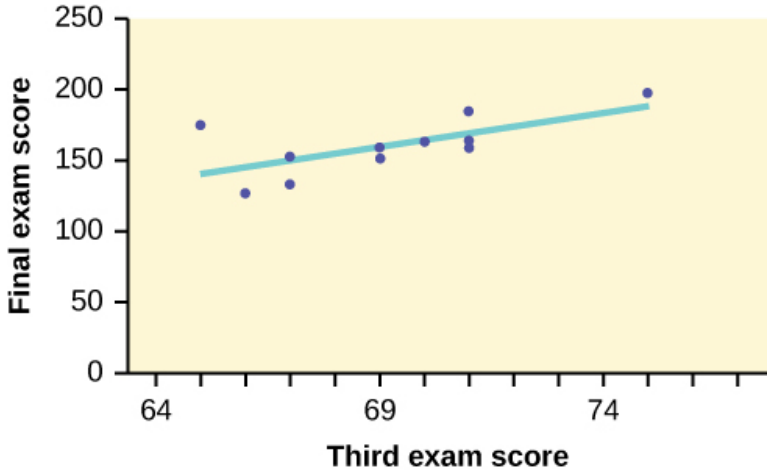


Figure 12.7

The least-squares regression line (best-fit line) for the third exam/final exam example has the equation

$$\hat{y} = -173.51 + 4.83x.$$

Understanding and Interpreting the *y*-intercept

The *y*-intercept, *a*, of the line describes where the plot line crosses the *y*-axis. The *y*-intercept of the best-fit line tells us the best value of the relationship when *x* is zero. In some cases, it does not make sense to figure out what *y* is when *x* = 0. For example, in the third exam vs. final exam example, the *y*-intercept occurs when the third exam score, or *x*, is zero. Since all the scores are grouped around a passing grade, there is no need to figure out what the final exam score, or *y*, would be when the third exam was zero.

However, the *y*-intercept is very useful in many cases. For many examples in science, the *y*-intercept gives the baseline reading when the experimental conditions aren''t applied to an experimental system. This baseline indicates how much the experimental condition affects the system. It could also be used to ensure that equipment and measurements are calibrated properly before starting the experiment.

In biology, the concentration of proteins in a sample can be measured using a chemical assay that changes color depending on how much protein is present. The more protein present, the darker the color. The amount of color can be measured by the absorbance reading. Table 12.5 shows the expected absorbance readings at different protein concentrations. This is called a *standard curve* for the assay.

Concentration (mM)	Absorbance (mAU)
125	0.021
250	0.023
500	0.068
750	0.086
1,000	0.105

Table 12.5

Concentration (mM)	Absorbance (mAU)
1,500	0.124
2,000	0.146

Table 12.5

The scatter plot Figure 12.8 includes the line of best fit.

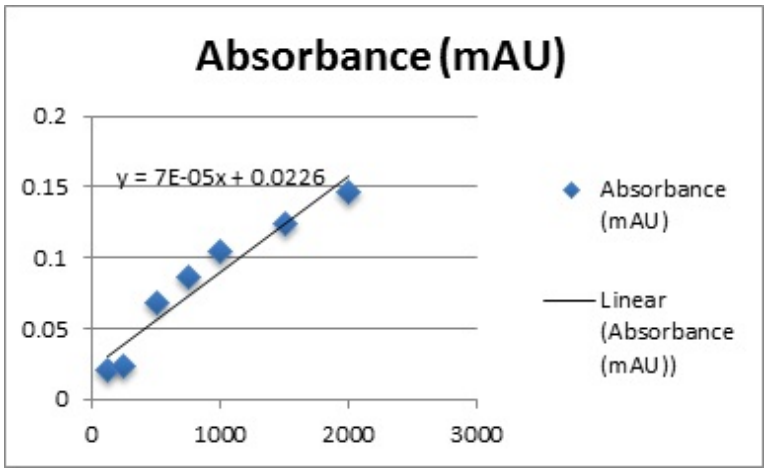


Figure 12.8

The *y*-intercept of this line occurs at 0.0226 mAU. This means the assay gives a reading of 0.0226 mAU when there is no protein present. That is, it is the baseline reading that can be attributed to something else, which, in this case, is some other non-protein chemicals that are absorbing light. We can tell that this line of best fit is reasonable because the *y*-intercept is small, close to zero. When there is no protein present in the sample, we expect the absorbance to be very small, or close to zero, as well.

Understanding Slope

The slope of the line, *b*, describes how changes in the variables are related. It is important to interpret the slope of the line in the context of the situation represented by the data. You should be able to write a sentence interpreting the slope in plain English.

Interpretation of the Slope: The slope of the best-fit line tells us how the dependent variable (*y*) changes for every one unit increase in the independent (*x*) variable, on average.

Third Exam vs. Final Exam Example
Slope: The slope of the line is $b = 4.83$.
Interpretation: For a 1-point increase in the score on the third exam, the final exam score increases by 4.83 points, on average.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Using the Linear Regression T Test: `LinRegTTest`

1. In the STAT list editor, enter the *x* data in list `L1` and the *y* data in list `L2`, paired so that the corresponding (*x*, *y*) values are next to each other in the lists. (If a particular pair of values is repeated, enter it as many times as it appears in the data.)
2. On the STAT TESTS menu, scroll down and select `LinRegTTest`. (Be careful to select `LinRegTTest`. Some calculators may also have a different item called `LinRegTInt`.)
3. On the `LinRegTTest` input screen, enter `Xlist: L1`, `Ylist: L2`, and `Freq: 1`.
4. On the next line, at the prompt β or ρ, highlight $\neq$ `0` and press ENTER.

5. Leave the line for RegEQ: blank.
6. Highlight Calculate and press ENTER.

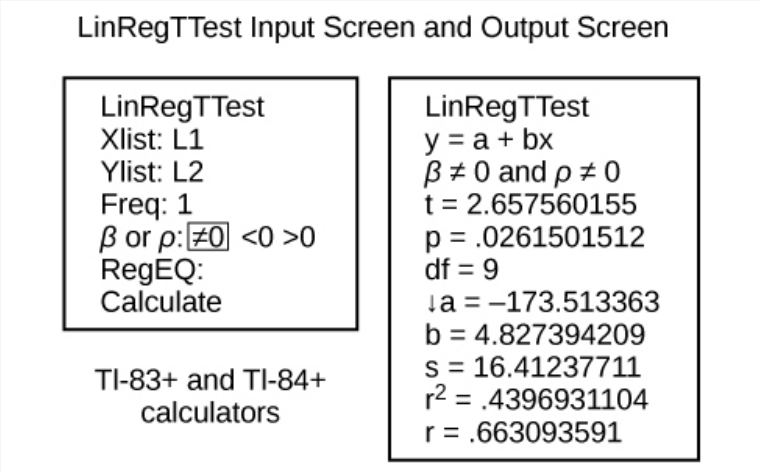


Figure 12.9

The output screen contains a lot of information. For now, let's focus on a few items from the output and return to the other items later.
The second line says $y = a + bx$. Scroll down to find the values $a = -173.513$ and $b = 4.8273$.

The equation of the best-fit line is $\hat{y} = -173.51 + 4.83x$.
The two items at the bottom are $r^2 = .43969$ and $r = .663$. For now, just note where to find these values; we examine them in the next two sections.

Graphing the Scatter Plot and Regression Line

1. We are assuming the x data are already entered in list L1 and the y data are in list L2.
2. Press 2nd STATPLOT ENTER to use Plot 1.
3. On the input screen for PLOT 1, highlight On, and press ENTER.
4. For TYPE, highlight the first icon, which is the scatter plot, and press ENTER.
5. Indicate Xlist: L1 and Ylist: L2.
6. For Mark, it does not matter which symbol you highlight.
7. Press the ZOOM key and then the number 9 (for menu item ZoomStat); the calculator fits the window to the data.
8. To graph the best-fit line, press the Y= key and type the equation –173.5 + 4.83X into equation Y1. (The X key is immediately left of the STAT key.) Press ZOOM 9 again to graph it.
9. Optional: If you want to change the viewing window, press the WINDOW key. Enter your desired window using Xmin, Xmax, Ymin, and Ymax.

NOTE

Another way to graph the line after you create a scatter plot is to use LinRegTTest.

1. Make sure you have done the scatter plot. Check it on your screen.
2. Go to LinRegTTest and enter the lists.
3. At RegEq, press VARS and arrow over to Y-VARS. Press 1 for 1:Function. Press 1 for 1:Y1. Then, arrow down to Calculate and do the calculation for the line of best fit.
4. Press Y= (you will see the regression equation).
5. Press GRAPH, and the line will be drawn.

The Correlation Coefficient *r*

Besides looking at the scatter plot and seeing that a line seems reasonable, how can you determine whether the line is a good predictor? Use the correlation coefficient as another indicator (besides the scatter plot) of the strength of the relationship between *x* and *y*.

The **correlation coefficient, *r*,** developed by Karl Pearson during the early 1900s, is numeric and provides a measure of the strength and direction of the linear association between the independent variable *x* and the dependent variable *y*.

If you suspect a linear relationship between *x* and *y*, then *r* can measure the strength of the linear relationship.

What the Value of *r* Tells Us

- The value of *r* is always between –1 and +1. In other words, $-1 \leq r \leq 1$.
- The size of the correlation *r* indicates the strength of the linear relationship between *x* and *y*. Values of *r* close to –1 or to +1 indicate a stronger linear relationship between *x* and *y*.
- If $r = 0$, there is absolutely no linear relationship between *x* and *y* (no linear correlation).
- If $r = 1$, there is perfect positive correlation. If $r = -1$, there is perfect negative correlation. In both these cases, all the original data points lie on a straight line. Of course, in the real world, this does not generally happen.

What the Sign of *r* Tells Us

- A positive value of *r* means that when *x* increases, *y* tends to increase and when *x* decreases, *y* tends to decrease (positive correlation).
- A negative value of *r* means that when *x* increases, *y* tends to decrease and when *x* decreases, *y* tends to increase (negative correlation).
- The sign of *r* is the same as the sign of the slope, *b*, of the best-fit line.

Note

A strong correlation does not suggest that *x* causes *y* or *y* causes *x*. We say correlation does not imply causation.

The correlation coefficient is calculated as the quantity of data points times the sum of the quantity of the *x*-coordinates times the *y*-coordinates, minus the quantity of the sum of the *x*-coordinates times the sum of the *y*-coordinates, all divided by the square root of the quantity of data points times the sum of the *x*-coordinates squared minus the square of the sum of the *x*-coordinates, times the number of data points times the sum of the *y*-coordinates squared minus the square of the sum of the *y*-coordinates. It can be summarized by the following equation:

$$r = \frac{n\Sigma(xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

where *n* is the number of data points.

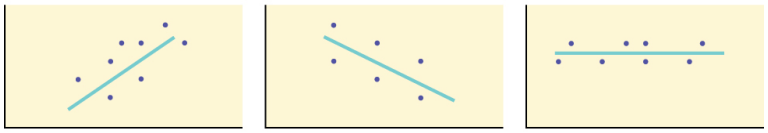


Figure 12.10 (a) A scatter plot showing data with a positive correlation: $0 < r < 1$. (b) A scatter plot showing data with a negative correlation: $-1 < r < 0$. (c) A scatter plot showing data with zero correlation: $r = 0$.

The formula for *r* looks formidable. However, computer spreadsheets, statistical software, and many calculators can calculate *r* quickly. The correlation coefficient, *r*, is the bottom item in the output screens for the LinRegTTest on the TI-83, TI-83+, or TI-84+ calculator (see previous section for instructions).

The Coefficient of Determination

The variable r^2 is called the **coefficient of determination** and it is the square of the correlation coefficient, but it is usually stated as a percentage, rather than in decimal form. It has an interpretation in the context of the data:

- r^2, when expressed as a percent, represents the percentage of variation in the dependent (predicted) variable *y* that can be explained by variation in the independent (explanatory) variable *x* using the regression (best-fit) line.
- $1 - r^2$, when expressed as a percentage, represents the percentage of variation in *y* that is *not* explained by

variation in x using the regression line. This can be seen as the scattering of the observed data points about the regression line.

Consider the third exam/final exam example introduced in the previous section.

- The line of best fit is: $\hat{y} = -173.51 + 4.83x$.
- The correlation coefficient is $r = .6631$.
- The coefficient of determination is $r^2 = .6631^2 = .4397$.

Interpret r^2 in the context of this example.

- Approximately 44 percent of the variation (0.4397 is approximately 0.44) in the final exam grades can be explained by the variation in the grades on the third exam, using the best-fit regression line.
- Therefore, the rest of the variation (1 – 0.44 = 0.56 or 56 percent) in the final exam grades cannot be explained by the variation of the grades on the third exam with the best-fit regression line. These are the variation of the points that are not as close to the regression line as others.

12.3 Testing the Significance of the Correlation Coefficient (Optional)

The correlation coefficient, r, tells us about the strength and direction of the linear relationship between x and y. However, the reliability of the linear model also depends on how many observed data points are in the sample. We need to look at both the correlation coefficient r and the sample size n, together.

We perform a hypothesis test of the *significance of the correlation coefficient* to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.

The sample data are used to compute r, the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient. But, because we have only sample data, we cannot calculate the population correlation coefficient. The sample correlation coefficient, r, is our estimate of the unknown population correlation coefficient.

The symbol for the population correlation coefficient is ρ, the Greek letter rho.
ρ = population correlation coefficient (unknown).
r = sample correlation coefficient (known; calculated from sample data).

The hypothesis test lets us decide whether the value of the population correlation coefficient ρ is *close to zero* or *significantly different from zero*. We decide this based on the sample correlation coefficient r and the sample size n.

If the test concludes the correlation coefficient is significantly different from zero, we say the correlation coefficient is *significant*.

- Conclusion: There is sufficient evidence to conclude there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero.
- What the conclusion means: There is a significant linear relationship between x and y. We can use the regression line to model the linear relationship between x and y in the population.

If the test concludes the correlation coefficient is not significantly different from zero (it is close to zero), we say the correlation coefficient is *not significant*.

- Conclusion: There is insufficient evidence to conclude there is a significant linear relationship between x and y because the correlation coefficient is not significantly different from zero.
- What the conclusion means: There is not a significant linear relationship between x and y. Therefore, we *cannot* use the regression line to model a linear relationship between x and y in the population.

Note

- If r is significant and the scatter plot shows a linear trend, the line can be used to predict the value of y for values of x that are within the domain of observed x values.
- If r is not significant *or* if the scatter plot does not show a linear trend, the line should not be used for prediction.
- If r is significant and the scatter plot shows a linear trend, the line may *not* be appropriate or reliable for prediction *outside* the domain of observed x values in the data.

Performing the Hypothesis Test

- Null hypothesis: H_0: $\rho = 0$.
- Alternate hypothesis: H_a: $\rho \neq 0$.

What the Hypothesis Means in Words:

- **Null hypothesis H_0:** The population correlation coefficient *is not* significantly different from zero. There *is not* a significant linear relationship (correlation) between x and y in the population.
- **Alternate hypothesis H_a:** The population correlation coefficient *is* significantly different from zero. There *is* a significant linear relationship (correlation) between x and y in the population.

Drawing a Conclusion:
There are two methods to make a conclusion. The two methods are equivalent and give the same result.

- **Method 1: Use the *p*-value.**
- **Method 2: Use a table of critical values.**

In this chapter, we will always use a significance level of 5 percent, $\alpha = 0.05$.

Note

Using the *p*-value method, you could choose any appropriate significance level you want; you are not limited to using $\alpha = 0.05$. But, the table of critical values provided in this textbook assumes we are using a significance level of 5 percent, $\alpha = 0.05$. If we wanted to use a significance level different from 5 percent with the critical value method, we would need different tables of critical values that are not provided in this textbook.

METHOD 1: Using a *p*-value to Make a Decision

USING THE TI-83, 83+, 84, 84+ CALCULATOR

To calculate the *p*-value using LinRegTTEST:

1. Complete the same steps as the LinRegTTest performed previously in this chapter, making sure on the line prompt for β or σ, $\neq$ 0 is highlighted.
2. When looking at the output screen, the *p*-value is on the line that reads p =.

If the *p*-value is less than the significance level ($\alpha = 0.05$):
- Decision: Reject the null hypothesis.
- Conclusion: There is sufficient evidence to conclude there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero.

If the *p*-value is *not* less than the significance level ($\alpha = 0.05$):
- Decision: Do not reject the null hypothesis.
- Conclusion: There is insufficient evidence to conclude there is a significant linear relationship between x and y because the correlation coefficient is not significantly different from zero.

You will use technology to calculate the *p*-value, but it is useful to know that the *p*-value is calculated using a t distribution with n – 2 degrees of freedom and that the *p*-value is the combined area in both tails.

An alternative way to calculate the *p*-value (p) given by LinRegTTest is the command 2*tcdf(abs(t),10^99, n–2) in 2nd DISTR.

Third Exam vs. Final Exam Example: *p*-value Method
- Consider the third exam/final exam example.
- The line of best fit is $\hat{y} = -173.51 + 4.83x$, with $r = 0.6631$, and there are $n = 11$ data points.
- Can the regression line be used for prediction? Given a third exam score (x value), can we use the line to predict the final exam score (predicted y value)?

H_0: $\rho = 0$

H_a: $\rho \neq 0$

$\alpha = 0.05$

- The p-value is 0.026 (from LinRegTTest on a calculator or from computer software).
- The p-value, 0.026, is less than the significance level of $\alpha = 0.05$.
- Decision: Reject the null hypothesis H_0.
- Conclusion: There is sufficient evidence to conclude there is a significant linear relationship between the third exam score (x) and the final exam score (y) because the correlation coefficient is significantly different from zero.

Because r is significant and the scatter plot shows a linear trend, the regression line can be used to predict final exam scores.

METHOD 2: Using a Table of Critical Values to Make a Decision

The **95 Percent Critical Values of the Sample Correlation Coefficient Table** (Table 12.9) can be used to give you a good idea of whether the computed value of r is significant. Use it to find the critical values using the degrees of freedom, $df = n - 2$. The table has already been calculated with $\alpha = 0.05$. The table tells you the positive critical value, but you should also make that number negative to have two critical values. If r is not between the positive and negative critical values, then the correlation coefficient is significant. If r is significant, then you may use the line for prediction. If r is not significant (between the critical values), you should not use the line to make predictions.

EXAMPLE 12.6

Suppose you computed $r = 0.801$ using $n = 10$ data points. The degrees of freedom would be 8 ($df = n - 2 = 10 - 2 = 8$). Using Table 12.9 with df = 8, we find that the critical value is 0.632. This means the critical values are really ±0.632. Since $r = 0.801$ and $0.801 > 0.632$, r is significant and the line may be used for prediction. If you view this example on a number line, it will help you to see that r is not between the two critical values.

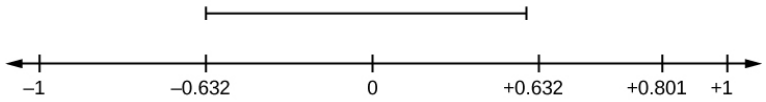


Figure 12.11 r is not between –0.632 and 0.632, so r is significant.

TRY IT 12.6

For a given line of best fit, you computed that $r = 0.6501$ using $n = 12$ data points, and the critical value found on the table is 0.576. Can the line be used for prediction? Why or why not?

EXAMPLE 12.7

Suppose you computed $r = –0.624$ with 14 data points, where $df = 14 - 2 = 12$. The critical values are –0.532 and 0.532. Since $–0.624 < –0.532$, r is significant and the line can be used for prediction.

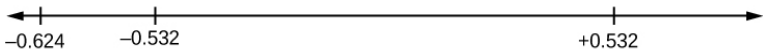


Figure 12.12 $r = –0.624$ and $–0.624 < –0.532$. Therefore, r is significant.

TRY IT 12.7

For a given line of best fit, you compute that $r = 0.5204$ using $n = 9$ data points, and the critical values are ±0.666. Can the line be used for prediction? Why or why not?

EXAMPLE 12.8

Suppose you computed $r = 0.776$ and $n = 6$, with $df = 6 -- 2 = 4$. The critical values are – 0.811 and 0.811. Since 0.776 is between the two critical values, r is not significant. The line should not be used for prediction.

Figure 12.13 $-0.811 < r = 0.776 < 0.811$. Therefore, r is not significant.

TRY IT 12.8

For a given line of best fit, you compute that $r = –0.7204$ using $n = 8$ data points, and the critical value is 0.707. Can the line be used for prediction? Why or why not?

Third Exam vs. Final Exam Example: Critical Value Method

Consider the third exam/final exam example. The line of best fit is: $\hat{y} = –173.51 + 4.83x$, with $r = .6631$, and there are $n = 11$ data points. Can the regression line be used for prediction? Given a third exam score (x value), can we use the line to predict the final exam score (predicted y value)?

H_0: $\rho = 0$
H_a: $\rho \neq 0$
$\alpha = 0.05$

- Use the 95 Percent Critical Values table for r with $df = n – 2 = 11 – 2 = 9$.
- Using the table with $df = 9$, we find that the critical value listed is 0.602. Therefore, the critical values are ±0.602.
- Since $0.6631 > 0.602$, r is significant.
- Decision: Reject the null hypothesis.
- Conclusion: There is sufficient evidence to conclude there is a significant linear relationship between the third exam score (x) and the final exam score (y) because the correlation coefficient is significantly different from zero.

Because r is significant and the scatter plot shows a linear trend, the regression line can be used to predict final exam scores.

EXAMPLE 12.9

Suppose you computed the following correlation coefficients. Using the table at the end of the chapter, determine whether r is significant and whether the line of best fit associated with each correlation coefficient can be used to predict a y value. If it helps, draw a number line.

a. $r = –0.567$ and the sample size, n, is 19.

   To solve this problem, first find the degrees of freedom. $df = n - 2 = 17$.
   Then, using the table, the critical values are ±0.456.
   $–0.567 < –0.456$, or you may say that –0.567 is not between the two critical values.
   r is significant and may be used for predictions.

b. $r = 0.708$ and the sample size, n, is 9.

   $df = n - 2 = 7$
   The critical values are ±0.666.
   $0.708 > 0.666$.
   r is significant and may be used for predictions.

c. $r = 0.134$ and the sample size, n, is 14.

   $df = 14 -- 2 = 12$.
   The critical values are ±0.532.
   0.134 is between –0.532 and 0.532.

r is not significant and may not be used for predictions.

d. *r* = 0 and the sample size, *n*, is 5.

It doesn''t matter what the degrees of freedom are because *r* = 0 will always be between the two critical values, so *r* is not significant and may not be used for predictions.

TRY IT 12.9

For a given line of best fit, you compute that *r* = 0 using *n* = 100 data points. Can the line be used for prediction? Why or why not?

Assumptions in Testing the Significance of the Correlation Coefficient

Testing the significance of the correlation coefficient requires that certain assumptions about the data be satisfied. The premise of this test is that the data are a sample of observed points taken from a larger population. We have not examined the entire population because it is not possible or feasible to do so. We are examining the sample to draw a conclusion about whether the linear relationship that we see between *x* and *y* in the sample data provides strong enough evidence that we can conclude there is a linear relationship between *x* and *y* in the population.

The regression line equation that we calculate from the sample data gives the best-fit line for our particular sample. We want to use this best-fit line for the sample as an estimate of the best-fit line for the population. Examining the scatter plot and testing the significance of the correlation coefficient helps us determine whether it is appropriate to do this.

The assumptions underlying the test of significance are as follows:

- There is a linear relationship in the population that models the sample data. Our regression line from the sample is our best estimate of this line in the population.
- The *y* values for any particular *x* value are normally distributed about the line. This implies there are more *y* values scattered closer to the line than are scattered farther away. Assumption 1 implies that these normal distributions are centered on the line; the means of these normal distributions of *y* values lie on the line.
- Normal distributions of all the *y* values have the same shape and spread about the line.
- The residual errors are mutually independent (no pattern).
- The data are produced from a well-designed, random sample or randomized experiment.

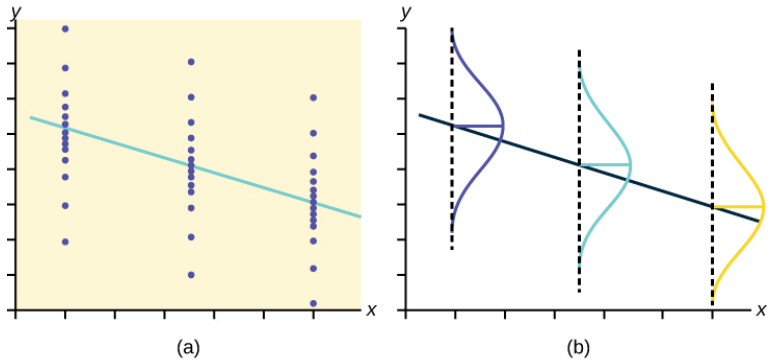


Figure 12.14 The *y* values for each *x* value are normally distributed about the line with the same standard deviation. For each *x* value, the mean of the *y* values lies on the regression line. More *y* values lie near the line than are scattered farther away from the line.

12.4 Prediction (Optional)

Recall the third exam/final exam example.

We found the equation of the best-fit line for the final exam grade as a function of the grade on the third exam. We can now use the least-squares regression line for prediction.

Suppose you want to estimate, or predict, the mean final exam score of statistics students who received a 73 on the third exam. The exam scores (*x* values) range from 65 to 75. Since 73 is between the *x* values 65 and 75, substitute *x* = 73 into

the equation. Then,

$$\hat{y} = -173.51 + 4.83(73) = 179.08.$$

We predict that statistics students who earn a grade of 73 on the third exam will earn a grade of 179.08 on the final exam, on average.

EXAMPLE 12.10

Recall the third exam/final exam example.

Problem

a. What would you predict the final exam score to be for a student who scored a 66 on the third exam?

Solution

a. 145.27

Problem

b. What would you predict the final exam score to be for a student who scored a 90 on the third exam?

Solution

b. The *x* values in the data are between 65 and 75. 90 is outside the domain of the observed *x* values in the data (independent variable), so you cannot reliably predict the final exam score for this student. Even though it is possible to enter 90 into the equation for *x* and calculate a corresponding *y* value, the *y* value that you get will not be reliable.

To understand how unreliable the prediction can be outside the *x* values observed in the data, make the substitution $x = 90$ into the equation:

$$\hat{y} = -173.51 + 4.83(90) = 261.19.$$

The final exam score is predicted to be 261.19. The most points that can be awarded for the final exam are 200.

TRY IT 12.10

Data are collected on the relationship between the number of hours per week practicing a musical instrument and scores on a math test. The line of best fit is as follows:

$\hat{y} = 72.5 + 2.8x$.
What would you predict the score on a math test will be for a student who practices a musical instrument for five hours a week?

12.5 Outliers

In some data sets, there are values (observed data points) called **outliers**. Outliers are observed data points that are far from the least-squares line. They have large errors, where the error or residual is not very close to the best-fit line.

Outliers need to be examined closely. Sometimes, they should not be included in the analysis of the data, like if it is possible that an outlier is a result of incorrect data. Other times, an outlier may hold valuable information about the population under study and should remain included in the data. The key is to examine carefully what causes a data point to be an outlier.

Besides outliers, a sample may contain one or a few points that are called influential points. Influential points are observed data points that are far from the other observed data points in the horizontal direction. These points may have a big effect on the slope of the regression line. To begin to identify an influential point, you can remove it from the data set and determine whether the slope of the regression line is changed significantly.

You also want to examine how the correlation coefficient, *r*, has changed. Sometimes, it is difficult to discern a significant change in slope, so you need to look at how the strength of the linear relationship has changed. Computers and many

calculators can be used to identify outliers and influential points. Regression analysis can determine if an outlier is, indeed, an influential point. The new regression will show how omitting the outlier will affect the correlation among the variables, as well as the fit of the line. A graph showing both regression lines helps determine how removing an outlier affects the fit of the model.

Identifying Outliers

We could guess at outliers by looking at a graph of the scatter plot and best-fit line. However, we would like some guideline regarding how far away a point needs to be to be considered an outlier. *As a rough rule of thumb, we can flag as an outlier any point that is located farther than two standard deviations above or below the best-fit line*. The standard deviation used is the standard deviation of the residuals or errors.

We can do this visually in the scatter plot by drawing an extra pair of lines that are two standard deviations above and below the best-fit line. Any data points outside this extra pair of lines are flagged as potential outliers. Or, we can do this numerically by calculating each residual and comparing it with twice the standard deviation. With regard to the TI-83, 83+, or 84+ calculators, the graphical approach is easier. The graphical procedure is shown first, followed by the numerical calculations. You would generally need to use only one of these methods.

EXAMPLE 12.11

Problem

In the third exam/final exam example, you can determine whether there is an outlier. If there is an outlier, as an exercise, delete it and fit the remaining data to a new line. For this example, the new line ought to fit the remaining data better. This means the SSE (sum of the squared errors) should be smaller and the correlation coefficient ought to be closer to 1 or –1.

Solution

Graphical Identification of Outliers

With the TI-83, 83+, or 84+ graphing calculators, it is easy to identify the outliers graphically and visually. If we were to measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance were equal to $2s$ or more, then we would consider the data point to be too far from the line of best fit. We need to find and graph the lines that are two standard deviations below and above the regression line. Any points that are outside these two lines are outliers. Let's call these lines Y2 and Y3.

As we did with the equation of the regression line and the correlation coefficient, we will use technology to calculate this standard deviation for us. Using the LinRegTTest with these data, scroll down through the output screens to find $s = 16.412$.

Line Y2 = $-173.5 + 4.83x - 2(16.4)$, and line Y3 = $-173.5 + 4.83x + 2(16.4)$,
where $\hat{y} = -173.5 + 4.83x$ is the line of best fit. Y2 and Y3 have the same slope as the line of best fit.

Graph the scatter plot with the best-fit line in equation Y1, then enter the two extra lines as Y2 and Y3 in the *Y*= equation editor. Press ZOOM-9 to get a good view. You will see that the only point that is not between Y2 and Y3 is the point (65, 175). On the calculator screen, it is barely outside these lines, but it is considered an outlier because it is more than two standard deviations away from the best-fit line. The outlier is the student who had a grade of 65 on the third exam and 175 on the final exam.

Sometimes a point is so close to the lines used to flag outliers on the graph that it is difficult to tell whether the point is between or outside the lines. On a computer, enlarging the graph may help; on a small calculator screen, zooming in may make the graph clearer. Note that when the graph does not give a clear enough picture, you can use the numerical comparisons to identify outliers.

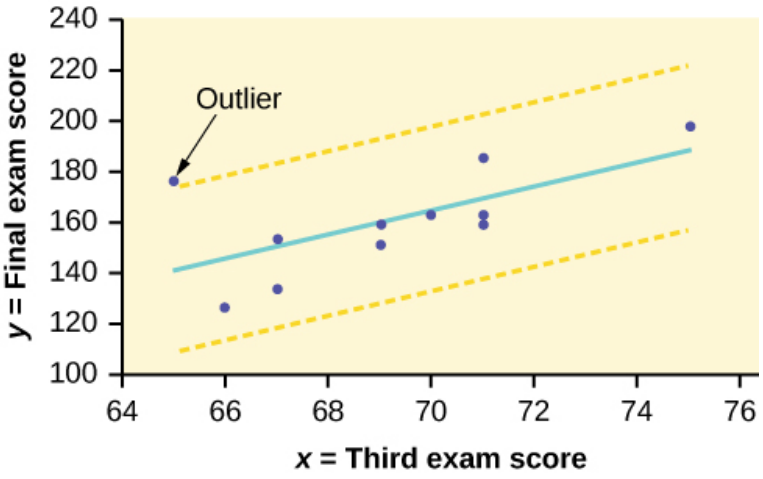


Figure 12.15

TRY IT 12.11

Identify the potential outlier in the scatter plot. The standard deviation of the residuals, or errors, is approximately 8.6.

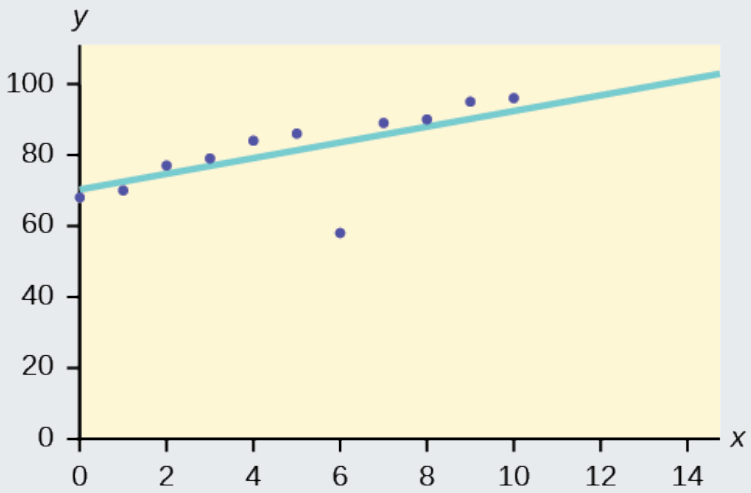


Figure 12.16

Numerical Identification of Outliers

In Table 12.6, the first two columns include the third exam and final exam data. The third column shows the predicted $\hat{y}$ values calculated from the line of best fit: $\hat{y} = -173.5 + 4.83x$. The residuals, or errors, that were mentioned in Section 3 of this chapter have been calculated in the fourth column of the table: Observed y value – predicted y value = $y - \hat{y}$.

s is the standard deviation of all the $y - \hat{y} = \varepsilon$ values, where n is the total number of data points. If each residual is calculated and squared, and the results are added, we get the SSE. The standard deviation of the residuals is calculated from the SSE as

$$s = \sqrt{\frac{SSE}{n-2}}.$$

Note

We divide by (n – 2) because the regression model involves two estimates.

Rather than calculate the value of s ourselves, we can find s using a computer or calculator. For this example, the calculator function LinRegTTest found s = 16.4 as the standard deviation of the residuals 35; –17; 16; –6; –19; 9; 3; –1; –10; –9; –1 .

x	y	$\hat{y}$	$y - \hat{y}$
65	175	140	175 – 140 = 35
67	133	150	133 – 150= –17
71	185	169	185 – 169 = 16
71	163	169	163 – 169 = –6
66	126	145	126 – 145 = –19
75	198	189	198 – 189 = 9
67	153	150	153 – 150 = 3
70	163	164	163 – 164 = –1
71	159	169	159 – 169 = –10
69	151	160	151 – 160 = –9
69	159	160	159 – 160 = –1

Table 12.6

We are looking for all data points for which the residual is greater than 2s = 2(16.4) = 32.8 or less than –32.8. Compare these values with the residuals in column four of the table. The only such data point is the student who had a grade of 65 on the third exam and 175 on the final exam; the residual for this student is 35.

How Does the Outlier Affect the Best-Fit Line?

Numerically and graphically, we have identified point (65, 175) as an outlier. Recall that recalculation of the least-squares regression line and summary statistics, following deletion of an outlier, may be used to determine whether an outlier is also an influential point. This process also allows you to compare the strength of the correlation of the variables and possible changes in the slope both before and after the omission of any outliers.

Compute a new best-fit line and correlation coefficient using the 10 remaining points.

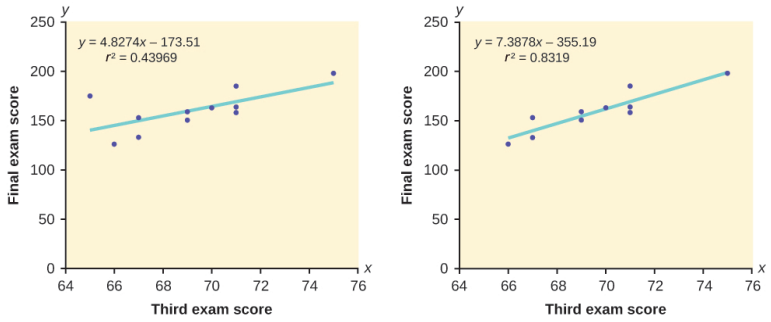
On the TI-83, TI-83+, or TI-84+ calculators, delete the outlier from L1 and L2. Using the LinRegTTest, found under Stat and Tests, the new line of best fit and correlation coefficient are the following:

$\hat{y} = -355.19 + 7.39x$ and $r = 0.9121$.

The slope is now 7.39, compared to the previous slope of 4.83. This seems significant, but we need to look at the change in r-values as well. The new line shows $r = 0.9121$, which indicates a stronger correlation than the original line, with $r = 0.6631$, since $r = 0.9121$ is closer to 1. This means the new line is a better fit to the data values. The line can better predict the final exam score given the third exam score. It also means the outlier of (65, 175) was an influential point, since there is a sizeable difference in r-values. We must now decide whether to delete the outlier. If the outlier was recorded erroneously, it should certainly be deleted. Because it produces such a profound effect on the correlation, the

new line of best fit allows for better prediction and an overall stronger model.

You may use Excel to graph the two least-squares regression lines and compare the slopes and fit of the lines to the data, as shown in Figure 12.17.



(a) Scatter plot of final exam score vs. third exam score with complete data set

(b) Scatter plot of final exam score vs. third exam score without student 1

Figure 12.17

You can see that the second graph shows less deviation from the line of best fit. It is clear that omission of the influential point produced a line of best fit that more closely models the data.

Numerical Identification of Outliers: Calculating *s* and Finding Outliers Manually

If you do not have the function LinRegTTest on your calculator, then you must calculate the outlier in the first example by doing the following.

First, square each $|y - \hat{y}|$.

The squares are 35^2; 17^2; 16^2; 6^2; 19^2; 9^2; 3^2; 1^2; 10^2; 9^2; 1^2.

Then, add (sum) all the $|y - \hat{y}|$ squared terms using the formula

$\sum_{i=1}^{11}(|y_i - \hat{y}_i|)^2 = \sum_{i=1}^{11}\varepsilon_i^2$ (Recall that $y_i - \hat{y}_i = \varepsilon_i$).

$= 35^2 + 17^2 + 16^2 + 6^2 + 19^2 + 9^2 + 3^2 + 1^2 + 10^2 + 9^2 + 1^2$

= 2,440 = SSE.

The result, **SSE**, is the sum of squared errors.

Next, calculate *s*, the standard deviation of all the $y - \hat{y} = \varepsilon$-values where *n* = the total number of data points.

The calculation is $s = \sqrt{\frac{\text{SSE}}{n-2}}$.

For the third exam/final exam example, $s = \sqrt{\frac{2440}{11-2}} = 16.47$.

Next, multiply *s* by 2:
(2)(16.47) = 32.94
32.94 is two standard deviations away from the mean of the $y - \hat{y}$ values.

If we were to measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance is at least 2*s*, then we would consider the data point to be too far from the line of best fit. We call that point a **potential outlier**.

For the example, if any of the $|y - \hat{y}|$ values are *at least* 32.94, the corresponding (*x*, *y*) data point is a potential outlier.

For the third exam/final exam example, all the $|y - \hat{y}|$ values are less than 31.29 except for the first one, which is 35:

35 > 31.29. That is, $|y - \hat{y}| \geq (2)(s)$.

The point that corresponds to $|y - \hat{y}|$ = 35 is (65, 175). *Therefore, the data point (65, 175) is a potential outlier.* For this example, we will delete it. (Remember, we do not always delete an outlier.)

Note

When outliers are deleted, the researcher should either record that data were deleted, and why, or the researcher should provide results both with and without the deleted data. If data are erroneous and the correct values are known (e.g., student 1 actually scored a 70 instead of a 65), then this correction can be made to the data.

The next step is to compute a new best-fit line using the 10 remaining points. The new line of best fit and the correlation coefficient are

$\hat{y} = -355.19 + 7.39x$ and $r = .9121$.

EXAMPLE 12.12

Problem

Using this new line of best fit (based on the remaining 10 data points in the third exam/final exam example), what would a student who receives a 73 on the third exam expect to receive on the final exam? Is this the same as the prediction made using the original line?

Solution

Using the new line of best fit, $\hat{y} = -355.19 + 7.39(73) = 184.28$. A student who scored 73 points on the third exam would expect to earn 184 points on the final exam.

The original line predicted that $\hat{y} = -173.51 + 4.83(73) = 179.08$, so the prediction using the new line with the outlier eliminated differs from the original prediction.

TRY IT 12.12

The data points for the graph from the third exam/final exam example are as follows: (1, 5), (2, 7), (2, 6), (3, 9), (4, 12), (4, 13), (5, 18), (6, 19), (7, 12), and (7, 21). Remove the outlier and recalculate the line of best fit. Find the value of $\hat{y}$ when $x = 10$.

EXAMPLE 12.13

The consumer price index (CPI) measures the average change over time in prices paid by urban consumers for consumer goods and services. The CPI affects nearly all Americans because of the many ways it is used. One of its biggest uses is as a measure of inflation. By providing information about price changes in the nation's economy to government, businesses, and labor forces, the CPI helps them make economic decisions. The president, U.S. Congress, and the Federal Reserve Board use CPI trends to form monetary and fiscal policies. In the following table, x is the year and y is the CPI.

x	y	x	y
1915	10.1	1969	36.7
1926	17.7	1975	49.3
1935	13.7	1979	72.6

Table 12.7

x	y	x	y
1940	14.7	1980	82.4
1947	24.1	1986	109.6
1952	26.5	1991	130.7
1964	31.0	1999	166.6

Table 12.7

Problem

a. Draw a scatter plot of the data.
b. Calculate the least-squares line. Write the equation in the form $\hat{y} = a + bx$.
c. Draw the line on a scatter plot.
d. Find the correlation coefficient. Is it significant?
e. What is the average CPI for the year 1990?

Solution

a. See Figure 12.18.
b. Using our calculator, $\hat{y} = -3204 + 1.662x$ is the equation of the line of best fit.
c. See Figure 12.18.
d. $r = 0.8694$. The number of data points is $n = 14$. Use the 95 Percent Critical Values of the Sample Correlation Coefficient table at the end of Chapter 12: In this case, $df = 12$. The corresponding critical values from the table are ±0.532. Since $0.8694 > 0.532$, r is significant. We can use the predicted regression line we found above to make the prediction for $x = 1990$.
e. $\hat{y} = -3204 + 1.662(1990) = 103.4$ CPI.

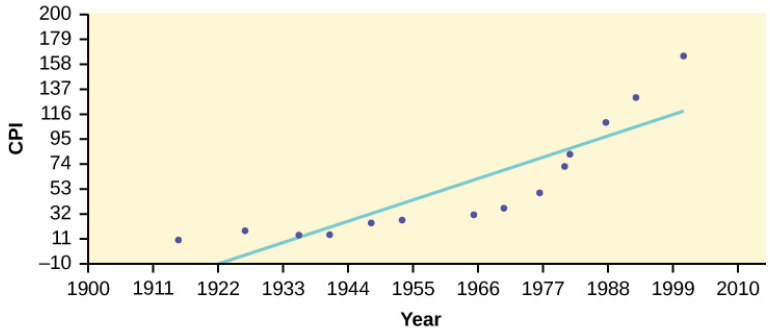


Figure 12.18

Note

In the example, notice the pattern of the points compared with the line. Although the correlation coefficient is significant, the pattern in the scatter plot indicates that a curve would be a more appropriate model to use than a line. In this example, a statistician would prefer to use other methods to fit a curve to these data, rather than model the data with the line we found. In addition to doing the calculations, it is always important to look at the scatter plot when deciding whether a linear model is appropriate.

If you are interested in seeing more years of data, visit the Bureau of Labor Statistics CPI website (ftp://ftp.bls.gov/pub/special.requests/cpi/cpiai.txt). Our data are taken from the column Annual Avg. (third column from the right). For example, you could add more current years of data. Try adding the more recent years: 2004, CPI = 188.9; 2008, CPI = 215.3; and 2011, CPI = 224.9. See how this affects the model. (Check: $\hat{y} = -4436 + 2.295x$; $r = 0.9018$. Is r significant? Is the

fit better with the addition of the new points?)

TRY IT 12.13

The following table shows economic development measured in per capita income (PCINC).

Year	PCINC	Year	PCINC
1870	340	1920	1,050
1880	499	1930	1,170
1890	592	1940	1,364
1900	757	1950	1,836
1910	927	1960	2,132

Table 12.8

a. What are the independent and dependent variables?
b. Draw a scatter plot.
c. Use regression to find the line of best fit and the correlation coefficient.
d. Interpret the significance of the correlation coefficient.
e. Is there a linear relationship between the variables?
f. Find the coefficient of determination and interpret it.
g. What is the slope of the regression equation? What does it mean?
h. Use the line of best fit to estimate PCINC for 1900 and for 2000.
i. Determine whether there are any outliers.

95 Percent Critical Values of the Sample Correlation Coefficient Table

Degrees of Freedom: $n - 2$	Critical Values: + and –
1	0.997
2	0.950
3	0.878
4	0.811
5	0.754
6	0.707
7	0.666
8	0.632

Table 12.9

Degrees of Freedom: $n - 2$	Critical Values: + and –
9	0.602
10	0.576
11	0.555
12	0.532
13	0.514
14	0.497
15	0.482
16	0.468
17	0.456
18	0.444
19	0.433
20	0.423
21	0.413
22	0.404
23	0.396
24	0.388
25	0.381
26	0.374
27	0.367
28	0.361
29	0.355
30	0.349
40	0.304
50	0.273
60	0.250

Table 12.9

Degrees of Freedom: $n - 2$	Critical Values: + and –
70	0.232
80	0.217
90	0.205
100	0.195

Table 12.9

12.6 Regression (Distance from School) (Optional)

Stats Lab

Regression (Distance From School)

Student Learning Outcomes

- The student will calculate and construct the line of best fit between two variables.
- The student will evaluate the relationship between two variables to determine whether that relationship is significant.

Collect the Data

Use eight members of your class for the sample. Collect bivariate data (distance an individual lives from school, the cost of supplies for the current term).

1. Complete the table.

Distance from School	Cost of Supplies This Term

Table 12.10

2. Which variable should be the dependent variable and which should be the independent variable? Why?
3. Graph *distance* vs. *cost*. Plot the points on the graph. Label both axes with words. Scale both axes.

Figure 12.19

Analyze the Data

Enter your data into a calculator or computer. Write the linear equation, rounding to four decimal places.

1. Calculate the following:
   a. a = _____
   b. b = _____
   c. correlation = _____
   d. n = _____
   e. equation: $\hat{y}$ = _____
   f. Is the correlation significant? Why or why not? (Answer in one to three complete sentences.)
2. Supply an answer for the following scenarios:
   a. For a person who lives eight miles from campus, predict the total cost of supplies this term.
   b. For a person who lives 80 miles from campus, predict the total cost of supplies this term.
3. Obtain the graph on a calculator or computer. Sketch the regression line.

Figure 12.20

Discussion Questions

1. Answer each question in complete sentences.
   a. Does the line seem to fit the data? Why?
   b. What does the correlation imply about the relationship between distance and cost?
2. Are there any outliers? If so, which point is an outlier?
3. Should the outlier, if it exists, be removed? Why or why not?

12.7 Regression (Textbook Cost) (Optional)

Stats Lab

Regression (Textbook Cost)

Student Learning Outcomes

- The student will calculate and construct the line of best fit between two variables.
- The student will evaluate the relationship between two variables to determine whether that relationship is significant.

Collect the Data

Survey 10 textbooks. Collect bivariate data (number of pages in a textbook, the cost of the textbook).

1. Complete the table.

Number of Pages	Cost of Textbook

Table 12.11

2. Which variable should be the dependent variable and which should be the independent variable? Why?
3. Graph *pages vs. cost*. Plot the points on the graph in Analyze the Data. Label both axes with words. Scale both axes.

Analyze the Data

Enter your data into a calculator or computer. Write the linear equation, rounding to four decimal places.

1. Calculate the following:
   a. a = _____
   b. b = _____
   c. correlation = _____
   d. n = _____
   e. equation: y = _____
   f. Is the correlation significant? Why or why not? (Answer in complete sentences.)
2. Supply an answer for the following scenarios:
   a. For a textbook with 400 pages, predict the cost.
   b. For a textbook with 600 pages, predict the cost.
3. Obtain the graph on a calculator or computer. Sketch the regression line.

Figure 12.21

Discussion Questions

1. Answer each question in complete sentences.
   a. Does the line seem to fit the data? Why?
   b. What does the correlation imply about the relationship between the number of pages and the cost?
2. Are there any outliers? If so, which point is an outlier?
3. Should the outlier, if it exists, be removed? Why or why not?

12.8 Regression (Fuel Efficiency) (Optional)

Stats Lab

Regression (Fuel Efficiency)

Student Learning Outcomes

- The student will calculate and construct the line of best fit between two variables.
- The student will evaluate the relationship between two variables to determine whether that relationship is significant.

Collect the Data

Find a reputable source that provides information on total fuel efficiency (in miles per gallon) and weight (in pounds) of new cars with an automatic transmission. You will use these data to determine the relationship, if any, between the fuel efficiency of a car and its weight.

1. Using your random-number generator, select 20 cars randomly from the list and record their weight and fuel efficiency into Table 12.12.

Weight	Fuel Efficiency

Table 12.12

Weight	Fuel Efficiency

Table 12.12

2. Which variable is the dependent variable and which is the independent variable? Why?
3. By hand, draw a scatter plot of *weight vs. fuel efficiency*. Plot the points on graph paper. Label both axes with words. Scale both axes accurately.

Figure 12.22

Analyze the Data

Enter your data into a calculator or computer. Write the linear equation, rounding to four decimal places.

1. Calculate the following:
   a. a = _____
   b. b = _____
   c. correlation = _____
   d. n = _____
   e. equation: $\hat{y}$ = _____
2. Obtain a graph of the regression line on a calculator. Sketch the regression line on the same axes as your scatter plot.

Discussion Questions

1. Is the correlation significant? Explain how you determined this in complete sentences.
2. Is the relationship a positive one or a negative one? Explain how you can tell and what this means in terms of weight and fuel efficiency.
3. In one or two complete sentences, what is the practical interpretation of the slope of the least-squares line in terms of fuel efficiency and weight?
4. For a car that weighs 4,000 pounds, predict its fuel efficiency. Include units.
5. Can we predict the fuel efficiency of a car that weighs 10,000 pounds using the least-squares line? Explain why or why not.
6. Answer each question in complete sentences.
   a. Does the line seem to fit the data? Why or why not?
   b. What does the correlation imply about the relationship between fuel efficiency and weight of a car? Is this what

you expected?

7. Are there any outliers? If so, which point is an outlier?

Key Terms

coefficient of correlation a measure developed by Karl Pearson during the early 1900s that gives the strength of association between the independent variable and the dependent variable;

$$r = \frac{n\sum xy - [\sum x][\sum y]}{\sqrt{(n\sum x^2 - [\sum x]^2)(n\sum y^2 - [\sum y]^2)}}$$

where n is the number of data points

The coefficient cannot be more than 1 and less than –1. The closer the coefficient is to ±1, the stronger the evidence of a significant linear relationship between x and y.

outlier an observation that does not fit the rest of the data

Chapter Review

12.1 Linear Equations

The most basic type of association is a linear association. This type of relationship can be defined algebraically by the equations used (numerically with actual or predicted data values) or graphically from a plotted curve. Lines are classified as straight curves. Algebraically, a linear equation typically takes the form $y = mx + b$, where m and b are constants, x is the independent variable, and y is the dependent variable. In a statistical context, a linear equation is written in the form $y = a + bx$, where a and b are the constants. This form is used to help you distinguish the statistical context from the algebraic context. In the equation $y = a + bx$, the constant b that multiplies the x variable (b is called a *coefficient*) is called the *slope*. The slope describes the rate of change between the independent and dependent variables; in other words, the rate of change describes the change that occurs in the dependent variable as the independent variable is changed. In the equation $y = a + bx$, the constant a is called the y-intercept. Graphically, the y-intercept is the y-coordinate of the point where the graph of the line crosses the y-axis. At this point, $x = 0$.

The slope of a line is a value that describes the rate of change between the independent and dependent variables. The slope tells us how the dependent variable (y) changes for every one-unit increase in the independent (x) variable, on average. The y-intercept is used to describe the dependent variable when the independent variable equals zero. Graphically, the slope is represented by three line types in elementary statistics.

12.2 The Regression Equation

A regression line, or a line of best fit, can be drawn on a scatter plot and used to predict outcomes for the x and y variables in a given data set or sample data. There are several ways to find a regression line, but usually the least-squares regression line is used because it creates a uniform line. Residuals, also called *errors*, measure the distance from the actual value of y and the estimated value of y. The sum of squared errors, or SSE, when set to its minimum, calculates the points on the line of best fit. Regression lines can be used to predict values within the given set of data but should not be used to make predictions for values outside the set of data.

The correlation coefficient, r, measures the strength of the linear association between x and y. The variable r has to be between –1 and +1. When r is positive, x and y tend to increase and decrease together. When r is negative, x increases and y decreases, or the opposite occurs: x decreases and y increases. The coefficient of determination, r^2, is equal to the square of the correlation coefficient. When expressed as a percentage, r^2 represents the percentage of variation in the dependent variable, y, that can be explained by variation in the independent variable, x, using the regression line.

12.3 Testing the Significance of the Correlation Coefficient (Optional)

Linear regression is a procedure for fitting a straight line of the form $\hat{y} = a + bx$ to data. The conditions for regression are as follows:

- Linear: In the population, there is a linear relationship that models the average value of y for different values of x.
- Independent: The residuals are assumed to be independent.
- Normal: The y values are distributed normally for any value of x.
- Equal variance: The standard deviation of the y values is equal for each x value.
- Random: The data are produced from a well-designed random sample or a randomized experiment.

The slope b and intercept a of the least-squares line estimate the slope β and intercept α of the population (true) regression line. To estimate the population standard deviation of y (σ) use the standard deviation of the residuals:

$s = \sqrt{\frac{SSE}{n-2}}$. The variable ρ (rho) is the population correlation coefficient. To test the null hypothesis, H_0: ρ = hypothesized value, use a linear regression *t*-test. The most common null hypothesis is H_0: $\rho = 0$, which indicates there is no linear relationship between *x* and *y* in the population. The TI-83, 83+, 84, 84+ calculator function LinRegTTest can perform this test (STATS, TESTS, LinRegTTest).

12.4 Prediction (Optional)

After determining the presence of a strong correlation coefficient and calculating the line of best fit, you can use the least-squares regression line to make predictions about your data.

12.5 Outliers

To determine whether a point is an outlier, do one of the following:

1. Input the following equations into the TI 83, 83+, 84, or 84+ calculator:

   $y_1 = a + bx$
   $y_2 = a + bx + 2s$
   $y_3 = a + bx - 2s$

   where *s* is the standard deviation of the residuals.

   If any point is above y_2 or below y_3, then the point is considered to be an outlier.
2. Use the residuals and compare their absolute values to 2*s*, where *s* is the standard deviation of the residuals. If the absolute value of any residual is greater than or equal to 2*s*, then the corresponding point is an outlier.
3. Note: The calculator function LinRegTTest (STATS, TESTS, LinRegTTest) calculates *s*.

Formula Review

12.1 Linear Equations

$y = a + bx$, where *a* is the *y*-intercept and *b* is the slope. The variable *x* is the independent variable and *y* is the dependent variable.

12.3 Testing the Significance of the Correlation Coefficient (Optional)

Least-Squares Line or Line of Best Fit:

$\hat{y} = a + bx$,

where *a* is the *y*-intercept and *b* is the slope.

Standard Deviation of the Residuals:

$s = \sqrt{\frac{SSE}{n-2}}$,

where SSE = sum of squared errors, and

n = the number of data points.

Practice

12.1 Linear Equations

Use the following information to answer the next three exercises. A vacation resort rents scuba equipment to certified divers. The resort charges an up-front fee of $25 and another fee of $12.50 an hour.

1. What are the dependent and independent variables?

2. Find the equation that expresses the total fee in terms of the number of hours the equipment is rented.

3. Graph the equation from Exercise 12.2.

Use the following information to answer the next two exercises. A credit card company charges $10 when a payment is late and $5 a day each day the payment remains unpaid.

4. Find the equation that expresses the total fee in terms of the number of days the payment is late.

5. Graph the equation from Exercise 12.4.

6. Is the equation $y = 10 + 5x - 3x^2$ linear? Why or why not?

7. Which of the following equations are linear?
a. $y = 6x + 8$
b. $y + 7 = 3x$
c. $y - x = 8x^2$
d. $4y = 8$

8. Does the graph in Figure 12.23 show a linear equation? Why or why not?

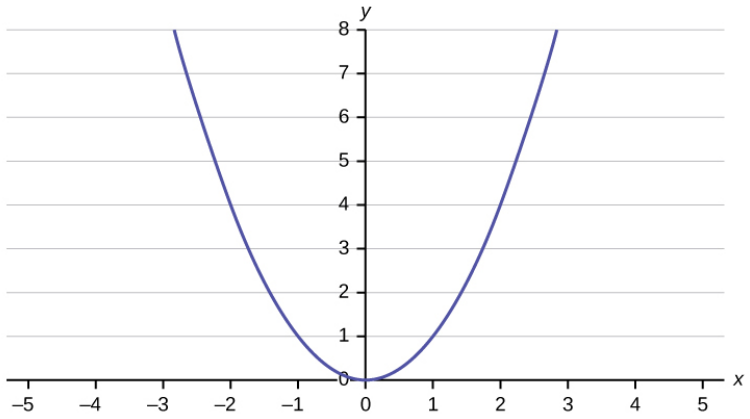


Figure 12.23

Use the following information to answer the next exercise. Table 12.13 contains real data for the first two decades of flu reporting.

Year	Number of Flu Cases Diagnosed	Number of Flu Deaths
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987
1987	28,564	16,162
1988	35,447	20,868
1989	42,674	27,591
1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456

Table 12.13

1996	59,347	38,510
1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	**802,118**	**489,093**

Table 12.13

9. Use the columns *Year* and *Number of Flu Cases Diagnosed*. Why is year the independent variable and number of flu cases diagnosed the dependent variable (instead of the reverse)?

Use the following information to answer the next two exercises. A specialty cleaning company charges an equipment fee and an hourly labor fee. A linear equation that expresses the total amount of the fee the company charges for each session is $y = 50 + 100x$.

10. What are the independent and dependent variables?

11. What is the *y*-intercept, and what is the slope? Interpret them using complete sentences.

Use the following information to answer the next three questions. As a result of erosion, a river shoreline is losing several thousand pounds of soil each year. A linear equation that expresses the total amount of soil lost per year is $y = 12{,}000x$.

12. What are the independent and dependent variables?

13. How many pounds of soil does the shoreline lose in a year?

14. What is the *y*-intercept? Interpret its meaning.

Use the following information to answer the next two exercises. The price of a single issue of stock can fluctuate throughout the day. A linear equation that represents the price of stock for Shipment Express is $y = 15 – 1.5x$, where x is the number of hours passed in an eight-hour day of trading.

15. What are the slope and *y*-intercept? Interpret their meaning.

16. If you owned this stock, would you want a positive or negative slope? Why?

12.2 The Regression Equation

17. Table 12.16 below represents the relationship between the number of hours spent studying and final exam grades.

x (number of hours spent studying)	*y* (final exam grades)
3	50
5	72
1	45
2	51
6	80
8	96
4	65
7	90

Table 12.14

Fill in the following chart as a first step in finding the line of best fit, using the median–median approach.

Group	*x* (no. of hours spent studying)	*y* (final exam grades)	Median *x* Value	Median *y* Value
1				
2				
3				

Table 12.15

Use the following information to answer the next five exercises. A random sample of 10 professional athletes produced the following data, where *x* is the number of endorsements the player has and *y* is the amount of money made, in millions of dollars.

x	*y*	*x*	*y*
0	2	5	12
3	8	4	9
2	7	3	9
1	3	0	3
5	13	4	10

Table 12.16

18. Draw a scatter plot of the data.

19. Use regression to find the equation for the line of best fit.

20. Draw the line of best fit on the scatter plot.

21. What is the slope of the line of best fit? What does it represent?

22. What is the *y*-intercept of the line of best fit? What does it represent?

23. What does an *r* value of zero mean?

24. When $n = 2$ and $r = 1$, are the data significant? Explain.

25. When $n = 100$ and $r = -0.89$, is there a significant correlation? Explain.

12.3 Testing the Significance of the Correlation Coefficient (Optional)

26. When testing the significance of the correlation coefficient, what is the null hypothesis?

27. When testing the significance of the correlation coefficient, what is the alternative hypothesis?

28. If the level of significance is 0.05 and the *p*-value is 0.04, what conclusion can you draw?

12.4 Prediction (Optional)

Use the following information to answer the next two exercises. An electronics retailer used regression to find a simple model to predict sales growth in the first quarter of the new year (January through March). The model is good for 90 days, where x is the day. The model can be written as $\hat{y} = 101.32 + 2.48x$, where $\hat{y}$ is in thousands of dollars.

29. What would you predict the sales to be on day 60?

30. What would you predict the sales to be on day 90?

Use the following information to answer the next three exercises. A landscaping company is hired to mow the grass for several large properties. The total area of the properties is 1,345 acres. The rate at which one person can mow is $\hat{y} = 1350 - 1.2x$, where x is the number of hours and $\hat{y}$ represents the number of acres left to mow.

31. How many acres are left to mow after 20 hours of work?

32. How many acres are left to mow after 100 hours of work?

33. How many hours does it take to mow all the lawns, or when is $\hat{y} = 0$?

Use the following information to answer the next 14 exercises. Table 12.17 contains real data for the first two decades of flu reporting.

Year	Number of Flu Cases Diagnosed	Number of Flu Deaths
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987
1987	28,564	16,162
1988	35,447	20,868
1989	42,674	27,591

Table 12.17 Adults and Adolescents Only, United States

1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456
1996	59,347	38,510
1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	**802,118**	**489,093**

Table 12.17 Adults and Adolescents Only, United States

34. Graph year versus number of flu cases diagnosed (plot the scatter plot). Do not include pre-1981 data.

35. Perform a linear regression. What is the linear equation? Round to the nearest whole number. Find the following:
Write the equations:
- *Linear equation:* _____
- a = _____
- b = _____
- r = _____
- n = _____

36. Solve.
a. When $x = 1985$, $\hat{y}$ = _____.
b. When $x = 1990$, $\hat{y}$ = _____.
c. When $x = 1970$, $\hat{y}$ = _____. Why doesn't this answer make sense?

37. Does the line seem to fit the data? Why or why not?

38. What does the correlation imply about the relationship between time (years) and the number of diagnosed flu cases reported in the United States?

39. Plot the two points on the graph. Then, connect the two points to form the regression line.

40. Write the equation: $\hat{y}$ = _____.

41. Hand-draw a smooth curve on the graph that shows the flow of the data.

42. Does the line seem to fit the data? Why or why not?

43. Do you think a linear fit is best? Why or why not?

44. What does the correlation imply about the relationship between time (years) and the number of diagnosed flu cases reported in the United States?

45. Graph *year* vs. *number flu cases diagnosed*. Do not include pre-1981. Label both axes with words. Scale both axes.

46. Enter your data into your calculator or computer. The pre-1981 data should not be included. Why is that so? Write the linear equation, rounding to four decimal places.

47. *Calculate the following:*

- $a =$ _____
- $b =$ _____
- correlation = _____
- $n =$ _____

12.5 Outliers

48. Marcus states that all outliers are influential points. Is he correct? Explain.

Use the following information to answer the next four exercises. The scatter plot shows the relationship between hours spent studying and exam scores. The line shown is the calculated line of best fit. The correlation coefficient is 0.69.

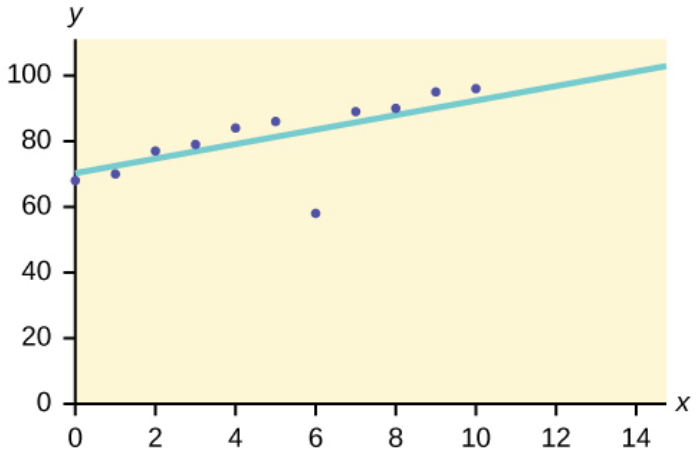


Figure 12.24

49. Do there appear to be any outliers?

50. A point is removed and the line of best fit is recalculated. The new correlation coefficient is 0.98. Does the point appear to have been an outlier? Why?

51. What effect did the potential outlier have on the line of best fit?

52. Are you more or less confident in the predictive ability of the new line of best fit?

53. The sum of squared errors (SSE) for a data set of 18 numbers is 49. What is the standard deviation?

54. The standard deviation for the SSE for a data set is 9.8. What is the cutoff for the vertical distance that a point can be from the line of best fit to be considered an outlier?

Homework

12.1 Linear Equations

55. For each of the following situations, state the independent variable and the dependent variable.

a. A study is done to determine whether elderly drivers are involved in more motor vehicle fatalities than other drivers. The number of fatalities per 100,000 drivers is compared with the age of drivers.
b. A study is done to determine whether the weekly grocery bill changes based on the number of family members.
c. Insurance companies base life insurance premiums partially on the age of the applicant.
d. Utility bills vary according to power consumption.
e. A study is done to determine whether a higher education reduces the crime rate in a population.

56. Piece-rate systems are widely debated incentive payment plans. In a recent study of loan officer effectiveness, the following piece-rate system was examined:

% of goal reached	< 80	80	100	120
Incentive	n/a	$4,000, with an additional $125 added per percentage point from 81% to 99%	$6,500, with an additional $125 added per percentage point from 101% to 119%	$9,500, with an additional $125 added per percentage point starting at 121%

Table 12.18

If a loan officer makes 95 percent of his or her goal, write the linear function that applies based on the incentive plan table. In context, explain the *y*-intercept and slope.

12.2 The Regression Equation

57. What is the process through which we can calculate a line that goes through a scatter plot with a linear pattern?

58. Explain what it means when a correlation has an r^2 value of .72.

59. Can a coefficient of determination be negative? Why or why not?

60. The table below represents the relationship between SAT scores on the math portion of the test and high school grade point averages (GPAs).
Use the median–-median line approach to find the equation for the line of best fit.

x (SAT math scores)	*y* (GPAs)
624	90
544	86
363	70
373	71
350	65
741	98
262	60
587	87
327	62
364	67
261	50

Table 12.19

12.4 Prediction (Optional)

61. Recently, the annual numbers of driver deaths per 100,000 people for the selected age groups are as follows:

Age (years)	Number of Driver Deaths (per 100,000 people)
16–19	38
20–24	36
25–34	24
35–54	20
55–74	18
75+	28

Table 12.20

a. For each age group, pick the midpoint of the interval for the *x* value. For the 75+ group, use 80.
b. Using *age* as the independent variable and *number of driver deaths per 100,000 people* as the dependent variable, make a scatter plot of the data.
c. Calculate the least-squares (best–fit) line. Put the equation in the form $\hat{y} = a + bx$.
d. Find the correlation coefficient. Is it significant?
e. Predict the number of deaths for ages 40 years and 60 years.
f. Based on the given data, is there a linear relationship between age of a driver and driver fatality rate?
g. What is the slope of the least-squares (best-fit) line? Interpret the slope.

62. Table 12.21 shows the life expectancy for an individual born in the United States in certain years.

Year of Birth	Life Expectancy in years
1930	59.7
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5
1987	75
1992	75.7
2010	78.7

Table 12.21

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the ordered pairs.
c. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
d. Find the correlation coefficient. Is it significant?
e. Find the estimated life expectancy for an individual born in 1950 and for one born in 1982.

f. Why aren't the answers to Part E the same as the values in Table 12.21 that correspond to those years?
g. Use the two points in Part E to plot the least-squares line on your graph from Part B.
h. Based on the data, is there a linear relationship between the year of birth and life expectancy?
i. Are there any outliers in the data?
j. Using the least-squares line, find the estimated life expectancy for an individual born in 1850. Does the least-squares line give an accurate estimate for that year? Explain why or why not.
k. What is the slope of the least-squares (best-fit) line? Interpret the slope.

63. The maximum discount value of the Entertainment® card for the *Fine Dining* section, 10th edition, for various pages is given in Table 12.22.

Page Number	Maximum Value ($)
4	16
14	19
25	15
32	17
43	19
57	15
72	16
85	15
90	17

Table 12.22

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the ordered pairs.
c. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
d. Find the correlation coefficient. Is it significant?
e. Find the estimated maximum values for the restaurants on page 10 and on page 70.
f. Does it appear that the restaurants giving the maximum value are placed in the beginning of the *Fine Dining* section? How did you arrive at your answer?
g. Suppose there are 200 pages of restaurants. What do you estimate to be the maximum value for a restaurant listed on page 200?
h. Is the least-squares line valid for page 200? Why or why not?
i. What is the slope of the least-squares (best-fit) line? Interpret the slope.

64. Table 12.23 gives the gold medal times for every other Summer Olympics for the women's 100-meter freestyle in swimming.

Year	Time in seconds
1912	82.2
1924	72.4

Table 12.23

Year	Time in seconds
1932	66.8
1952	66.8
1960	61.2
1968	60.0
1976	55.65
1984	55.92
1992	54.64
2000	53.8
2008	53.1

Table 12.23

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is the decrease in times significant?
f. Find the estimated gold medal time for 1932. Find the estimated time for 1984.
g. Why are the answers from Part F different from the chart values?
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Use the least-squares line to estimate the gold medal time for the next Summer Olympics. Do you think your answer is reasonable? Why or why not?

65.

State	No. of Letters in Name	Year Entered the Union	Rank for Entering the Union	Area in square miles
Alabama	7	1819	22	52,423
Colorado	8	1876	38	104,100
Hawaii	6	1959	50	10,932
Iowa	4	1846	29	56,276
Maryland	8	1788	7	12,407
Missouri	8	1821	24	69,709
New Jersey	9	1787	3	8,722
Ohio	4	1803	17	44,828

Table 12.24

State	No. of Letters in Name	Year Entered the Union	Rank for Entering the Union	Area in square miles
South Carolina	13	1788	8	32,008
Utah	4	1896	45	84,904
Wisconsin	9	1848	30	65,499

Table 12.24

We are interested in whether the number of letters in a state name depends on the year the state entered the Union.

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
e. Find the correlation coefficient. What does it imply about the significance of the relationship?
f. Find the estimated number of letters (to the nearest integer) a state name would have if it entered the Union in 1900. Find the estimated number of letters a state name would have if it entered the Union in 1940.
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Use the least-squares line to estimate the number of letters for a new state that enters the Union this year. Can the least-squares line be used to predict it? Why or why not?

12.5 Outliers

66. Given the information in Table 12.30, which represents the relationship between final exam math grades and final exam history grades, decide whether point (56, 95) is an influential point. Explain how you arrived at your decision. Show all work.

x (final exam math grades)	*y* (final exam history grades)
54	60
56	68
77	82
74	78
63	69
51	55
88	97
72	77
69	78
56	95

Table 12.25

67. In Table 12.31, the height (sidewalk to roof) of notable tall buildings in America is compared with the number of stories of the building (beginning at street level).

Height (in feet)	Stories
1,050	57
428	28
362	26
529	40
790	60
401	22
380	38
1,454	110
1,127	100
700	46

Table 12.26

a. Using *stories* as the independent variable and *height* as the dependent variable, make a scatter plot of the data.
b. Does it appear from inspection that there is a relationship between the variables?
c. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
d. Find the correlation coefficient. Is it significant?
e. Find the estimated heights for a building that has 32 stories and for a building that has 94 stories.
f. Based on the data in Table 12.26, is there a linear relationship between the number of stories in tall buildings and the height of the buildings?
g. Are there any outliers in the data? If so, which point(s)?
h. What is the estimated height of a building with six stories? Does the least-squares line give an accurate estimate of height? Explain why or why not.
i. Based on the least-squares line, adding an extra story is predicted to add about how many feet to a building?
j. What is the slope of the least-squares (best-fit) line? Interpret the slope.

68. Ornithologists (scientists who study birds) tag sparrow hawks in 13 different colonies to study their population. They gather data for the percentage of new sparrow hawks in each colony and the percentage of those that have returned from migration.
Percent return: 74, 66, 81, 52, 73, 62, 52, 45, 62, 46, 60, 46, 38
Percent new: 5, 6, 8, 11, 12, 15, 16, 17, 18, 18, 19, 20, 20

a. Enter the data into a calculator and make a scatter plot.
b. Use the calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot from Part A.
c. Explain what the slope and *y*-intercept of the regression line tell us.
d. How well does the regression line fit the data? Explain your response.
e. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.
f. An ecologist wants to predict how many birds will join another colony of sparrow hawks to which 70 percent of the adults from the previous year have returned. What is the prediction?

69. The following table shows data on average per capita coffee consumption and death rate from heart disease in a

random sample of 10 countries.

Yearly Coffee Consumption (liters)	2.5	3.9	2.9	2.4	2.9	0.8	9.1	2.7	0.8	0.7
No. of Deaths from Heart Disease	221	167	131	191	220	297	71	172	211	300

Table 12.27

a. Enter the data into a calculator and make a scatter plot.
b. Use the calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot from Part A.
c. Explain what the slope and *y*-intercept of the regression line tell us.
d. How well does the regression line fit the data? Explain your response.
e. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.
f. Do the data provide convincing evidence that there is a linear relationship between the amount of coffee consumed and the heart disease death rate? Carry out an appropriate test at a significance level of 0.05 to help answer this question.

70. The following table consists of one student athlete's time (in minutes) to swim 2,000 yards and the student's heart rate (beats per minute) after swimming on a random sample of 10 days.

Swim Time	Heart Rate
34.12	144
35.72	152
34.72	124
34.05	140
34.13	152
35.73	146
36.17	128
35.57	136
35.37	144
35.57	148

Table 12.28

a. Enter the data into a calculator and make a scatter plot.
b. Use the calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot from Part A.
c. Explain what the slope and *y*-intercept of the regression line tell us.
d. How well does the regression line fit the data? Explain your response.
e. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.

71. A researcher is investigating whether population impacts homicide rate. He uses demographic data from Detroit, Michigan, to compare homicide rates and the population.

Population Size	Homicide Rate per 100,000 People
558,724	8.6
538,584	8.9
519,171	8.52
500,457	8.89
482,418	13.07
465,029	14.57
448,267	21.36
432,109	28.03
416,533	31.49
401,518	37.39
387,046	46.26
373,095	47.24
359,647	52.33

Table 12.29

a. Use a calculator to construct a scatter plot of the data. What is the independent variable? Why?
b. Use the calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot.
c. Discuss what the following mean in context:
   i. The slope of the regression equation
   ii. The y-intercept of the regression equation
   iii. The correlation coefficient, r
   iv. The coefficient of determination, r^2
d. Do the data provide convincing evidence that there is a linear relationship between population size and homicide rate? Carry out an appropriate test at a significance level of 0.05 to help answer this question.

72.

School	Mid-Career Salary (in thousands of U.S. dollars)	Yearly Tuition (in U.S. dollars)
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
U.S. Naval Academy	122	0

Table 12.30

School	Mid-Career Salary (in thousands of U.S. dollars)	Yearly Tuition (in U.S. dollars)
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

Table 12.30

Use the data in the Table 12.35 to determine the linear regression line equation with the outliers removed. Is there a linear correlation for the data set with outliers removed? Justify your answer.

Bringing It Together: Homework

73. The average number of people in a family who attended college for various years is given in Table 12.31.

Year	No. of Family Members Attending College
1969	4.0
1973	3.6
1975	3.2
1979	3.0
1983	3.0
1988	3.0
1991	2.9

Table 12.31

a. Using *year* as the independent variable and *number of family members attending college* as the dependent variable, draw a scatter plot of the data.
b. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
c. Does the y-intercept, a, have any meaning here?
d. Find the correlation coefficient. Is it significant?
e. Pick two years between 1969 and 1991 and find the estimated number of family members attending college.
f. Based on the data in Table 12.31, is there a linear relationship between the year and the average number of family members attending college?
g. Using the least-squares line, estimate the number of family members attending college for 1960 and 1995. Does the least-squares line give an accurate estimate for those years? Explain why or why not.
h. Are there any outliers in the data?
i. What is the estimated average number of family members attending college for 1986? Does the least-squares line give an accurate estimate for that year? Explain why or why not.
j. What is the slope of the least-squares (best-fit) line? Interpret the slope.

74. The percent of female wage and salary workers who are paid hourly rates is given in Table 12.32 for the years 1979 to 1992.

Year	Percent of Workers Paid Hourly Rates
1979	61.2
1980	60.7
1981	61.3
1982	61.3
1983	61.8
1984	61.7
1985	61.8
1986	62.0
1987	62.7
1990	62.8
1992	62.9

Table 12.32

a. Using *year* as the independent variable and *percent of workers paid hourly rates* as the dependent variable, draw a scatter plot of the data.
b. Does it appear from inspection that there is a relationship between the variables? Why or why not?
c. Does the *y*-intercept, *a*, have any meaning here?
d. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is it significant?
f. Find the estimated percentages for 1991 and 1988.
g. Based on the data, is there a linear relationship between the year and the percentage of female wage and salary earners who are paid hourly rates?
h. Are there any outliers in the data?
i. What is the estimated percentage for the year 2050? Does the least-squares line give an accurate estimate for that year? Explain why or why not.
j. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Use the following information to answer the next two exercises. The cost of a leading liquid laundry detergent in different sizes is given in Table 12.33.

Size (ounces)	Cost ($)	Cost per Ounce
16	3.99	
32	4.99	

Table 12.33

Size (ounces)	Cost ($)	Cost per Ounce
64	5.99	
200	10.99	

Table 12.33

75. a. Using *size* as the independent variable and *cost* as the dependent variable, draw a scatter plot.
b. Does it appear from inspection that there is a relationship between the variables? Why or why not?
c. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
d. Find the correlation coefficient. Is it significant?
e. If the laundry detergent were sold in a 40 oz. size, what is the estimated cost?
f. If the laundry detergent were sold in a 90 oz. size, what is the estimated cost?
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Are there any outliers in the given data?
i. Is the least-squares line valid for predicting what a 300 oz. size of the laundry detergent would cost? Why or why not?
j. What is the slope of the least-squares (best-fit) line? Interpret the slope.

76. a. Complete Table 12.33 for the cost per ounce of the different sizes of laundry detergent.
b. Using *size* as the independent variable and *cost per ounce* as the dependent variable, draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is it significant?
f. If the laundry detergent were sold in a 40 oz. size, what is the estimated cost per ounce?
g. If the laundry detergent were sold in a 90 oz. size, what is the estimated cost per ounce?
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Are there any outliers in the the data?
j. Is the least-squares line valid for predicting what a 300 oz. size of the laundry detergent would cost per ounce? Why or why not?
k. What is the slope of the least-squares (best-fit) line? Interpret the slope.

77. According to a flyer published by Prudential Insurance Company, the costs of approximate probate fees and taxes for selected net taxable estates are as follows:

Net Taxable Estate ($)	Approximate Probate Fees and Taxes ($)
600,000	30,000
750,000	92,500
1,000,000	203,000
1,500,000	438,000
2,000,000	688,000
2,500,000	1,037,000
3,000,000	1,350,000

Table 12.34

a. Decide which variable should be the independent variable and which should be the dependent variable.

b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is it significant?
f. Find the estimated total cost for a net taxable estate of $1,000,000. Find the cost for $2,500,000.
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Are there any outliers in the data?
i. Based on these results, what would be the probate fees and taxes for an estate that does not have any assets?
j. What is the slope of the least-squares (best-fit) line? Interpret the slope.

78. The following are advertised sale prices of color televisions at Anderson's:

Size (inches)	Sale Price ($)
9	147
20	197
27	297
31	447
35	1,177
40	2,177
60	2,497

Table 12.35

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is it significant?
f. Find the estimated sale price for a 32-inch television. Find the cost for a 50-inch television.
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Are there any outliers in the data?
i. What is the slope of the least-squares (best-fit) line? Interpret the slope.

79. Table 12.36 shows the average heights for American boys in 1990.

Age (years)	Height (centimeters)
Birth	50.8
2	83.8
3	91.4
5	106.6
7	119.3

Table 12.36

Age (years)	Height (centimeters)
10	137.1
14	157.5

Table 12.36

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is it significant?
f. Find the estimated average height for a 1-year-old. Find the estimated average height for an 11-year-old.
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Are there any outliers in the data?
i. Use the least-squares line to estimate the average height for a 62-year-old man. Do you think that your answer is reasonable? Why or why not?
j. What is the slope of the least-squares (best-fit) line? Interpret the slope.

80.

State	No. of Letters in Name	Year Entered the Union	Rank for Entering the Union	Area (square miles)
Alabama	7	1819	22	52,423
Colorado	8	1876	38	104,100
Hawaii	6	1959	50	10,932
Iowa	4	1846	29	56,276
Maryland	8	1788	7	12,407
Missouri	8	1821	24	69,709
New Jersey	9	1787	3	8,722
Ohio	4	1803	17	44,828
South Carolina	13	1788	8	32,008
Utah	4	1896	45	84,904
Wisconsin	9	1848	30	65,499

Table 12.37

We are interested in whether there is a relationship between the ranking of a state and the area of the state.
a. What are the independent and dependent variables?
b. What do you think the scatter plot will look like? Make a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
e. Find the correlation coefficient. What does it imply about the significance of the relationship?

f. Find the estimated areas for Alabama and for Colorado. Are they close to the actual areas?
g. Use the two points in Part F to plot the least-squares line on your graph from Part B.
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Are there any outliers?
j. Use the least-squares line to estimate the area of a new state that enters the Union. Can the least-squares line be used to predict it? Why or why not?
k. Delete *Hawaii* and substitute *Alaska* for it. Alaska is a state with an area of 656,424 square miles.
l. Calculate the new least-squares line.
m. Find the estimated area for Alabama. Is it closer to the actual area with this new least-squares line or with the previous one that included Hawaii? Why do you think that's the case?
n. Do you think that, in general, newer states are larger than the original states?

References

12.1 Linear Equations

Centers for Disease Control and Prevention. (n.d.). Retrieved from https://www.cdc.gov/

National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention. (n.d.). Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/nchhstp/default.htm

12.4 Prediction (Optional)

Centers for Disease Control and Prevention. (n.d.). Retrieved from https://www.cdc.gov/

National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention. (n.d.). Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/nchhstp/default.htm

National Center for Health Statistics. (n.d.). Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/nchs/index.htm

U.S. Census Bureau. (n.d.). Retrieved from http://www.census.gov/compendia/statab/cats/transportation/motor_vehicle_accidents_and_fatalities.html

12.5 Outliers

Committee on Ways and Means, U.S. House of Representatives. (n.d.). Washington, DC: U.S. Department of Health and Human Services.

Microsoft Bookshelf. (n.d.).

Physician's Desk Reference Staff. (1990). *Physician's desk reference*. Ohio: Medical Economics Company.

U.S. Bureau of Labor Statistics. (n.d.). Retrieved from https://www.bls.gov/

Solutions

1. dependent variable: fee amount
independent variable: time

3.

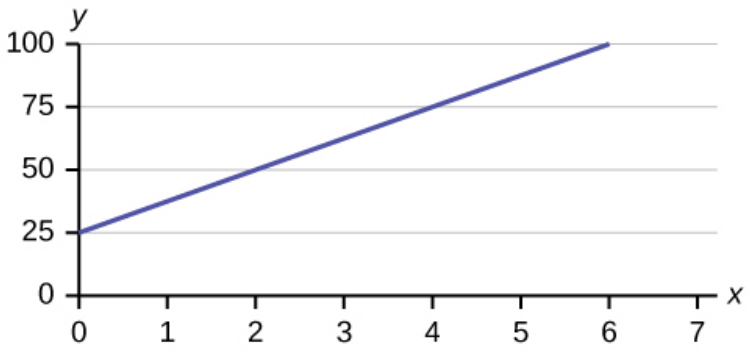


Figure 12.25

5.

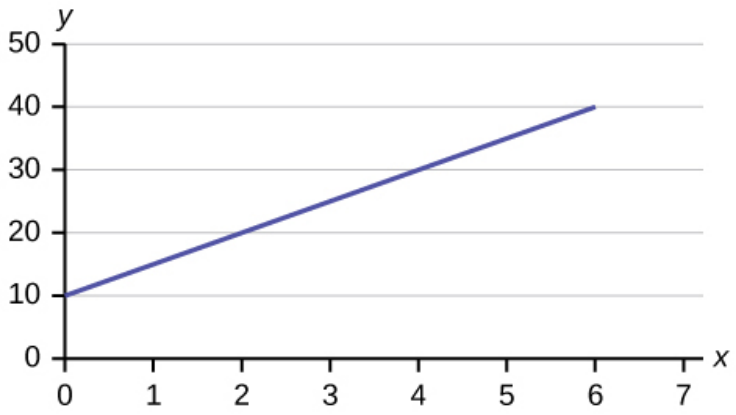


Figure 12.26

7. $y = 6x + 8$, $4y = 8$, and $y + 7 = 3x$ are all linear equations.

9. The number of flu cases depends on the year. Therefore, year becomes the independent variable and the number of flu cases is the dependent variable.

11. The y-intercept is 50 ($a = 50$). At the start of the cleaning, the company charges a one-time fee of $50 (this is when $x = 0$). The slope is 100 ($b = 100$). For each session, the company charges $100 for each hour they clean.

13. 12,000 lb of soil

15. The slope is –1.5 ($b = –1.5$). This means the stock is losing value at a rate of $1.50 per hour. The y-intercept is $15 ($a = 15$). This means the price of stock before the trading day was $15.

17.

Group	x (no. of hours spent studying)	y (final exam grades)	Median x value	Median y value
1	1 2 3	45 50 51	2	50
2	4 5	65 72	4.5	68.5
3	6 7 8	80 90 96	7	90

Table 12.38

19. $\hat{y} = 2.23 + 1.99x$

21. The slope is 1.99 ($b = 1.99$). It means that for every endorsement deal a professional player gets, he gets an average of another $1.99 million in pay each year.

23. It means that there is no correlation between the data sets.

25. Yes. There are enough data points and the value of r is strong enough to show there is a strong negative correlation between the data sets.

27. $H_a: \rho \neq 0$

29. $250,120

31. 1326 acres

33. 1125 hours, or when $x = 1125$

35. Check student solution.

36. a. When $x = 1985$, $\hat{y} = 25{,}52$.
b. When $x = 1990$, $\hat{y} = 34{,}275$.
c. When $x = 1970$, $\hat{y} = -725$. Why doesn't this answer make sense? The range of x values was 1981 to 2002; the year 1970 is not in this range. The regression equation does not apply, because predicting for the year 1970 is extrapolation, which requires a different process. Also, a negative number does not make sense in this context, when we are predicting flu cases diagnosed.

38. Also, the correlation $r = 0.4526$. If r is compared with the value in the 95 Percent Critical Values of the Sample Correlation Coefficient Table, because $r > 0.423$, r is significant, and you would think that the line could be used for prediction. But, the scatter plot indicates otherwise.

39. Check student' solution.

40. $\hat{y} = 3{,}448{,}225 + 1750x$

42. There was an increase in flu cases diagnosed until 1993. From 1993 through 2002, the number of flu cases diagnosed declined each year. It is not appropriate to use a linear regression line to fit to the data.

44. Because there is no linear association between year and number of flu cases diagnosed, it is not appropriate to calculate a linear correlation coefficient. When there is a linear association and it is appropriate to calculate a correlation, we cannot say that one variable *causes* the other variable.

46. We don't know if the pre-1981 data were collected from a single year. So, we don't have an accurate x value for this figure.
Regression equation: $\hat{y}$ (number of flu cases) = –3,448,225 + 1749.777 (year).

	Coefficients
Intercept	–3,448,225
x Variable 1	1,749.777

Table 12.39

47.
- $a = -3{,}488{,}225$
- $b = 1{,}750$
- correlation = 0.4526
- $n = 22$

48. No, he is not correct. An outlier is only an influential point if it significantly impacts the slope of the least-squares regression line and the correlation coefficient, r. If omission of this data point from the calculation of the regression line does not show much impact on the slope or r-value, then the outlier is not considered an influential point. For different reasons, it still may be determined that the data point must be omitted from the data set.

49. Yes. There appears to be an outlier at (6, 58).

51. The potential outlier flattened the slope of the line of best fit because it was below the data set. It made the line of best fit less accurate as a predictor for the data.

53. $s = 1.75$

55. a. independent variable: age; dependent variable: fatalities
b. independent variable: number of family members; dependent variable: grocery bill
c. independent variable: age of applicant; dependent variable: insurance premium
d. independent variable: power consumption; dependent variable: utility
e. independent variable: higher education (years); dependent variable: crime rates

58. It means that 72 percent of the variation in the dependent variable (y) can be explained by the variation in the independent variable (x).

60.

x (SAT math scores)	y (GPAs)
261	50
262	60
327	62
350	65
363	70
364	67
373	71
544	86
587	87
624	90
741	98

Table 12.40

We must remember to check the order of the y values within each group as well. We notice that the y values in the second group are not in order from the least value to the greatest value; these values thus must be reordered, meaning the median y value for that group is 70.

Group	*x* (SAT math scores)	*y* (GPAs)	Median *x* value	Median *y* value
1	261 262 327 350	50 60 62 65	294.5	61
2	363 364 373	67 70 71	364	70
3	544 587 624 741	86 87 90 98	605.5	88.5

Table 12.41

The ordered pairs are (294.5, 61), (364, 70), and (605.5, 88.5).
The slope can be calculated using the formula $m = \frac{y_3 - y_1}{x_3 - x_1}$. Substituting the median *x* and *y* values, from the first and third groups gives $m = \frac{88.5 - 61}{605.5 - 294.5}$, which simplifies to $m \approx 0.09$.
The y-intercept may be found using the formula $b = \frac{\sum y - m \sum x}{3}$. The sum of the median *x* values is 1264, and the sum of the median *y* values is 219.5. Substituting these sums and the slope into the formula gives $b = \frac{219.5 - 0.09(1264)}{3}$, which simplifies to $b \approx 35.25$.
The line of best fit is represented as $y = mx + b$. Thus, the equation can be written as $y = 0.09x + 35.25$.

61. b. Check student solution.

c. $\hat{y} = 35.5818045 - 0.19182491x$

d. $r = -0.57874$
For four degrees of freedom and alpha = 0.05, the LinRegTTest gives a *p* value of 0.2288, so we do not reject the null hypothesis; there is not a significant linear relationship between deaths and age.
Using the table of critical values for the correlation coefficient, with four degrees of freedom, the critical value is 0.811. The correlation coefficient $r = -0.57874$ is not less than –0.811, so we do not reject the null hypothesis.

f. There is not a linear relationship between the two variables, as evidenced by a *p* value greater than 0.05.

63. a. We wonder if the better discounts appear earlier in the book, so we select page as *x* and discount as *y*.

b. Check student solution.

c. $\hat{y} = 17.21757 - 0.01412x$

d. $r = -0.2752$
For seven degrees of freedom and alpha = 0.05, LinRegTTest gives a *p* value = 0.4736, so we do not reject; there is a not a significant linear relationship between page and discount.
Using the table of critical values for the correlation coefficient, with seven gives degrees of freedom, the critical value is 0.666. The correlation coefficient $xi = -0.2752$ is not less than 0.666, so we do not reject.

f. There is not a significant linear correlation so it appears there is no relationship between the page and the amount of the discount.
As the page number increases by one page, the discount decreases by $0.01412.

65. a. Year is the independent or *x* variable; the number of letters is the dependent or *y* variable.

b. Check student's solution.

c. No.

d. $\hat{y} = 47.03 - 0.0216x$

e. –0.4280. The *r* value indicates that there is not a significant correlation between the year the state entered the Union and the number of letters in the name.

g. No. The relationship does not appear to be linear; the correlation is not significant.

66. Using LinRegTTest, the output for the original least-squares regression line is $y = 26.14 + 0.7539x$ and $r = 0.6657$. The output for the new least-squares regression line, after omitting the outlier of (56, 95), is $\hat{y} = 6.36 + 1.0045x$ and $r = 0.9757$.
The slope of the new line is quite a bit different from the slope of the original least-squares regression line, but the larger change is shown in the *r*-values, such that the new line has an *r*-value that has increased to a value that is almost equal to one.
Thus, it may be stated that the outlier (56, 95) is also an influential point.

68. a. and b. Check student solution.
c. The slope of the regression line is –0.3031 with a *y*-intercept of 31.93. In context, the *y*-intercept indicates that when there are no returning sparrow hawks, there will be almost 32 percent new sparrow hawks, which doesn't make sense, because if there are no returning birds, then the new percentage would have to be 100% (this is an example of why we do not extrapolate). The slope tells us that for each percentage increase in returning birds, the percentage of new birds in the colony decreases by 30.3 percent.
d. If we examine r_2, we see that only 57.52 percent of the variation in the percentage of new birds is explained by the model and the correlation coefficient, $r = -.7584$ only indicates a somewhat strong correlation between returning and new percentages.
e. The ordered pair (66, 6) generates the largest residual of 6.0. This means that when the observed return percentage is 66 percent, our observed new percentage, 6 percent, is almost 6 percent less than the predicted new value of 11.98 percent. If we remove this data pair, we see only an adjusted slope of –0.2789 and an adjusted intercept of 30.9816. In other words, although these data generate the largest residual, it is not an outlier, nor is the data pair an influential point.
f. If there are 70 percent returning birds, we would expect to see $y = -0.2789(70) + 30.9816 = 0.114$ or 11.4 percent new birds in the colony.

70.
a. Check student solution.
b. Check student solution.
c. We have a slope of –1.4946 with a *y*-intercept of 193.88. The slope, in context, indicates that for each additional minute added to the swim time, the heart rate decreases by 1.5 beats per minute. If the student is not swimming at all, the *y*-intercept indicates that his heart rate will be 193.88 beats per minute. Although the slope has meaning (the longer it takes to swim 2000 m, the less effort the heart puts out), the *y*-intercept does not make sense. If the athlete is not swimming (resting), then his heart rate should be very low.
d. Because only 1.5 percent of the heart rate variation is explained by this regression equation, we must conclude that this association is not explained with a linear relationship.
e. Point (34.72, 124) generates the largest residual: –11.82. This means that our observed heart rate is almost 12 beats less than our predicted rate of 136 beats per minute. When this point is removed, the slope becomes –2.953, with the *y*-intercept changing to 247.1616. Although the linear association is still very weak, we see that the removed data pair can be considered an influential point in the sense that the *y*-intercept becomes more meaningful.

72. If we remove the two service academies (the tuition is $0.00), we construct a new regression equation of $y = -0.0009x + 160$, with a correlation coefficient of 0.71397 and a coefficient of determination of 0.50976. This allows us to say there is a fairly strong linear association between tuition costs and salaries if the service academies are removed from the data set.

73. c. No. The *y*-intercept would occur at year 0, which doesn't exist.

74. a. Check student's solution.
b. Yes.
c. No, the *y*-intercept would occur at year 0, which doesn't exist.
d. $\hat{y} = -266.8863 + 0.1656x$.
e. 0.9448, yes.
f. 62.8233, 62.3265.
g. Yes.
h. No, (1987, 62.7).
i. 72.5937, no.
j. Slope = 0.1656. As the year increases by one, the percent of workers paid hourly rates tends to increase by 0.1656.

76. a.

Size (ounces)	Cost ($)	Cost per ounce
16	3.99	24.94
32	4.99	15.59
64	5.99	9.36
200	10.99	5.50

Table 12.42

b. Check student solution.
c. There is a linear relationship for the sizes 16 through 64, but that linear trend does not continue to the 200-oz size.
d. $\hat{y} = 20.2368 - 0.0819x$
e. $r = -.8086$
f. 40-oz: 16.96 cents/oz
g. 90-oz: 12.87 cents/oz
h. The relationship is not linear; the least-squares line is not appropriate.
i. There are no outliers.
j. No. You would be extrapolating. The 300-oz size is outside the range of *x*.
k. X = –0.08194. For each additional ounce in size, the cost per ounce decreases by 0.082 cents.

78. a. Size is *x*, the independent variable, and price is *y*, the dependent variable.
b. Check student solution.
c. The relationship does not appear to be linear.
d. $\hat{y} = -745.252 + 54.75569x$.
e. $r = .8944$ and yes, it is significant.
f. 32-inch: $1006.93, 50-inch: $1992.53.
g. No, the relationship does not appear to be linear. However, *r* is significant.
h. No, the 60-inch TV.
i. For each additional inch, the price increases by $54.76.

80. a. Rank is the independent variable and area is the dependent variable.
b. Check student solution.
c. There appears to be a linear relationship, with one outlier.
d. $\hat{y}$(area) = 24177.06 + 1010.478*x*
e. $r = .50047$. *r* is not significant, so there is no relationship between the variables.
f. Alabama: 46,407.576 square miles, Colorado: 62,575.224 square miles.
g. The Alabama estimate is closer than the Colorado estimate.

h. If the outlier is removed, there is a linear relationship.
i. There is one outlier (Hawaii).
j. rank 51: 75,711.4 square miles, no.
k.

Alabama	7	1819	22	52,423
Colorado	8	1876	38	104,100
Hawaii	6	1959	50	10,932
Iowa	4	1846	29	56,276
Maryland	8	1788	7	12,407
Missouri	8	1821	24	69,709
New Jersey	9	1787	3	8,722
Ohio	4	1803	17	44,828
South Carolina	13	1788	8	32,008
Utah	4	1896	45	84,904
Wisconsin	9	1848	30	65,499

Table 12.43

l. $\hat{y} = -87065.3 + 7828.532x$.
m. Alabama: 85,162.404; the prior estimate was closer. Alaska is an outlier.
n. Yes, with the exception of Hawaii.

13

F DISTRIBUTION AND ONE-WAY ANOVA

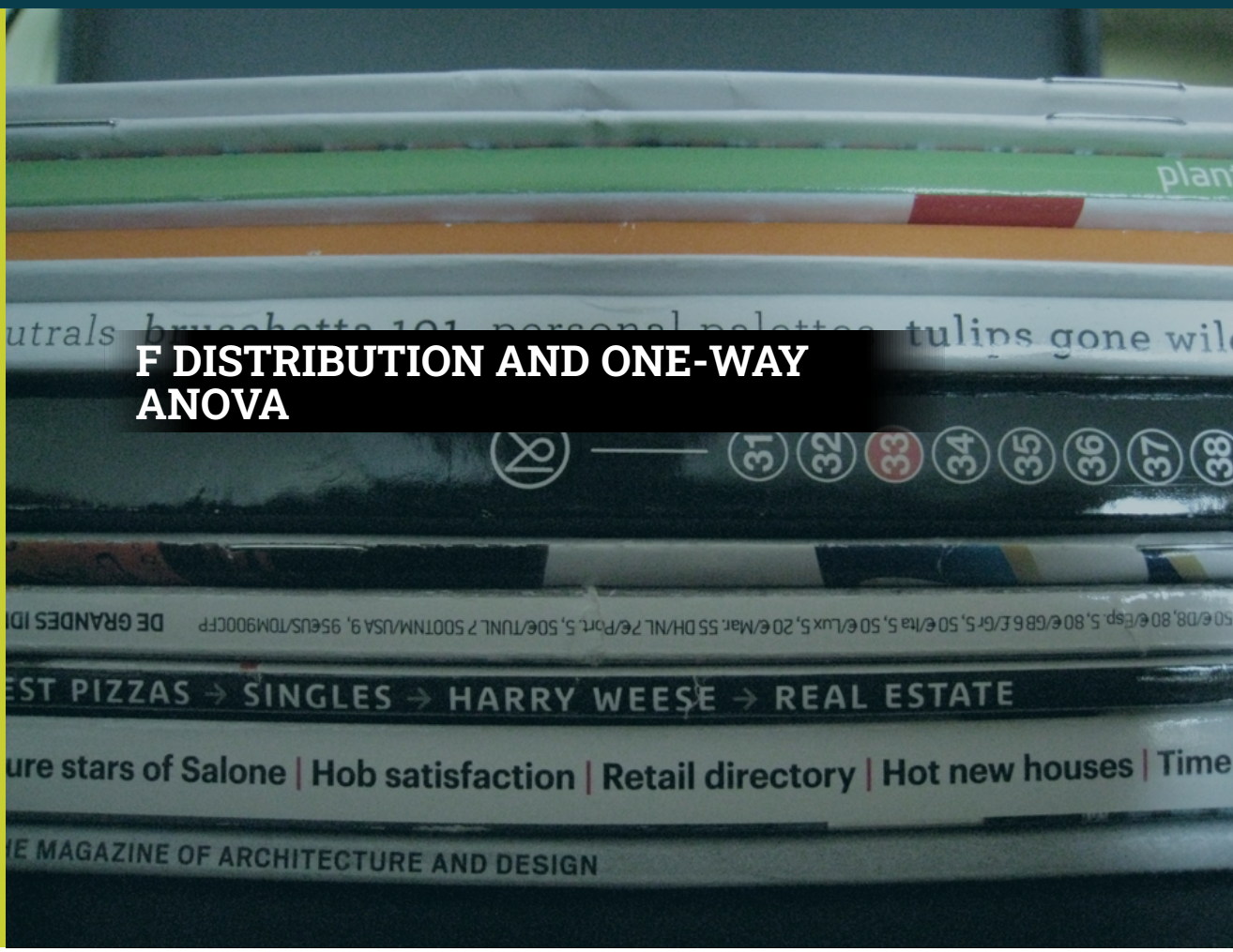


Figure 13.1 One-way ANOVA is used to measure information from several groups.

Chapter Objectives

By the end of this chapter, the student should be able to do the following:

- Interpret the *F* probability distribution as the number of groups and the sample size change
- Discuss two uses for the *F* distribution: one-way ANOVA and the test of two variances
- Conduct and interpret one-way ANOVA
- Conduct and interpret hypothesis tests of two variances

Introduction

Many statistical applications in psychology, social science, business administration, and the natural sciences involve several groups. For example, an environmentalist is interested in knowing if the average amount of pollution varies among several bodies of water. A sociologist is interested in knowing if the amount of income a person earns varies according to his or her upbringing. A consumer looking for a new car might compare the average gas mileage of several models.

For hypothesis tests comparing averages across more than two groups, statisticians have developed a method called *analysis of variance* (abbreviated ANOVA). In this chapter, you will study the simplest form of ANOVA called single factor or one-way ANOVA. You will also study the *F* distribution, used for one-way ANOVA, and the test of two variances. This is a very brief overview of one-way ANOVA. You will study this topic in much greater detail in future statistics courses. One-way ANOVA, as it is presented here, relies heavily on a calculator or computer.

13.1 One-Way ANOVA

The purpose of a one-way ANOVA test is to determine the existence of a statistically significant difference among several group means. The test uses **variances** to help determine if the means are equal or not. To perform a one-way ANOVA test, there are five basic assumptions to be fulfilled:

- Each population from which a sample is taken is assumed to be normal.
- All samples are randomly selected and independent.
- The populations are assumed to have equal standard deviations (or variances).
- The factor is a categorical variable.
- The response is a numerical variable.

The Null and Alternative Hypotheses

The null hypothesis is that all the group population means are the same. The alternative hypothesis is that at least one pair of means is different. For example, if there are k groups

H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$

H_a: At least two of the group means $\mu_1, \mu_2, \mu_3, \ldots, \mu_k$ are not equal. That is, $\mu_i \neq \mu_j$ for some $i \neq j$.

The graphs, a set of box plots representing the distribution of values with the group means indicated by a horizontal line through the box, help in the understanding of the hypothesis test. In the first graph (red box plots), H_0: $\mu_1 = \mu_2 = \mu_3$ and the three populations have the same distribution if the null hypothesis is true. The variance of the combined data is approximately the same as the variance of each of the populations.

If the null hypothesis is false, then the variance of the combined data is larger, which is caused by the different means as shown in the second graph (green box plots).

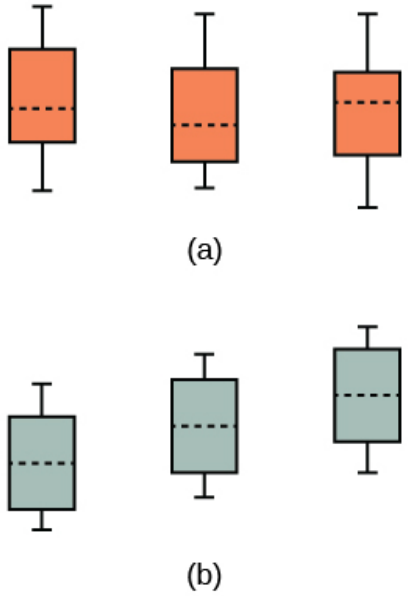


Figure 13.2 (a) We fail to reject H_0 as it may be true. All the means are about the same; the differences may be due to random variation. (b) We reject H_0 as all the means are not the same; the differences are too large to be due to random variation.

13.2 The F Distribution and the F Ratio

The distribution used for the hypothesis test is a new one. It is called the ***F* distribution**, named after Sir Ronald Fisher, an English statistician. The F statistic is a ratio (a fraction). There are two sets of degrees of freedom: one for the numerator and one for the denominator.

For example, if F follows an F distribution and the number of degrees of freedom for the numerator is 4, and the number of degrees of freedom for the denominator is 10, then $F \sim F_{4,10}$.

Note

The *F* distribution is derived from the Student's *t*-distribution. The values of the *F* distribution are squares of the corresponding values of the *t*-distribution. One-way ANOVA expands the *t*-test for comparing more than two groups. The scope of that derivation is beyond the level of this course. It is preferable to use ANOVA when there are more than two groups instead of performing pairwise *t*-tests because performing multiple tests introduces the likelihood of making a Type 1 error.

To calculate the ***F* ratio**, two estimates of the variance are made.

1. Variance between samples: an estimate of σ^2 that is the variance of the sample means multiplied by n, when the sample sizes are the same. If the samples are different sizes, the variance between samples is weighted to account for the different sample sizes. The variance is also called *variation due to treatment* or *explained variation.*
2. Variance within samples: an estimate of σ^2 that is the average of the sample variances, also known as a *pooled variance*. When the sample sizes are different, the variance within samples is weighted. The variance is also called the *variation due to error* or *unexplained variation.*

- $SS_{between}$ = the sum of squares that represents the variation among the different samples
- SS_{within} = the sum of squares that represents the variation within samples that is due to chance

To find a *sum of squares* mean, add together squared quantities which, in some cases, may be weighted. We used sum of squares to calculate the sample variance and the sample standard deviation in Descriptive Statistics.

MS means *mean square*. $MS_{between}$ is the variance between groups, and MS_{within} is the variance within groups.

Calculation of Sum of Squares and Mean Square

- k = the number of different groups
- n_j = the size of the j^{th} group
- s_j = the sum of the values in the j^{th} group
- n = total number of all the values combined (total sample size: $\sum n_j$)
- x = one value: $\sum x = \sum s_j$
- Sum of squares of all values from every group combined: $\sum x^2$
- Between group variability: $SS_{total} = \sum x^2 - \frac{\left(\sum x^2\right)}{n}$
- Total sum of squares: $\sum x^2 - \frac{(\sum x)^2}{n}$
- Explained variation: sum of squares representing variation among the different samples
  $SS_{(between)} = \sum \left[\frac{(s_j)^2}{n_j}\right] - \frac{\left(\sum s_j\right)^2}{n}$
- Unexplained variation: sum of squares representing variation within samples due to chance
  $SS_{within} = SS_{total} - SS_{between}$
- *df*s for different groups (*df*s for the numerator): $df = k - 1$
- Equation for errors within samples (*df*s for the denominator): $df_{within} = n - k$
- Mean square (variance estimate) explained by the different groups: $MS_{between} = \frac{SS_{between}}{df_{between}}$
- Mean square (variance estimate) that is due to chance (unexplained): $MS_{within} = \frac{SS_{within}}{df_{within}}$

$MS_{between}$ and MS_{within} can be written as follows:

- $MS_{between} = \frac{SS_{between}}{df_{between}} = \frac{SS_{between}}{k-1}$
- $MS_{within} = \frac{SS_{within}}{df_{within}} = \frac{SS_{within}}{n-k}$

The one-way ANOVA test depends on the fact that $MS_{between}$ can be influenced by population differences among means of the several groups. Since MS_{within} compares values of each group to its own group mean, the fact that group means might be different does not affect MS_{within}.

The null hypothesis says that all groups are samples from populations having the same normal distribution. The alternate hypothesis says that at least two of the sample groups come from populations with different normal distributions. If the null hypothesis is true, $MS_{between}$ and MS_{within} should both estimate the same value.

Note

The null hypothesis says that all the group population means are equal. The hypothesis of equal means implies that the populations have the same normal distribution because it is assumed that the populations are normal and that they have equal variances.

***F* Ratio or *F* Statistic**

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

If MS_{between} and MS_{within} estimate the same value, following the belief that H_0 is true, then the F ratio should be approximately equal to 1. Mostly, just sampling errors would contribute to variations away from 1. As it turns out, MS_{between} consists of the population variance plus a variance produced from the differences between the samples. MS_{within} is an estimate of the population variance. Since variances are always positive, if the null hypothesis is false, MS_{between} will generally be larger than MS_{within}. Then the F ratio will be larger than 1. However, if the population effect is small, it is not unlikely that MS_{within} will be larger in a given sample.

The previous calculations were done with groups of different sizes. If the groups are the same size, the calculations simplify somewhat and the F ratio can be written as follows:

***F* Ratio formula when the groups are the same size**

$$F = \frac{n \cdot s_{\bar{x}}^2}{s^2_{\text{pooled}}}$$

where

- n = the sample size
- $df_{\text{numerator}} = k - 1$
- $df_{\text{denominator}} = n - k$
- s^2 pooled = the mean of the sample variances (pooled variance)
- $s_{\bar{x}}^2$ = the variance of the sample means

Data is typically put into a table for easy viewing. One-way ANOVA results are often displayed in this manner by computer software.

Source of Variation	Sum of Squares (*SS*)	Degrees of Freedom (*df*)	Mean Square (*MS*)	*F*
Factor (Between)	*SS*(Factor)	$k - 1$	*MS*(Factor) = *SS*(Factor)/(k – 1)	*F* = *MS*(Factor)/*MS*(Error)
Error (Within)	*SS*(Error)	$n - k$	*MS*(Error) = *SS*(Error)/(n – k)	
Total	*SS*(Total)	$n - 1$		

Table 13.1

EXAMPLE 13.1

Three different diet plans are to be tested for mean weight loss. The entries in the table are the weight losses for the different plans. The one-way ANOVA results are shown in Table 13.2.

Plan 1: $n_1 = 4$	Plan 2: $n_2 = 3$	Plan 3: $n_3 = 3$
5	3.5	8
4.5	7	4
4		3.5
3	4.5	

Table 13.2

$s_1 = 16.5$, $s_2 = 15$, $s_3 = 15.5$

Following are the calculations needed to fill in the one-way ANOVA table. The table is used to conduct a hypothesis test.

$$SS(between) = \sum \left[\frac{(s_j)^2}{n_j} \right] - \frac{\left(\sum s_j\right)^2}{n} \quad 13.1$$

$$= \frac{s_1^2}{4} + \frac{s_2^2}{3} + \frac{s_3^2}{3} - \frac{(s_1 + s_2 + s_3)^2}{10} \quad 13.2$$

where $n_1 = 4$, $n_2 = 3$, $n_3 = 3$, and $n = n_1 + n_2 + n_3 = 10$

$$= \frac{(16.5)^2}{4} + \frac{(15)^2}{3} + \frac{(15.5)^2}{3} - \frac{(16.5 + 15 + 15.5)^2}{10} \quad 13.3$$

$$SS(between) = 2.2458 \quad 13.4$$

$$S(total) = \sum x^2 - \frac{\left(\sum x\right)^2}{n} \quad 13.5$$

$$= \left(5^2 + 4.5^2 + 4^2 + 3^2 + 3.5^2 + 7^2 + 4.5^2 + 8^2 + 4^2 + 3.5^2\right) \quad 13.6$$

$$- \frac{(5 + 4.5 + 4 + 3 + 3.5 + 7 + 4.5 + 8 + 4 + 3.5)^2}{10} \quad 13.7$$

$$= 244 - \frac{47^2}{10} = 244 - 220.9 \quad 13.8$$

$$SS(total) = 23.1 \quad 13.9$$

$$SS(within) = SS(total) - SS(between) \quad 13.10$$

$$= 23.1 - 2.2458 \quad 13.11$$

$$SS(within) = 20.8542 \quad 13.12$$

USING THE TI-83, 83+, 84, 84+ CALCULATOR

One-way ANOVA Table: The formulas for *SS*(Total), *SS*(Factor) = *SS*(Between), and *SS*(Error) = *SS*(Within) as shown previously. The same information is provided by the TI calculator hypothesis test function ANOVA in STAT TESTS (syntax is ANOVA[L1, L2, L3] where L1, L2, L3 have the data from Plan 1, Plan 2, Plan 3, respectively).

Source of Variation	Sum of Squares (*SS*)	Degrees of Freedom (*df*)	Mean Square (*MS*)	*F*
Factor (Between)	*SS*(Factor) = *SS*(Between) = 2.2458	$k - 1$ = 3 groups – 1 = 2	*MS*(Factor) = *SS*(Factor)/(k – 1) = 2.2458/2 = 1.1229	*F* = *MS*(Factor)/*MS*(Error) = 1.1229/2.9792 = 0.3769
Error (Within)	*SS*(Error) = *SS*(Within) = 20.8542	$n - k$ = 10 total data – 3 groups = 7	*MS*(Error) = *SS*(Error)/(n – k) = 20.8542/7 = 2.9792	
Total	*SS*(Total) = 2.2458 + 20.8542 = 23.1	$n - 1$ = 10 total data – 1 = 9		

Table 13.3

TRY IT 13.1

As part of an experiment to see how different types of soil cover would affect slicing tomato production, Marist College students grew tomato plants under different soil cover conditions. Groups of three plants each had one of the following treatments:

- Bare soil
- A commercial ground cover
- Black plastic
- Straw
- Compost

All plants grew under the same conditions and were the same variety. Students recorded the weight in grams of tomatoes produced by each of the $n = 15$ plants, as seen in Table 13.4.

Bare: $n_1 = 3$	Ground Cover: $n_2 = 3$	Plastic: $n_3 = 3$	Straw: $n_4 = 3$	Compost: $n_5 = 3$
2,625	5,348	6,583	7,285	6,277
2,997	5,682	8,560	6,897	7,818
4,915	5,482	3,830	9,230	8,677

Table 13.4

Create the one-way ANOVA table.

The one-way ANOVA hypothesis test is always right-tailed because larger *F* values are way out in the right tail of the *F* distribution curve and tend to make us reject H_0.

Notation

The notation for the F distribution is $F \sim F_{df(num),df(denom)}$,

where $df(num) = df_{between}$ and $df(denom) = df_{within}$.

The mean for the F distribution is $\mu = \frac{df(\text{denom})}{df(\text{denom})-2}$.

13.3 Facts About the F Distribution

The following are facts about the F distribution:

- The curve is not symmetrical but skewed to the right.
- There is a different curve for each set of dfs.
- The F statistic is greater than or equal to zero.
- As the degrees of freedom for the numerator and for the denominator get larger, the curve approximates the normal.
- Other uses for the F distribution include comparing two variances and two-way analysis of variance. Two-way analysis is beyond the scope of this chapter.

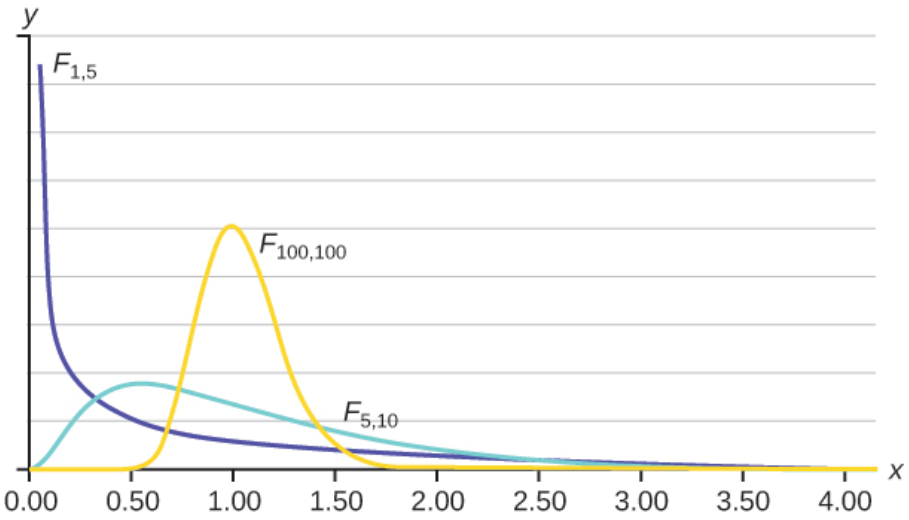


Figure 13.3

EXAMPLE 13.2

Problem

Let's return to the slicing tomato exercise in Try It 13.1. The means of the tomato yields under the five mulching conditions are represented by μ_1, μ_2, μ_3, μ_4, μ_5. We will conduct a hypothesis test to determine if all means are the same or at least one is different. Using a significance level of 5 percent, test the null hypothesis that there is no difference in mean yields among the five groups against the alternative hypothesis that at least one mean is different from the rest.

Solution

The null and alternative hypotheses are as follows:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H_a: $\mu_i \neq \mu_j$ for some $i \neq j$

The one-way ANOVA results are shown in Table 13.5

Source of Variation	Sum of Squares (*SS*)	Degrees of Freedom (*df*)	Mean Square (*MS*)	*F*
Factor (Between)	36,648,561	5 – 1 = 4	$\frac{36,648,561}{4} = 9,162,140$	$\frac{9,162,140}{2,044,672.6} = 4.4810$
Error (Within)	20,446,726	15 – 5 = 10	$\frac{20,446,726}{10} = 2,044,672.6$	
Total	57,095,287	15 – 1 = 14		

Table 13.5

Distribution for the test: $F_{4,10}$

df(*num*) = 5 – 1 = 4

df(*denom*) = 15 – 5 = 10

Test statistic: *F* = 4.4810

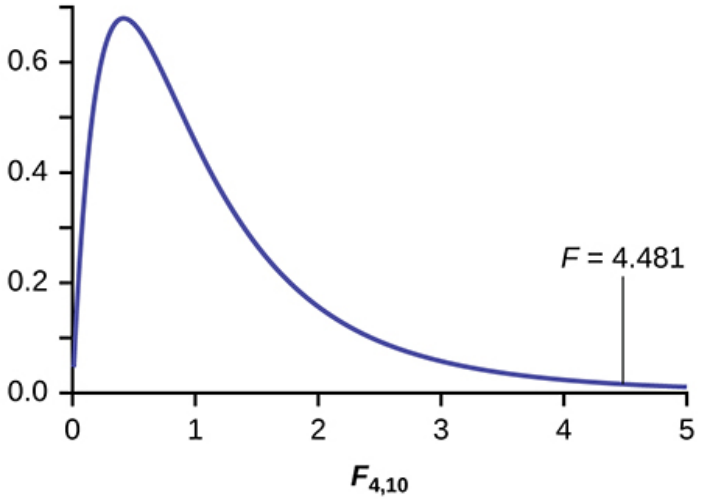


Figure 13.4

Probability statement: *p*-value = *P*(*F* > 4.481) = 0.0248

Compare *α* and the *p*-value: $\alpha = 0.05$, *p*-value = 0.0248

Make a decision: Since $\alpha >$ *p*-value, we reject H_0.

Conclusion: At the 5 percent significance level, we have reasonably strong evidence that differences in mean yields for slicing tomato plants grown under different mulching conditions are unlikely to be due to chance alone. We may conclude that at least some of the mulches led to different mean yields.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

To find these results on the calculator:

Press STAT. Press `1:EDIT`. Put the data into the lists `L1,L2,L3,L4,L5`.

Press STAT, arrow over to TESTS, and arrow down to `ANOVA`. Press ENTER, and then enter (`L1,L2,L3,L4,L5`). Press ENTER. You will see that the values in the foregoing ANOVA table are easily produced by the calculator, including the test statistic and the *p*-value of the test.

The calculator displays:
$F = 4.4810$
$p = 0.0248$ (p-value)
Factor
$df = 4$
$SS = 36648560.9$
$MS = 9162140.23$
Error
$df = 10$
$SS = 20446726$
$MS = 2044672.6$

TRY IT 13.2

MRSA, or *Staphylococcus aureus*, can cause serious bacterial infections in hospital patients. Table 13.6 shows various colony counts from different patients who may or may not have MRSA. The data from the table is plotted in Figure 13.5.

Conc = 0.6	Conc = 0.8	Conc = 1.0	Conc = 1.2	Conc = 1.4
9	16	22	30	27
66	93	147	199	168
98	82	120	148	132

Table 13.6

Plot of the data for the different concentrations:

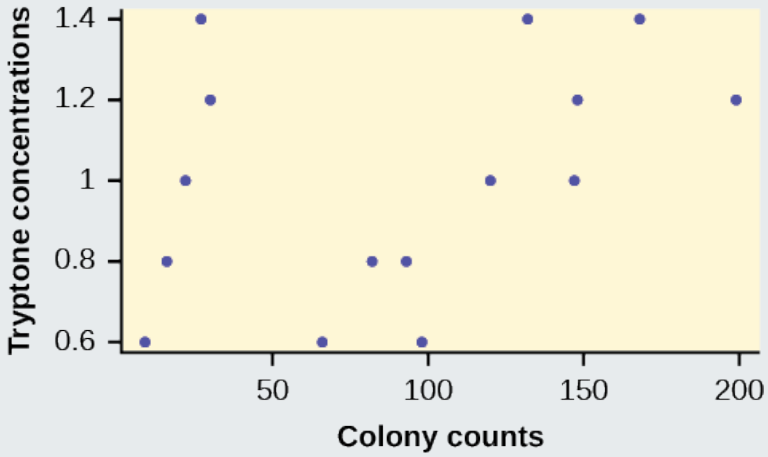


Figure 13.5

Test whether the mean numbers of colonies are the same or are different. Construct the ANOVA table by hand or by using a TI-83, 83+, or 84+ calculator, find the p-value, and state your conclusion. Use a 5 percent significance level.

EXAMPLE 13.3

Four sororities took a random sample of sisters regarding their grade means for the past term. The results are shown in Table 13.7.

Sorority 1	Sorority 2	Sorority 3	Sorority 4
2.17	2.63	2.63	3.79
1.85	1.77	3.78	3.45
2.83	3.25	4.00	3.08
1.69	1.86	2.55	2.26
3.33	2.21	2.45	3.18

Table 13.7 Mean Grades for Four Sororities

Problem

Using a significance level of 1 percent, is there a difference in mean grades among the sororities?

Solution

Let μ_1, μ_2, μ_3, μ_4 be the population means of the sororities. Remember that the null hypothesis claims that the sorority groups are from the same normal distribution. The alternate hypothesis says that at least two of the sorority groups come from populations with different normal distributions. Notice that the four sample sizes are each five.

Note

This is an example of a *balanced design*, because each factor (i.e., sorority) has the same number of observations.

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a: Not all of the means μ_1, μ_2, μ_3, μ_4 are equal.

Distribution for the test: $F_{3,16}$

where $k = 4$ groups and $n = 20$ samples in total.

$df(num) = k - 1 = 4 - 1 = 3$

$df(denom) = n - k = 20 - 4 = 16$

Calculate the test statistic: $F = 2.23$

Graph

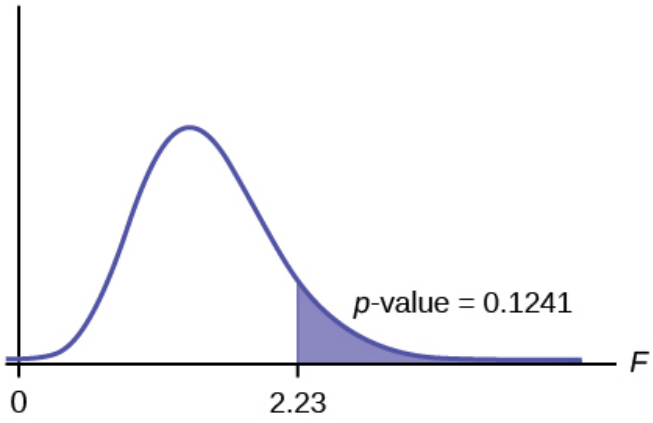


Figure 13.6

Probability statement: p-value = $P(F > 2.23) = 0.1241$

Compare α and the p-value: $\alpha = 0.01$
p-value = 0.1241
$\alpha < p$-value

Make a decision: Since $\alpha < p$-value, we cannot reject H_0.

Conclusion: There is not sufficient evidence to conclude that there is a difference among the mean grades for the sororities.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Put the data into lists L_1, L_2, L_3, and L_4. Press STAT and arrow over to TESTS. Arrow down to F:ANOVA. Press ENTER and enter (L1, L2, L3, L4).

The calculator displays the F statistic, the p-value, and the values for the one-way ANOVA table:
$F = 2.2303$
$p = 0.1241$ (p-value)
Factor
$df = 3$
$SS = 2.88732$
$MS = 0.96244$
Error
$df = 16$
$SS = 6.9044$
$MS = 0.431525$

TRY IT 13.3

Four sports teams took a random sample of players regarding their GPAs for the last year. The results are shown in Table 13.8.

Basketball	Baseball	Hockey	Lacrosse
3.6	2.1	4.0	2.0
2.9	2.6	2.0	3.6
2.5	3.9	2.6	3.9
3.3	3.1	3.2	2.7
3.8	3.4	3.2	2.5

Table 13.8 GPAs for four sports teams

Use a significance level of 5 percent and determine if there is a difference in GPA among the teams.

EXAMPLE 13.4

A fourth-grade class is studying the environment. One of the assignments is to grow bean plants in different soils. Tommy chose to grow his bean plants in soil found outside his classroom mixed with dryer lint. Tara chose to grow her bean plants in potting soil bought at the local nursery. Nick chose to grow his bean plants in soil from his mother's garden. No chemicals were used on the plants, only water. They were grown inside the classroom next to a large window. Each child grew five plants. At the end of the growing period, each plant was measured, producing the data in inches in Table 13.9.

Tommy's Plants	Tara's Plants	Nick's Plants
24	25	23
21	31	27
23	23	22
30	20	30
23	28	20

Table 13.9

Problem

Does it appear that the three soils in which the bean plants were grown produce the same mean height? Test at a 3 percent level of significance.

Solution

This time, we will perform the calculations that lead to the F' statistic. Notice that each group has the same number of plants, so we will use the formula $F' = \frac{n \cdot s_{\bar{x}}^2}{s^2_{\text{pooled}}}$.

First, calculate the sample mean and sample variance of each group.

	Tommy's Plants	Tara's Plants	Nick's Plants
Sample Mean	24.2	25.4	24.4
Sample Variance	11.7	18.3	16.3

Table 13.10

Next, calculate the variance of the three group means by calculating the variance of 24.2, 25.4, and 24.4. Variance of the group means = 0.413 = $s_{\bar{x}}^2$,

then $MS_{between} = ns_{\bar{x}}^2$ = (5)(0.413) where n = 5 is the sample size (number of plants each child grew).

Calculate the mean of the three sample variances (11.7, 18.3, and 16.3). Mean of the sample variances = 15.433 = s^2_{pooled},

then $MS_{within} = s^2_{pooled}$ = 15.433.

The F statistic (or F ratio) is $F = \frac{MS_{\text{between}}}{MS_{\text{within}}} = \frac{ns_{\bar{x}}^2}{s^2_{pooled}} = \frac{(5)(0.413)}{15.433} = 0.134.$

The dfs for the numerator = the number of groups – 1 = 3 – 1 = 2.

The dfs for the denominator = the total number of samples – the number of groups = 15 – 3 = 12.

The distribution for the test is $F_{2,12}$ and the F statistic is F = 0.134.

The p-value is $P(F > 0.134) = 0.8759$.

Decision: Since $\alpha = 0.03$ and the p-value = 0.8759, we do not reject H_0. Why?

Conclusion: With a 3 percent level of significance from the sample data, the evidence is not sufficient to conclude that the mean heights of the bean plants are different.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

To calculate the p-value:

- Press 2nd DISTR,
- Arrow down to Fcdf and press ENTER,
- Enter 0.134, E99, 2, 12, and
- Press ENTER.

The p-value is 0.8759.

TRY IT 13.4

Another fourth grader also grew bean plants, but in a jelly-like mass. The heights were (in inches) 24, 28, 25, 30, and 32. Do a one-way ANOVA test on the four groups. Are the heights of the bean plants different? Use the same method as shown in Example 13.4.

COLLABORATIVE EXERCISE

From the class, create four groups of the same size as follows: men under 22, men at least 22, women under 22, women at least 22. Have each member of each group record the number of states in the United States he or she has visited. Run an ANOVA test to determine if the average number of states visited in the four groups are the same. Test at a 1 percent level of significance. Use one of the solution sheets in Appendix E.

13.4 Test of Two Variances

Another use of the *F* distribution is testing two variances. It is often desirable to compare two variances rather than two averages. For instance, college administrators would like two college professors grading exams to have the same variation in their grading. For a lid to fit a container, the variation in the lid and the container should be the same. A supermarket might be interested in the variability of check-out times for two checkers.

To perform a *F* test of two variances, it is important that the following are true:

- The populations from which the two samples are drawn are normally distributed.
- The two populations are independent of each other.

Unlike most other tests in this book, the *F* test for equality of two variances is very sensitive to deviations from normality. If the two distributions are not normal, the test can give higher *p*-values than it should, or lower ones, in ways that are unpredictable. Many texts suggest that students not use this test at all, but in the interest of completeness we include it here.

Suppose we sample randomly from two independent normal populations. Let σ_1^2 and σ_2^2 be the population variances and s_1^2 and s_2^2 be the sample variances. Let the sample sizes be n_1 and n_2. Since we are interested in comparing the two sample variances, we use the *F* ratio

$$F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(\sigma_2)^2}\right]}.$$

F has the distribution $F \sim F(n_1 - 1, n_2 - 1)$,

where $n_1 - 1$ are the degrees of freedom for the numerator and $n_2 - 1$ are the degrees of freedom for the denominator.

If the null hypothesis is $\sigma_1^2 = \sigma_2^2$, then the *F* ratio becomes $F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(\sigma_2)^2}\right]} = \frac{(s_1)^2}{(s_2)^2}$.

Note

The *F* ratio could also be $\frac{(s_2)^2}{(s_1)^2}$. It depends on H_a and on which sample variance is larger.

If the two populations have equal variances, then s_1^2 and s_2^2 are close in value and $F = \frac{(s_1)^2}{(s_2)^2}$ is close to 1. But if the two population variances are very different, s_1^2 and s_2^2 tend to be very different, too. Choosing s_1^2 as the larger sample variance causes the ratio $\frac{(s_1)^2}{(s_2)^2}$ to be greater than 1. If s_1^2 and s_2^2 are far apart, then $F = \frac{(s_1)^2}{(s_2)^2}$ is a large number.

Therefore, if *F* is close to 1, the evidence favors the null hypothesis (the two population variances are equal). But if *F* is much larger than 1, then the evidence is against the null hypothesis. A test of two variances may be left-tailed, right-tailed, or two-tailed.

EXAMPLE 13.5

Problem

Two college instructors are interested in whethe there is any variation in the way they grade math exams. They each grade the same set of 30 exams. The first instructor's grades have a variance of 52.3. The second instructor's grades have a variance of 89.9. Test the claim that the first instructor's variance is smaller. In most colleges, it is desirable for the variances of exam grades to be nearly the same among instructors. The level of significance is 10 percent.

Solution

Let 1 and 2 be the subscripts that indicate the first and second instructor, respectively.

$n_1 = n_2 = 30$.

H_0: $\sigma_1^2 = \sigma_2^2$ and H_a: $\sigma_1^2 < \sigma_2^2$.

Calculate the test statistic: By the null hypothesis ($\sigma_1^2 = \sigma_2^2$), the F statistic is

$$F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(\sigma_2)^2}\right]} = \frac{(s_1)^2}{(s_2)^2} = \frac{52.3}{89.9} = 0.5818.$$

Distribution for the test: $F_{29,29}$ where $n_1 - 1 = 29$ and $n_2 - 1 = 29$.

Graph: This test is left-tailed.

Draw the graph, labeling and shading appropriately.

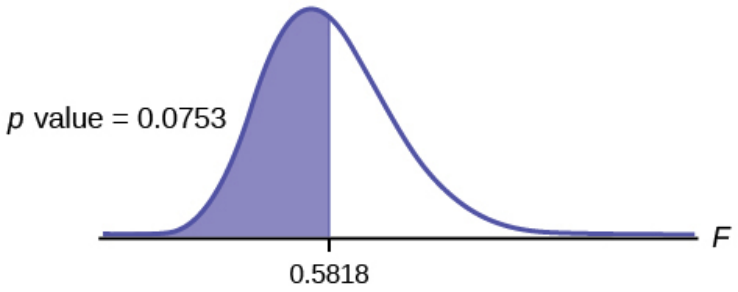


Figure 13.7

Probability statement: p-value = $P(F < 0.5818) = 0.0753$.

Compare α and the p-value: $\alpha = 0.10$; $\alpha > p$-value.

Make a decision: Since $\alpha > p$-value, reject H_0.

Conclusion: With a 10 percent level of significance from the data, there is sufficient evidence to conclude that the variance in grades for the first instructor is smaller.

USING THE TI-83, 83+, 84, 84+ CALCULATOR

Press STAT and arrow over to TESTS. Arrow down to `D:2-SampFTest`. Press ENTER. Arrow to `Stats` and press ENTER. For `Sx1`, `n1`, `Sx2`, and `n2`, enter $\sqrt{(52.3)}$, `30`, $\sqrt{(89.9)}$, and `30`. Press ENTER after each. Arrow to `σ1:` and `<σ2`. Press ENTER. Arrow down to `Calculate` and press ENTER. $F = 0.5818$ and p-value = 0.0753. Do the procedure again and try `Draw` instead of `Calculate`.

TRY IT 13.5

The New York Choral Society divides male singers into four categories from highest voices to lowest: Tenor1, Tenor2, Bass1, and Bass2. In the table are heights of the men in the Tenor1 and Bass2 groups. One suspects that taller men will have lower voices, and that the variance of height may go up with the lower voices as well. Do we have good evidence that the variance of the heights of singers in each of these two groups (Tenor1 and Bass2) are different?

Tenor1	Bass2	Tenor1	Bass2	Tenor1	Bass2
69	72	67	72	68	67
72	75	70	74	67	70
71	67	65	70	64	70
66	75	72	66		69
76	74	70	68		72
74	72	68	75		71
71	72	64	68		74
66	74	73	70		75
68	72	66	72		

Table 13.11

13.5 Lab: One-Way ANOVA

Stats Lab

One-Way ANOVA

Student Learning Outcome

- The student will conduct a simple one-way ANOVA test involving three variables.

Collect the Data

1. Record the price per pound of eight fruits, eight vegetables, and eight breads in your local supermarket.

Fruits	Vegetables	Breads

Table 13.12

2. Explain how you could try to collect the data randomly.

Analyze the Data and Conduct a Hypothesis Test

1. State the null hypothesis and the alternative hypothesis.
2. Compute the following:
   a. Fruit
      i. $\overline{x}$ = _____
      ii. s_x = _____
      iii. n = _____
   b. Vegetables
      i. $\overline{x}$ = _____
      ii. s_x = _____
      iii. n = _____
   c. Bread
      i. $\overline{x}$ = _____
      ii. s_x = _____
      iii. n = _____
3. Find the following:
   a. $df(num)$ = _____
   b. $df(denom)$ = _____
4. State the approximate distribution for the test.
5. Test statistic: F = _____
6. Sketch a graph of this situation. Clearly label and scale the horizontal axis and shade the region(s) corresponding to the p-value.
7. p-value = _____
8. Test at $\alpha = 0.05$. State your decision and conclusion.
9. a. Decision: why did you make this decision?
   b. Conclusion (write a complete sentence):
   c. Based on the results of your study, is there a need to investigate any of the food groups's prices? Why or why not?

Key Terms

analysis of variance also referred to as ANOVA; a method of testing whether the means of three or more populations are equal
The method is applicable if
- all populations of interest are normally distributed,
- the populations have equal standard deviations, and
- samples (not necessarily of the same size) are randomly and independently selected from each population.

The test statistic for analysis of variance is the *F* ratio.

one-way ANOVA a method of testing whether the means of three or more populations are equal; the method is applicable if
- all populations of interest are normally distributed,
- the populations have equal standard deviations,
- samples (not necessarily of the same size) are randomly and independently selected from each population, and
- there is one independent variable and one dependent variable.

The test statistic for analysis of variance is the *F* ratio

variance mean of the squared deviations from the mean; the square of the standard deviation
For a set of data, a deviation can be represented as $x - \overline{x}$ where x is a value of the data and $\overline{x}$ is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and 1.

Chapter Review

13.1 One-Way ANOVA

Analysis of variance extends the comparison of two groups to several, each a level of a categorical variable (factor). Samples from each group are independent and must be randomly selected from normal populations with equal variances. We test the null hypothesis of equal means of the response in every group versus the alternative hypothesis of one or more group means being different from the others. A one-way ANOVA hypothesis test determines if several population means are equal. The distribution for the test is the *F* distribution with two different degrees of freedom.

Assumptions:
- Each population from which a sample is taken is assumed to be normal.
- All samples are randomly selected and independent.
- The populations are assumed to have equal standard deviations (or variances).

13.2 The F Distribution and the F Ratio

Analysis of variance compares the means of a response variable for several groups. ANOVA compares the variation within each group to the variation of the mean of each group. The ratio of these two is the *F* statistic from an *F* distribution with (number of groups – 1) as the numerator degrees of freedom and (number of observations – number of groups) as the denominator degrees of freedom. These statistics are summarized in the ANOVA table.

13.3 Facts About the F Distribution

The graph of the *F* distribution is always positive and skewed right, though the shape can be mounded or exponential depending on the combination of numerator and denominator degrees of freedom. The *F* statistic is the ratio of a measure of the variation in the group means to a similar measure of the variation within the groups. If the null hypothesis is correct, then the numerator should be small compared to the denominator. A small *F* statistic will result, and the area under the *F* curve to the right will be large, representing a large *p*-value. When the null hypothesis of equal group means is incorrect, then the numerator should be large compared to the denominator, giving a large *F* statistic and a small area (small *p*-value) to the right of the statistic under the *F* curve.

When the data have unequal group sizes (unbalanced data), then techniques from The F Distribution and the F Ratio need to be used for hand calculations. In the case of balanced data, where the groups are the same size, simplified calculations based on group means and variances may be used. In practice, software is usually employed in the analysis. As in any analysis, graphs of various sorts should be used in conjunction with numerical techniques. Always look at your data!

13.4 Test of Two Variances

The *F* test for the equality of two variances rests heavily on the assumption of normal distributions. The test is unreliable if this assumption is not met. If both distributions are normal, then the ratio of the two sample variances is distributed as an *F* statistic, with numerator and denominator degrees of freedom that are one less than the samples sizes of the corresponding two groups. A *test of two variances* hypothesis test determines if two variances are the same. The distribution for the hypothesis test is the *F* distribution with two different degrees of freedom.

Assumptions:

- The populations from which the two samples are drawn are normally distributed.
- The two populations are independent of each other.

Formula Review

13.2 The F Distribution and the F Ratio

$$SS_{\text{between}} = \sum \left[\frac{(s_j)^2}{n_j}\right] - \frac{\left(\sum s_j\right)^2}{n}$$

$$SS_{\text{total}} = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

$$SS_{\text{within}} = SS_{\text{total}} - SS_{\text{between}}$$

$df_{\text{between}} = df(num) = k - 1$

$df_{\text{within}} = df(denom) = n - k$

$$MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}}$$

$$MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}}$$

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

F ratio when the groups are the same size: $F = \frac{ns_{\bar{x}}{}^2}{s^2{}_{pooled}}$

Mean of the *F* distribution: $\mu = \frac{df(num)}{df(denom)-1}$

where

- k = the number of groups
- n_j = the size of the j^{th} group
- s_j = the sum of the values in the j^{th} group
- n = the total number of all values (observations) combined
- x = one value (one observation) from the data
- $s_{\bar{x}}{}^2$ = the variance of the sample means
- $s^2{}_{pooled}$ = the mean of the sample variances (pooled variance)

13.4 Test of Two Variances

F has the distribution $F \sim F(n_1 - 1, n_2 - 1)$

$$F = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}}$$

If $\sigma_1 = \sigma_2$, then $F = \frac{s_1{}^2}{s_2{}^2}$

Practice

13.1 One-Way ANOVA

Use the following information to answer the next five exercises. There are five basic assumptions that must be fulfilled to perform a one-way ANOVA test. What are they?

1. Write one assumption.
2. Write another assumption.
3. Write a third assumption.
4. Write a fourth assumption.
5. Write the final assumption.
6. State the null hypothesis for a one-way ANOVA test if there are four groups.
7. State the alternative hypothesis for a one-way ANOVA test if there are three groups.
8. When do you use an ANOVA test?

13.2 The F Distribution and the F Ratio

Use the following information to answer the next seven exercises. Groups of men from three different areas of the country are to be tested for mean weight. The entries in Table 13.13 are the weights for the different groups.

Group 1	Group 2	Group 3
216	202	170
198	213	165
240	284	182
187	228	197
176	210	201

Table 13.13

9. What is the sum of squares factor?

10. What is the sum of squares error?

11. What is the *df* for the numerator?

12. What is the *df* for the denominator?

13. What is the mean square factor?

14. What is the mean square error?

15. What is the *F* statistic?

Use the following information to answer the next eight exercises. Girls from four different soccer teams are to be tested for mean goals scored per game. The entries in Table 13.14 are the goals per game for the different teams.

Team 1	Team 2	Team 3	Team 4
1	2	0	3
2	3	1	4
0	2	1	4
3	4	0	3
2	4	0	2

Table 13.14

16. What is $SS_{between}$?

17. What is the *df* for the numerator?

18. What is $MS_{between}$?

19. What is SS_{within}?

20. What is the *df* for the denominator?

21. What is MS_{within}?

22. What is the *F* statistic?

23. Judging by the *F* statistic, do you think it is likely or unlikely that you will reject the null hypothesis?

13.3 Facts About the F Distribution

24. An *F* statistic can have what values?

25. What happens to the curves as the degrees of freedom for the numerator and the denominator get larger?

Use the following information to answer the next seven exercises. Four basketball teams took a random sample of players regarding how high each player can jump (in inches). The results are shown in Table 13.15.

Team 1	Team 2	Team 3	Team 4	Team 5
36	32	48	38	41
42	35	50	44	39
51	38	39	46	40

Table 13.15

26. What is the *df(num)*?

27. What is the *df(denom)*?

28. What are the sum of squares and mean squares factors?

29. What are the sum of squares and mean squares errors?

30. What is the *F* statistic?

31. What is the *p*-value?

32. At the 5 percent significance level, is there a difference in the mean jump heights among the teams?

Use the following information to answer the next seven exercises. A video game developer is testing a new game on three different groups. Each group represents a different target market for the game. The developer collects scores from a random sample from each group. The results are shown in Table 13.16.

Group A	Group B	Group C
101	151	101
108	149	109
98	160	198
107	112	186
111	126	160

Table 13.16

33. What is the *df(num)*?

34. What is the *df(denom)*?

35. What are the $SS_{between}$ and $MS_{between}$?

36. What are the SS_{within} and MS_{within}?

37. What is the *F* Statistic?

38. What is the *p*-value?

39. At the 10 percent significance level, are the scores among the different groups different?

Use the following information to answer the next three exercises. Suppose a group is interested in determining whether teenagers obtain their drivers licenses at approximately the same average age across the country. Suppose that the following data are randomly collected from five teenagers in each region of the country. The numbers represent the age at which teenagers obtained their drivers licenses.

	Northeast	South	West	Central	East
	16.3	16.9	16.4	16.2	17.1
	16.1	16.5	16.5	16.6	17.2
	16.4	16.4	16.6	16.5	16.6
	16.5	16.2	16.1	16.4	16.8
$\overline{x} =$	_____	_____	_____	_____	_____
$s^2 =$	_____	_____	_____	_____	_____

Table 13.17

Enter the data into your calculator or computer.

40. *p*-value = _____

State the decisions and conclusions (in complete sentences) for the following preconceived levels of α.

41. $\alpha = 0.05$
a. Decision: _____
b. Conclusion: _____

42. $\alpha = 0.01$
a. Decision: _____
b. Conclusion: _____

13.4 Test of Two Variances

Use the following information to answer the next two exercises. There are two assumptions that must be true to perform an *F* test of two variances.

43. Name one assumption that must be true.

44. What is the other assumption that must be true?

Use the following information to answer the next seven exercises. Two coworkers commute from the same building. They are interested in whether there is any variation in the time it takes them to drive to work. They each record their times for 20 commutes. The first worker's times have a variance of 12.1. The second worker's times have a variance of 16.9. The first worker thinks that he is more consistent with his commute times. Test the claim at the 10 percent level. Assume that commute times are normally distributed.

45. State the null and alternative hypotheses.

46. What is s_1 in this problem?

47. What is s_2 in this problem?

48. What is *n*?

49. What is the *F* statistic?

50. What is the *p*-value?

51. Is the claim accurate?

Use the following information to answer the next four exercises. Two students are interested in whether there is variation in their test scores for math class. There are 15 total math tests they have taken so far. The first student's grades have a standard deviation of 38.1. The second student's grades have a standard deviation of 22.5. The second student thinks his scores are more consistent.

52. State the null and alternative hypotheses.

53. What is the F statistic?

54. What is the p-value?

55. At the 5 percent significance level, do we reject the null hypothesis?

Use the following information to answer the next three exercises. Two cyclists are comparing the variances of their overall paces going uphill. Each cyclist records his or her speeds going up 35 hills. The first cyclist has a variance of 23.8, and the second cyclist has a variance of 32.1. The cyclists want to see if their variances are the same or different. Assume that speeds are normally distributed.

56. State the null and alternative hypotheses.

57. What is the F statistic?

58. At the 5 percent significance level, what can we say about the cyclists' variances?

Homework

13.1 One-Way ANOVA

59. Three different traffic routes are tested for mean driving time. The entries in the Table 13.18 are the driving times in minutes on the three different routes.

Route 1	Route 2	Route 3
30	27	16
32	29	41
27	28	22
35	36	31

Table 13.18

State $SS_{between}$, SS_{within}, and the F statistic.

60. Suppose a group is interested in determining whether teenagers obtain their drivers licenses at approximately the same average age across the country. Suppose that the following data are randomly collected from five teenagers in each region of the country. The numbers represent the age at which teenagers obtained their drivers licenses.

	Northeast	South	West	Central	East
	16.3	16.9	16.4	16.2	17.1
	16.1	16.5	16.5	16.6	17.2
	16.4	16.4	16.6	16.5	16.6

Table 13.19

	Northeast	South	West	Central	East
	16.5	16.2	16.1	16.4	16.8
$\overline{x} =$	_____	_____	_____	_____	_____
$s^2 =$	_____	_____	_____	_____	_____

Table 13.19

State the hypotheses.
H_0: _____
H_a: _____

13.2 The F Distribution and the F Ratio

Use the following information to answer the next three exercises. Suppose a group is interested in determining whether teenagers obtain their drivers licenses at approximately the same average age across the country. Suppose that the following data are randomly collected from five teenagers in each region of the country. The numbers represent the age at which teenagers obtained their drivers licenses.

	Northeast	South	West	Central	East
	16.3	16.9	16.4	16.2	17.1
	16.1	16.5	16.5	16.6	17.2
	16.4	16.4	16.6	16.5	16.6
	16.5	16.2	16.1	16.4	16.8
$\overline{x} =$	_____	_____	_____	_____	_____
$s^2 =$	_____	_____	_____	_____	_____

Table 13.20

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

Ha: At least any two of the group means $\mu_1, \mu_2, ..., \mu_5$ are not equal.

61. degrees of freedom – numerator: $df(num)$ = _____

62. degrees of freedom – denominator: $df(denom)$ = _____

63. F statistic = _____

13.3 Facts About the F Distribution

DIRECTIONS

Use a solution sheet to conduct the following hypothesis tests. The solution sheet can be found in Appendix E.

64. Three students, Linda, Tuan, and Javier, are given five laboratory rats each for a nutritional experiment. Each rat's weight is recorded in grams. Linda feeds her rats Formula A, Tuan feeds his rats Formula B, and Javier feeds his rats Formula C. At the end of a specified time period, each rat is weighed again, and the net gain in grams is recorded. Using a significance level of 10 percent, test the hypothesis that the three formulas produce the same mean weight gain.

Linda's Rats (g)	Tuan's Rats (g)	Javier's Rats (g)
43.5	47.0	51.2
39.4	40.5	40.9
41.3	38.9	37.9
46.0	46.3	45.0
38.2	44.2	48.6

Table 13.21

65. A grassroots group opposed to a proposed increase in the gas tax claimed that the increase would hurt working-class people the most since they commute the farthest to work. Suppose that the group randomly surveyed 24 individuals and asked them their daily one-way commuting mileage. The results are in Table 13.22. Using a 5 percent significance level, test the hypothesis that the three mean commuting mileages are the same.

Working-Class	Professional (middle incomes)	Professional (wealthy)
17.8	16.5	8.5
26.7	17.4	6.3
49.4	22.0	4.6
9.4	7.4	12.6
65.4	9.4	11.0
47.1	2.1	28.6
19.5	6.4	15.4
51.2	13.9	9.3

Table 13.22

Use the following information to answer the next two exercises. Table 13.23 lists the number of pages in four different types of magazines.

Home Decorating	News	Health	Computer
172	87	82	104
286	94	153	136
163	123	87	98

Table 13.23

Home Decorating	News	Health	Computer
205	106	103	207
197	101	96	146

Table 13.23

66. Using a significance level of 5 percent, test the hypothesis that the four magazine types have the same mean length.

67. Eliminate one magazine type that you now feel has a mean length different from the others. Redo the hypothesis test, testing that the remaining three means are statistically the same. Use a new solution sheet. Based on this test, are the mean lengths for the remaining three magazines statistically the same?

68. A researcher wants to know if the mean times (in minutes) that people watch their favorite news station are the same. Suppose that Table 13.24 shows the results of a study.

CNN	FOX	Local
45	15	72
12	43	37
18	68	56
38	50	60
23	31	51
35	22	

Table 13.24

Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a level of significance of 0.05.

69. Are the means for the final exams the same for all statistics class delivery types? Table 13.25 shows the scores on final exams from several randomly selected classes that used the different delivery types.

Online	Hybrid	Face-to-Face
72	83	80
84	73	78
77	84	84
80	81	81
81		86

Table 13.25

Online	Hybrid	Face-to-Face
		79
		82

Table 13.25

Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a level of significance of 0.05.

70. Are the mean number of times a month a person eats out the same for whites, blacks, Hispanics, and Asians? Suppose that Table 13.26 shows the results of a study.

White	Black	Hispanic	Asian
6	4	7	8
8	1	3	3
2	5	5	5
4	2	4	1
6		6	7

Table 13.26

Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a level of significance of 0.05.

71. Are the mean numbers of daily visitors to a ski resort the same for the three types of snow conditions? Suppose that Table 13.27 shows the results of a study.

Powder	Machine Made	Hard Packed
1,210	2,107	2,846
1,080	1,149	1,638
1,537	862	2,019
941	1,870	1,178
	1,528	2,233
	1,382	

Table 13.27

Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a level of significance of 0.05.

72. Sanjay made identical paper airplanes out of three different weights of paper: light, medium, and heavy. He made four airplanes from each of the weights and launched them himself across the room. Here are the distances (in meters) that his planes flew.

Paper Type/Trial	Trial 1	Trial 2	Trial 3	Trial 4
Heavy	5.1 meters	3.1 meters	4.7 meters	5.3 meters
Medium	4 meters	3.5 meters	4.5 meters	6.1 meters
Light	3.1 meters	3.3 meters	2.1 meters	1.9 meters

Table 13.28

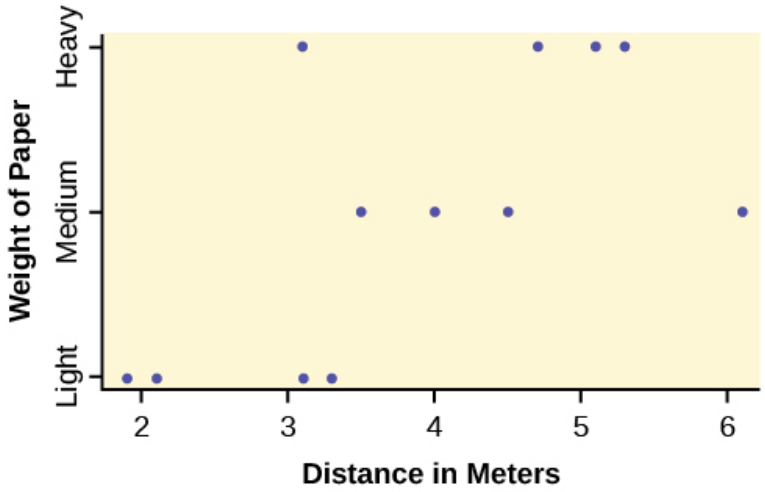


Figure 13.8

a. Take a look at the data in the graph. Look at the spread of data for each group (light, medium, heavy). Does it seem reasonable to assume a normal distribution with the same variance for each group?
b. Why is this a balanced design?
c. Calculate the sample mean and sample standard deviation for each group.
d. Does the weight of the paper have an effect on how far the plane will travel? Use a 1 percent level of significance. Complete the test using the method shown in the bean plant example in Example 13.4.
   - Variance of the group means _____
   - $MS_{between}$= _____
   - Mean of the three sample variances _____
   - MS_{within} = _____
   - F statistic = _____
   - $df(num)$ = _____, $df(denom)$ = _____
   - Number of groups _____
   - Number of observations _____
   - p-value = _____ ($P(F >$ _____) = _____)
   - Graph the p-value.
   - Decision: _____
   - Conclusion: _____

73. DDT is a pesticide that has been banned from use in the United States and most other areas of the world. It is quite effective but persisted in the environment and over time proved to be harmful to higher-level organisms. Famously, egg shells of eagles and other raptors were believed to be thinner and prone to breakage in the nest because of ingestion of DDT in the food chain of the birds.
An experiment was conducted on the number of eggs (fecundity) laid by female fruit flies. There are three groups of flies. One group was bred to be resistant to DDT (the RS group). Another was bred to be especially susceptible to DDT (SS). The third group was a control line of nonselected or typical fruit flies (NS). Here are the data:

RS	SS	NS	RS	SS	NS
12.8	38.4	35.4	22.4	23.1	22.6
21.6	32.9	27.4	27.5	29.4	40.4
14.8	48.5	19.3	20.3	16	34.4
23.1	20.9	41.8	38.7	20.1	30.4
34.6	11.6	20.3	26.4	23.3	14.9
19.7	22.3	37.6	23.7	22.9	51.8
22.6	30.2	36.9	26.1	22.5	33.8
29.6	33.4	37.3	29.5	15.1	37.9
416.4	26.7	228.2	38.6	31	29.5
20.3	39	23.4	44.4	16.9	42.4
29.3	12.8	33.7	23.2	16.1	36.6
914.9	14.6	29.2	23.6	10.8	47.4
27.3	12.2	41.7			

Table 13.29

The values are the average number of eggs laid daily for each of 75 flies (25 in each group) over the first 14 days of their lives. Using a 1 percent level of significance, are the mean rates of egg selection for the three strains of fruit fly different? If so, in what way? Specifically, the researchers were interested in whether the selectively bred strains were different from the nonselected line, and whether the two selected lines were different from each other.
Here is a chart of the three groups:

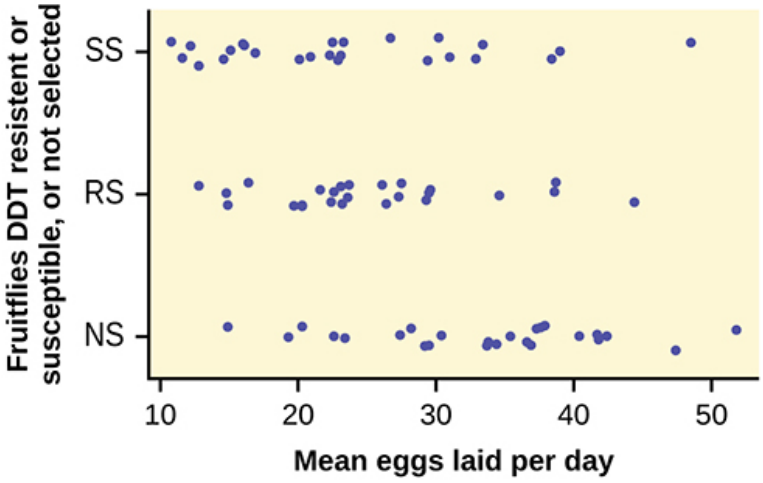


Figure 13.9

74. The data shown is the recorded body temperatures of 130 subjects as estimated from available histograms. Traditionally, we are taught that the normal human body temperature is 98.6 °F. This is not quite correct for everyone. Are the mean temperatures among the four groups different?
Calculate 95 percent confidence intervals for the mean body temperature in each group and comment about the confidence intervals.

FL	FH	ML	MH	FL	FH	ML	MH
96.4	96.8	96.3	96.9	98.4	98.6	98.1	98.6
96.7	97.7	96.7	97	98.7	98.6	98.1	98.6
97.2	97.8	97.1	97.1	98.7	98.6	98.2	98.7
97.2	97.9	97.2	97.1	98.7	98.7	98.2	98.8
97.4	98	97.3	97.4	98.7	98.7	98.2	98.8
97.6	98	97.4	97.5	98.8	98.8	98.2	98.8
97.7	98	97.4	97.6	98.8	98.8	98.3	98.9
97.8	98	97.4	97.7	98.8	98.8	98.4	99
97.8	98.1	97.5	97.8	98.8	98.9	98.4	99
97.9	98.3	97.6	97.9	99.2	99	98.5	99
97.9	98.3	97.6	98	99.3	99	98.5	99.2
98	98.3	97.8	98		99.1	98.6	99.5
98.2	98.4	97.8	98		99.1	98.6	
98.2	98.4	97.8	98.3		99.2	98.7	
98.2	98.4	97.9	98.4		99.4	99.1	
98.2	98.4	98	98.4		99.9	99.3	
98.2	98.5	98	98.6		100	99.4	
98.2	98.6	98	98.6		100.8		

Table 13.30

13.4 Test of Two Variances

75. Three students, Linda, Tuan, and Javier, are given five laboratory rats each for a nutritional experiment. Each rat's weight is recorded in grams. Linda feeds her rats Formula A, Tuan feeds his rats Formula B, and Javier feeds his rats Formula C. At the end of a specified time period, each rat is weighed again and the net gain in grams is recorded.

Linda's Rats	Tuan's Rats	Javier's Rats
43.5	47.0	51.2
39.4	40.5	40.9
41.3	38.9	37.9
46.0	46.3	45.0
38.2	44.2	48.6

Table 13.31

Determine whether the variance in weight gain is statistically the same between Javier's and Linda's rats. Test at a significance level of 10 percent.

76. A grassroots group opposed to a proposed increase in the gas tax claimed that the increase would hurt working-class people the most since they commute the farthest to work. Suppose that the group randomly surveyed 24 individuals and asked them their daily one-way commuting mileage. The results are as follows.

Working-Class	Professional (middle incomes)	Professional (wealthy)
17.8	16.5	8.5
26.7	17.4	6.3
49.4	22.0	4.6
9.4	7.4	12.6
65.4	9.4	11.0
47.1	2.1	28.6
19.5	6.4	15.4
51.2	13.9	9.3

Table 13.32

Determine whether the variance in mileage driven is statistically the same between the working class and professional (middle income) groups. Use a 5 percent significance level.

Use the following information to answer the next two exercises. The following table lists the number of pages in four different types of magazines.

Home Decorating	News	Health	Computer
172	87	82	104
286	94	153	136
163	123	87	98

Table 13.33

Home Decorating	News	Health	Computer
205	106	103	207
197	101	96	146

Table 13.33

77. Which two magazine types do you think have the same variance in length?

78. Which two magazine types do you think have different variances in length?

79. Is the variance for the amount of money, in dollars, that shoppers spend on Saturdays at the mall the same as the variance for the amount of money that shoppers spend on Sundays at the mall? Suppose that Table 13.34 shows the results of a study.

Saturday	Sunday	Saturday	Sunday
75	44	62	137
18	58	0	82
150	61	124	39
94	19	50	127
62	99	31	141
73	60	118	73
	89		

Table 13.34

80. Are the variances for incomes on the East Coast and the West Coast the same? Suppose that Table 13.35 shows the results of a study. Income is shown in thousands of dollars. Assume that both distributions are normal. Use a level of significance of 0.05.

East	West
38	71
47	126
30	42
82	51
75	44
52	90

Table 13.35

East	West
115	88
67	

Table 13.35

81. Thirty men in college were taught a method of finger tapping. They were randomly assigned to three groups of 10, with each receiving one of three doses of caffeine: 0 mg, 100 mg, or 200 mg. This is approximately the amount in zero, one, or two cups of coffee. Two hours after ingesting the caffeine, the men had the rate of finger tapping per minute recorded. The experiment was double blind, so neither the recorders nor the students knew which group they were in. Does caffeine affect the rate of tapping, and if so how?
Here are the data:

0 mg	100 mg	200 mg	0 mg	100 mg	200 mg
242	248	246	245	246	248
244	245	250	248	247	252
247	248	248	248	250	250
242	247	246	244	246	248
246	243	245	242	244	250

Table 13.36

82. King Manuel I Komnenos ruled the Byzantine Empire from Constantinople (Istanbul) during the years A.D. 1145–1170. The empire was very powerful during his reign but declined significantly afterward. Coins minted during his era were found in Cyprus, an island in the eastern Mediterranean Sea. Nine coins were from his first coinage, seven from the second, four from the third, and seven from the fourth. These spanned most of his reign. We have data on the silver content of the coins:

First Coinage	Second Coinage	Third Coinage	Fourth Coinage
5.9	6.9	4.9	5.3
6.8	9.0	5.5	5.6
6.4	6.6	4.6	5.5
7.0	8.1	4.5	5.1
6.6	9.3		6.2
7.7	9.2		5.8
7.2	8.6		5.8

Table 13.37

First Coinage	Second Coinage	Third Coinage	Fourth Coinage
6.9			
6.2			

Table 13.37

Did the silver content of the coins change over the course of Manuel's reign? Here are the means and variances of each coinage. The data are unbalanced.

	First	Second	Third	Fourth
Mean	6.7444	8.2429	4.875	5.6143
Variance	0.2953	1.2095	0.2025	0.1314

Table 13.38

83. The American League and the National League of Major League Baseball are each divided into three divisions: East, Central, and West. Many years, fans talk about some divisions being stronger (having better teams) than other divisions. This may have consequences for the postseason. For instance, in 2012 Tampa Bay won 90 games and did not play in the postseason, while Detroit won only 88 and did play in the postseason. This may have been an oddity, but is there good evidence that in the 2012 season, the American League divisions were significantly different in overall records? Use the following data to test whether the mean number of wins per team in the three American League divisions were the same. Note that the data are not balanced, as two divisions had five teams, while one had only four.

Division	Team	Wins
East	NY Yankees	95
East	Baltimore	93
East	Tampa Bay	90
East	Toronto	73
East	Boston	69

Table 13.39

Division	Team	Wins
Central	Detroit	88
Central	Chicago Sox	85
Central	Kansas City	72

Table 13.40

Division	Team	Wins
Central	Cleveland	68
Central	Minnesota	66

Table 13.40

Division	Team	Wins
West	Oakland	94
West	Texas	93
West	LA Angels	89
West	Seattle	75

Table 13.41

References

13.2 The F Distribution and the F Ratio

Marist College School of Science. (n.d.). *Tomato data* (Unpublished student research). Marist College School of Science, Poughkeepsie, NY.

13.3 Facts About the F Distribution

ESPN. (2012). *MLB standings – 2012.* Retrieved from http://espn.go.com/mlb/standings/_/year/2012.

Hand, D. J. et al. (1994). *A Handbook of Small Datasets: Data for Fruitfly Fecundity.* London: Chapman & Hall.

Hand, D. J. et al. (1994). *A Handbook of Small Datasets.* London: Chapman & Hall, p. 50.

Hand. *A Handbook of Small Datasets.* p. 118.

Mackowiak, P. A.,Wasserman, S. S., & Levine, M. M. (1992). A critical appraisal of 98.6 degrees F, the upper limit of the normal body temperature, and other legacies of Carl Reinhold August Wunderlich. *Journal of the American Medical Association, 268,* 1578–1580.

Private K–12 school in San Jose, CA. (1994). *Data from a fourth grade classroom.*

13.4 Test of Two Variances

ESPN. (2012). *MLB standings – 2012.* Retrieved from http://espn.go.com/mlb/standings/_/year/2012/type/vs-division/order/true.

Solutions

1. Each population from which a sample is taken is assumed to be normal.

3. The populations are assumed to have equal standard deviations (or variances).

5. The response is a numerical value.

7. H_a: At least two of the group means μ_1, μ_2, μ_3 are not equal.

9. 4,939.2

11. 2

13. 2,469.6

15. 3.7416

17. 3

19. 13.2

21. 0.825

23. Because a one-way ANOVA test is always right-tailed, a high F statistic corresponds to a low p value, so it is likely that we will reject the null hypothesis.

25. The curves approximate the normal distribution.

27. 10

29. $SS = 237.33$; $MS = 23.73$

31. 0.1614

33. two

35. $SS = 5{,}700.4$;
$MS = 2{,}850.2$

37. 3.6101

39. Yes, there is enough evidence to show that the scores among the groups are statistically significant at the 10 percent level.

43. The populations from which the two samples are drawn are normally distributed.

45. H_0: $\sigma_1 = \sigma_2$
H_a: $\sigma_1 < \sigma_2$
or
H_0: $\sigma_1^2 = \sigma_2^2$
H_a: $\sigma_1^2 < \sigma_2^2$

47. 4.11

49. 0.7159

51. No, at the 10 percent level of significance, we do not reject the null hypothesis and state that the data do not show that the variation in drive times for the first worker is less than the variation in drive times for the second worker.

53. 2.8674

55. Reject the null hypothesis. There is enough evidence to say that the variance of the grades for the first student is higher than the variance in the grades for the second student.

57. 0.7414

59. $SS_{\text{between}} = 26$
$SS_{\text{within}} = 441$
$F = 0.2653$

62. $df(denom) = 15$

64. a. H_0: $\mu_L = \mu_T = \mu_J$
b. H_a: at least any two of the means are different
c. $df(num) = 2$; $df(denom) = 12$
d. F distribution
e. 0.67
f. 0.5305
g. Check student's solution.
h. Decision: Do not reject null hypothesis.
i. Conclusion: There is insufficient evidence to conclude that the means are different.

67. a. H_a: $\mu_c = \mu_n = \mu_h$
b. At least any two of the magazines have different mean lengths.
c. $df(num) = 2$, $df(denom) = 12$
d. F distribtuion
e. $F = 15.28$
f. p-value = 0.0005
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Reject the null hypothesis.
iii. Reason for decision: p-value < alpha
iv. Conclusion: There is sufficient evidence to conclude that the mean lengths of the magazines are different.

69. a. H_0: $\mu_o = \mu_h = \mu_f$
b. At least two of the means are different.
c. $df(n) = 2$, $df(d) = 13$
d. $F_{2,13}$
e. 0.64
f. 0.5437
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: p-value > alpha
iv. Conclusion: The mean scores of different class delivery are not different.

71. a. H_0: $\mu_p = \mu_m = \mu_h$
b. At least any two of the means are different.
c. $df(n) = 2$, $df(d) = 12$
d. $F_{2,12}$
e. 3.13
f. 0.0807
g. Check student's solution.
h. i. Alpha: 0.05
ii. Decision: Do not reject the null hypothesis.
iii. Reason for decision: p-value > alpha
iv. Conclusion: There is not sufficient evidence to conclude that the mean numbers of daily visitors are different.

73. The data appear normally distributed from the chart and of similar spread. There do not appear to be any serious outliers, so we may proceed with our ANOVA calculations, to see if we have good evidence of a difference between the three groups.
H_0: $\mu_1 = \mu_2 = \mu_3$
H_a: $\mu_i \neq \mu_j$ some $i \neq j$
Define μ_1, μ_2, μ_3, as the population mean number of eggs laid by the three groups of fruitflies.
F statistic = 8.6657
p-value = 0.0004

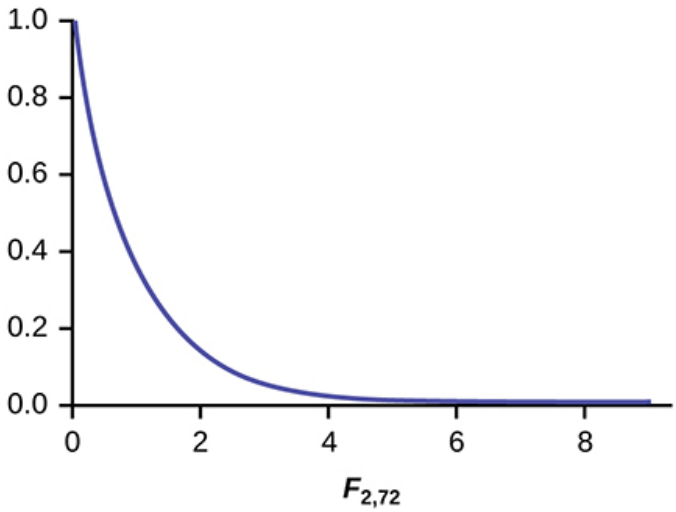


Figure 13.10

Decision: Since the *p*-value is less than the level of significance of 0.01, we reject the null hypothesis.
Conclusion: We have good evidence that the average number of eggs laid during the first 14 days of life for these three strains of fruitflies are different.
Interestingly, if you perform a two sample *t* test to compare the RS and NS groups they are significantly different (*p* = 0.0013). Similarly, SS and NS are significantly different (*p* = 0.0006). However, the two selected groups, RS and SS are *not* significantly different (*p* = 0.5176). Thus we appear to have good evidence that selection either for resistance or for susceptibility involves a reduced rate of egg production (for these specific strains) as compared to flies that were not selected for resistance or susceptibility to DDT. Here, genetic selection has apparently involved a loss of fecundity.

75.
a. $H_0: \sigma_1^2 = \sigma_2^2$
b. $H_a: \sigma_1^2 \neq \sigma_1^2$
c. *df*(*num*) = 4; *df*(*denom*) = 4
d. $F_{4, 4}$
e. 3.00
f. 2(0.1563) = 0.3126. Using the TI-83+/84+ function 2-SampFtest, you get the test statistic as 2.9986 and *p*-value directly as 0.3127. If you input the lists in a different order, you get a test statistic of 0.3335 but the *p*-value is the same because this is a two-tailed test.
g. Check student's solution.
h. Decision: Do not reject the null hypothesis.
i. Conclusion: There is insufficient evidence to conclude that the variances are different.

78. The answers may vary. Sample answer: Home decorating magazines and news magazines have different variances.

80.
a. $H_0: = \sigma_1^2 = \sigma_2^2$
b. $H_a: \sigma_1^2 \neq \sigma_1^2$
c. *df*(*n*) = 7, *df*(*d*) = 6
d. $F_{7,6}$
e. 0.8117
f. 0.7825
g. Check student's solution.
h.
   i. Alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: *p*-value > alpha
   iv. Conclusion: There is not sufficient evidence to conclude that the variances are different.

82. Here is a strip chart of the silver content of the coins:

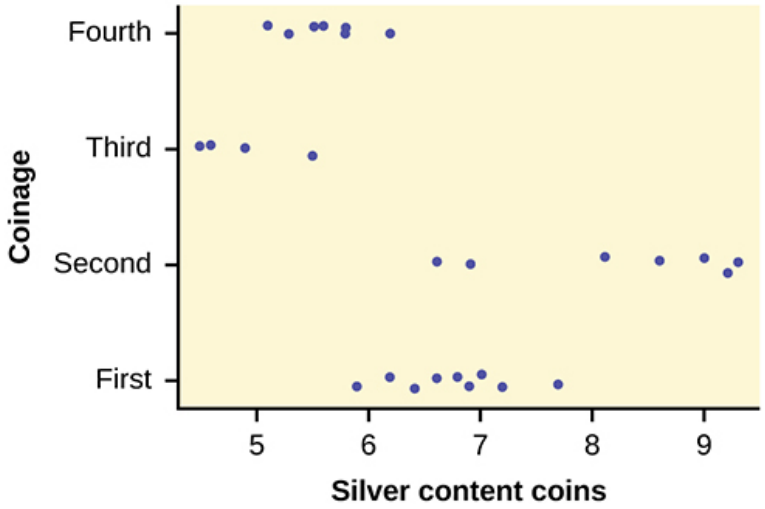


Figure 13.11

While there are differences in spread, it is not unreasonable to use ANOVA techniques. Here is the completed ANOVA table:

Source of Variation	Sum of Squares (*SS*)	Degrees of Freedom (*df*)	Mean Square (*MS*)	*F*
Factor (between)	37.748	4 – 1 = 3	12.5825	26.272
Error (within)	11.015	27 – 4 = 23	0.4789	
Total	48.763	27 – 1 = 26		

Table 13.42

$P(F > 26.272) = 0$.
Reject the null hypothesis for any alpha. There is sufficient evidence to conclude that the mean silver content among the four coinages are different. From the strip chart, it appears that the first and second coinages had higher silver contents than the third and fourth.

83. Here is a stripchart of the number of wins for the 14 teams in the AL for the 2012 season.

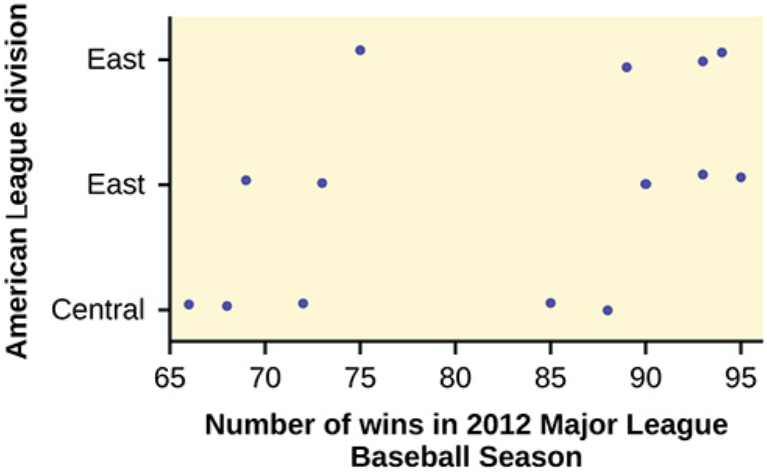


Figure 13.12

While the spread seems similar, there may be some question about the normality of the data, given the wide gaps in the middle near the 0.500 mark of 82 games (teams play 162 games each season in MLB). However, one-way ANOVA is robust.

Here is the ANOVA table for the data:

Source of Variation	Sum of Squares (*SS*)	Degrees of Freedom (*df*)	Mean Square (*MS*)	*F*
Factor (between)	344.16	3 – 1 = 2	172.08	
Error (within)	1,219.55	14 – 3 = 11	110.87	1.5521
Total	1,563.71	14 – 1 = 13		

Table 13.43

$P(F > 1.5521) = 0.2548$
Since the *p*-value is so large, there is not good evidence against the null hypothesis of equal means. We decline to reject the null hypothesis. Thus, for 2012, there is not any good evidence of a significant difference in mean number of wins between the divisions of the American League.

A APPENDIX A REVIEW EXERCISES (CH 3–13)

These review exercises are designed to provide extra practice on concepts learned before a particular chapter. For example, the review exercises for Chapter 3 cover material learned in Chapters 1 and 2.

Chapter 3

Use the following information to answer the next six exercises. In a survey of 100 stocks on NASDAQ, the average percent increase for the past year was 9 percent for NASDAQ stocks.

1. The *average increase* for all NASDAQ stocks is the —

A. population
B. statistic
C. parameter
D. sample
E. variable

2. All of the NASDAQ stocks are —

A. population
B. statistics
C. parameter
D. sample
E. variable

3. Nine percent is —

A. population
B. statistics
C. parameter
D. sample
E. variable

4. The 100 NASDAQ stocks in the survey are —

A. population
B. statistic
C. parameter
D. sample
E. variable

5. The percent increase for one stock in the survey is —

A. population
B. statistic
C. parameter
D. sample
E. variable

6. Would the data collected by qualitative, quantitative discrete, or quantitative continuous?

Use the following information to answer the next two exercises. Thirty people spent two weeks around Mardi Gras in New Orleans. Their two-week weight gain is below. Note—a loss is shown by a negative weight gain.

Weight Gain	Frequency
–2	3
–1	5
0	2
1	4
4	13
6	2
11	1

Table A1

7. Calculate the following values:

A. The average weight gain for the two weeks
B. The standard deviation
C. The first, second, and third quartiles

8. Construct a histogram and box plot of the data.

Chapter 4

Use the following information to answer the next two exercises. A recent poll concerning credit cards found that 35 percent of respondents use a credit card that gives them a mile of air travel for every dollar they charge. Thirty percent of the respondents charge more than $2,000 per month. Of those respondents who charge more than $2,000, 80 percent use a credit card that gives them a mile of air travel for every dollar they charge.

9. What is the probability that a randomly selected respondent will spend more than $2,000 *and* use a credit card that gives them a mile of air travel for every dollar they charge?

A. (.30)(.35)
B. (.80)(.35)
C. (.80)(.30)
D. (.80)

10. Are using a credit card that gives a mile of air travel for each dollar spent *and* charging more than $2,000 per month independent events?

A. Yes
B. No, and they are not mutually exclusive either
C. No, but they are mutually exclusive
D. Not enough information given to determine the answer

11. A sociologist wants to know the opinions of employed adult women about government funding for day care. She obtains a list of 520 members of a local business and professional women's club and mails a questionnaire to 100 of these women selected at random. Sixty-eight questionnaires are returned. What is the population in this study?

A. All employed adult women
B. All the members of a local business and professional women's club
C. The 100 women who received the questionnaire
D. All employed women with children

Use the following information to answer the next two exercises. An article from the *San Jose Mercury News* was concerned with the racial mix of the 1,500 students at Prospect High School in Saratoga, CA. The table summarizes the results. Male and female values are approximate. Suppose one Prospect High School student is randomly selected.

Gender/Ethnic Group	White	Asian	Hispanic	Black	American Indian
Male	400	468	115	35	16
Female	440	132	140	40	14

Table A2

12. Find the probability that a student is Asian or male.

13. Find the probability that a student is black given that the student is female.

14. A sample of pounds lost, in a certain month, by individual members of a weight reducing clinic produced the following statistics:

- Mean = 5 lbs
- Median = 4.5 lbs
- Mode = 4 lbs
- Standard deviation = 3.8 lbs
- First quartile = 2 lbs
- Third quartile = 8.5 lbs

What is the correct statement?

A. One fourth of the members lost exactly two pounds.
B. The middle 50 percent of the members lost from two to 8.5 lbs.
C. Most people lost 3.5 to 4.5 lbs.
D. All of the choices above are correct.

15. What does it mean when a data set has a standard deviation equal to zero?

A. All values of the data appear with the same frequency.
B. The mean of the data is also zero.
C. All of the data have the same value.
D. There are no data to begin with.

16. Which statement describes the illustration?

Figure A1

A. The mean is equal to the median.
B. There is no first quartile.
C. The lowest data value is the median.
D. The median equals $\frac{Q_1+Q_3}{2}$.

17. According to a recent article in the *San Jose Mercury News* the average number of babies born with significant hearing loss—deafness—is approximately 2 per 1,000 babies in a healthy baby nursery. The number climbs to an average of 30 per 1,000 babies in an intensive care nursery. Suppose that 1,000 babies from healthy baby nurseries were

randomly surveyed. Find the probability that exactly two babies were born deaf.

18. A *friend* offers you the following *deal*: For a $10 fee, you may pick an envelope from a box containing 100 seemingly identical envelopes. However, each envelope contains a coupon for a free gift.

- Ten of the coupons are for a free gift worth $6.
- Eighty of the coupons are for a free gift worth $8.
- Six of the coupons are for a free gift worth $12.
- Four of the coupons are for a free gift worth $40.

Based upon the financial gain or loss over the long run, should you play the game?

A. Yes, I expect to come out ahead in money.
B. No, I expect to come out behind in money.
C. It doesn't matter. I expect to break even.

Use the following information to answer the next four exercises. Recently, a nurse commented that when a patient calls the medical advice line claiming to have the flu, the chance that he/she truly has the flu—and not just a nasty cold—is only about 4 percent. Of the next 25 patients calling in claiming to have the flu, we are interested in how many actually have the flu.

19. Define the random variable and list its possible values.

20. State the distribution of X.

21. Find the probability that at least four of the 25 patients actually have the flu.

22. On average, for every 25 patients calling in, how many do you expect to have the flu?

Use the following information to answer the next two exercises. Different types of writing can sometimes be distinguished by the number of letters in the words used. A student interested in this fact wants to study the number of letters of words used by Tom Clancy in his novels. She opens a Clancy novel at random and records the number of letters of the first 250 words on the page.

23. What kind of data was collected?

A. Qualitative
B. Quantitative continuous
C. Quantitative discrete

24. What is the population under study?

Chapter 5

Use the following information to answer the next five exercises. A recent study of mothers of junior high school children in Santa Clara County reported that 76 percent of the mothers are employed in paid positions. Of those mothers who are employed, 64 percent work full-time—more than 35 hours per week—and 36 percent work part-time. However, out of all of the mothers in the population, 49 percent work full-time. The population under study is made up of mothers of junior high school children in Santa Clara County. Let E = employed and F = full-time employment.

25.

A. Find the percent of all mothers in the population that are *not* employed.
B. Find the percent of mothers in the population that are employed part-time.

26. The *type of employment* is considered to be what type of data?

27. Find the probability that a randomly selected mother works part-time given that she is employed.

28. Find the probability that a randomly selected person from the population will be employed or work full-time.

29. Being employed and working part-time—

A. mutually exclusive events? Why or why not?
B. independent events? Why or why not?

Use the following additional information to answer the next two exercises. We randomly pick 10 mothers from the above population. We are interested in the number of the mothers that are employed. Let X = number of mothers that are employed.

30. State the distribution for X.

31. Find the probability that at least six are employed.

32. We expect the statistics discussion board to have, on average, 14 questions posted to it per week. We are interested in the number of questions posted to it per day.

A. Define X.
B. What are the values that the random variable may take on?
C. State the distribution for X.
D. Find the probability that from 10 to 14—inclusive—questions are posted to the listserv on a randomly picked day.

33. A person invests $1,000 into stock of a company that hopes to go public in one year. The probability that the person will lose all his money after one year, that is, his stock will be worthless, is 35 percent. The probability that the person's stock will still have a value of $1,000 after one year, that is, no profit and no loss, is 60 percent. The probability that the person's stock will increase in value by $10,000 after one year, that is, will be worth $11,000, is 5 percent. Find the expected profit after one year.

34. Rachel's piano cost $3,000. The average cost for a piano is $4,000 with a standard deviation of $2,500. Becca's guitar cost $550. The average cost for a guitar is $500 with a standard deviation of $200. Matt's drums cost $600. The average cost for drums is $700 with a standard deviation of $100. Whose cost was lowest when compared to his or her own instrument?

Figure A2

35. Explain why each statement is either true or false given the box plot in Figure A2.

A. Twenty-five percent of the data are at most five.
B. There is the same amount of data from 4–5 as there is from 5–7.
C. There are no data values of three.
D. Fifty percent of the data are four.

Using the following information to answer the next two exercises. 64 faculty members were asked the number of cars they owned—including spouse and children's cars. The results are given in the following graph.

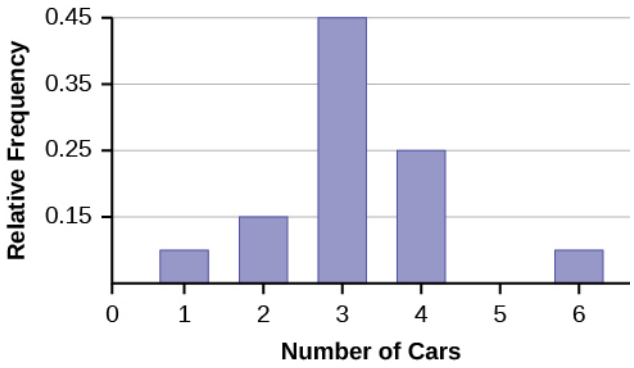


Figure A3

36. Find the approximate number of responses that were three.

37. Find the first, second, and third quartiles. Use them to construct a box plot of the data.

Use the following information to answer the next three exercises. Table A3 shows data gathered from 15 girls on the Snow Leopard soccer team when they were asked how they liked to wear their hair. Supposed one girl from the team is randomly selected.

Hair Style/Hair Color	Blond	Brown	Black
Ponytail	3	2	5
Plain	2	2	1

Table A3

38. Find the probability that the girl has black hair GIVEN that she wears a ponytail.

39. Find the probability that the girl wears her hair plain OR has brown hair.

40. Find the probability that the girl has blond hair AND that she wears her hair plain.

Chapter 6

Use the following information to answer the next two exercises. $X \sim U(3, 13)$

41. Explain which of the following are false and which are true.

A. $f(x) = \frac{1}{10}$, $3 \leq x \leq 13$
B. There is no mode.
C. The median is less than the mean.
D. $P(x > 10) = P(x \leq 6)$

42. Calculate

A. the mean,
B. the median, and
C. the 65^{th} percentile.

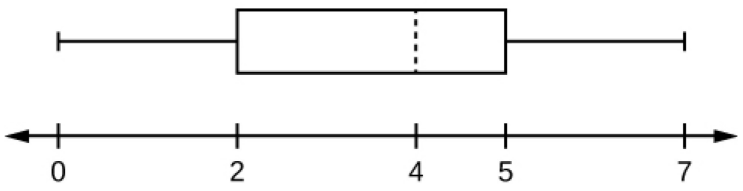


Figure A4

43. Which of the following is true for the box plot in Figure A4?

A. Twenty-five percent of the data are at most five.
B. There is about the same amount of data from 4–5 as there is from 5–7.
C. There are no data values of three.
D. Fifty percent of the data are four.

44. If $P(G|H) = P(G)$, then which of the following is correct?

A. G and H are mutually exclusive events.
B. $P(G) = P(H)$
C. Knowing that H has occurred will affect the chance that G will happen.
D. G and H are independent events.

45. If $P(J) = .3$, $P(K) = .63$, and J and K are independent events, then explain which are correct and which are incorrect.

A. $P(J \text{ AND } K) = 0$
B. $P(J \text{ OR } K) = .9$
C. $P(J \text{ OR } K) = .72$
D. $P(J) \neq P(J|K)$

46. On average, five students from each high school class get full scholarships to four-year colleges. Assume that most high school classes have about 500 students. X = the number of students from a high school class that get full scholarships to four-year schools. Which of the following is the distribution of X?

A. $P(5)$
B. $B(500, 5)$
C. $Exp\left(\frac{1}{5}\right)$
D. $N\left(5, \frac{(.01)(.99)}{500}\right)$

Chapter 7

Use the following information to answer the next three exercises. Richard's Furniture Company delivers furniture from 10 a.m. to 2 p.m. continuously and uniformly. We are interested in how long—in hours—past the 10 a.m. start time that individuals wait for their delivery.

47. $X \sim$ _____

A. $U(0, 4)$
B. $U(10, 20)$
C. $Exp(2)$
D. $N(2, 1)$

48. The average wait time is —

A. one hour
B. two hours
C. two and a half hours
D. four hours

49. Suppose that it is now past noon on a delivery day. The probability that a person must wait at least 1.5 more hours is —

A. $\frac{1}{4}$
B. $\frac{1}{2}$

C. $\frac{3}{4}$
D. $\frac{3}{8}$

50. Given $X \sim Exp\left(\frac{1}{3}\right)$

A. Find $P(x > 1)$.
B. Calculate the minimum value for the upper quartile.
C. Find $P\left(x = \frac{1}{3}\right)$

51.

- Forty percent of full-time students took four years to graduate.
- Thirty percent of full-time students took five years to graduate.
- Twenty percent of full-time students took six years to graduate.
- Ten percent of full-time students took seven years to graduate.

The expected time for full-time students to graduate is —

A. four years
B. four and a half years
C. five years
D. five and a half years

52. Which of the following distributions is described by the following example?
Many people can run a short distance of under two miles, but as the distance increases, fewer people can run that far.

A. binomial
B. uniform
C. exponential
D. normal

53. The length of time to brush one's teeth is generally thought to be exponentially distributed with a mean of $\frac{3}{4}$ minutes. Find the probability that a randomly selected person brushes his or her teeth less than $\frac{3}{4}$ minutes.

A. .5
B. $\frac{3}{4}$
C. .43
D. .63

54. Which distribution accurately describes the following situation?
The chance that a teenage boy regularly gives his mother a kiss goodnight is about 20 percent. Fourteen teenage boys are randomly surveyed. Let X = the number of teenage boys that regularly give their mother a kiss goodnight.

A. $B(14,.20)$
B. $P(2.8)$
C. $N(2.8,2.24)$
D. $Exp\left(\frac{1}{.20}\right)$

55. A 2008 report on technology use states that approximately 20 percent of U.S. households have never sent an email. Suppose that we select a random sample of fourteen U.S. households. Let X = the number of households in a 2008 sample of 14 households that have never sent an email.

A. $B(14,.20)$

B. $P(2.8)$
C. $N(2.8,2.24)$
D. $Exp\left(\frac{1}{.20}\right)$

Chapter 8

Use the following information to answer the next three exercises. Suppose that a sample of 15 randomly chosen people were put on a special weight-loss diet. The amount of weight lost, in pounds, follows an unknown distribution with mean equal to 12 pounds and standard deviation equal to three pounds. Assume that the distribution for the weight loss is normal.

56. To find the probability that the mean amount of weight lost by 15 people is no more than 14 pounds, the random variable should be _____.

A. number of people who lost weight on the special weight-loss diet
B. the number of people who were on the diet
C. the mean amount of weight lost by 15 people on the special weight-loss diet
D. the total amount of weight lost by 15 people on the special weight-loss diet

57. Find the probability asked for in Question 56.

58. Find the 90^{th} percentile for the mean amount of weight lost by 15 people.

Using the following information to answer the next three exercises. The time of occurrence of the first accident during rush-hour traffic at a major intersection is uniformly distributed between the three hour interval 4 p.m. to 7 p.m. Let X = the amount of time—hours—it takes for the first accident to occur.

59. What is the probability that the time of occurrence is within the first half-hour or the last hour of the period from 4 to 7 p.m.?

A. It cannot be determined from the information given.
B. $\frac{1}{6}$
C. $\frac{1}{2}$
D. $\frac{1}{3}$

60. The 20^{th} percentile occurs after how many hours?

A. .20
B. .60
C. .50
D. 1

61. Assume Ramon has kept track of the times for the first accidents to occur for 40 different days. Let C = the total cumulative time. Then C follows which distribution?

A. $U(0,3)$
B. $Exp(13)$
C. $N(60, 5.477)$
D. $N(1.5, .01875)$

62. Using the information in Question 61, find the probability that the total time for all first accidents to occur is more than 43 hours.

Use the following information to answer the next two exercises. The length of time a parent must wait for his children to clean their rooms is uniformly distributed in the time interval from one to 15 days.

63. How long must a parent expect to wait for his children to clean their rooms?

A. 8 days
B. 3 days
C. 14 days
D. 6 days

64. What is the probability that a parent will wait more than six days given that the parent has already waited more than three days?

A. .5174
B. .0174
C. .7500
D. .2143

Use the following information to answer the next five exercises. Twenty percent of the students at a local community college live in within five miles of the campus. Thirty percent of the students at the same community college receive some kind of financial aid. Of those who live within five miles of the campus, 75 percent receive some kind of financial aid.

65. Find the probability that a randomly chosen student at the local community college does not live within five miles of the campus.

A. 80 percent
B. 20 percent
C. 30 percent
D. Cannot be determined

66. Find the probability that a randomly chosen student at the local community college lives within five miles of the campus or receives some kind of financial aid.

A. 50 percent
B. 35 percent
C. 27.5 percent
D. 75 percent

67. Are living in student housing within five miles of the campus and receiving some kind of financial aid mutually exclusive?

A. Yes
B. No
C. Cannot be determined

68. The interest rate charged on the financial aid is _____ data.

A. Quantitative discrete
B. Quantitative continuous
C. Qualitative discrete
D. Qualitative

69. The following information is about the students who receive financial aid at the local community college.

- 1st quartile = $250
- 2nd quartile = $700
- 3rd quartile = $1,200

These amounts are for the school year. If a sample of 200 students is taken, how many are expected to receive $250 or

more?

A. 50
B. 250
C. 150
D. Cannot be determined

Use the following information to answer the next two exercises. $P(A) = .2$, $P(B) = .3$; A and B are independent events.

70. $P(A \text{ AND } B) =$ —

A. .5
B. .6
C. 0
D. .06

71. $P(A \text{ OR } B) =$ —

A. .56
B. .5
C. .44
D. 1

72. If H and D are mutually exclusive events, $P(H) = .25$, $P(D) = .15$, then $P(H|D)$.

A. 1
B. 0
C. .40
D. .0375

Chapter 9

73. Rebecca and Matt are 14 year old twins. Matt's height is two standard deviations below the mean for 14 year old boys' height. Rebecca's height is .10 standard deviations above the mean for 14 year old girls' height. Interpret this.

A. Matt is 2.1 inches shorter than Rebecca.
B. Rebecca is very tall compared to other 14 year old girls.
C. Rebecca is taller than Matt.
D. Matt is shorter than the average 14 year old boy.

74. Construct a histogram of the IPO data (see Appendix C Data Sets).

Use the following information to answer the next three exercises. Ninety homeowners were asked the number of estimates they obtained before having their homes fumigated. Let X = the number of estimates.

x	Relative Frequency	Cumulative Relative Frequency
1	.3	
2	.2	
4	.4	
5	.1	

Table A4

75. Complete the cumulative frequency column.

76. Calculate the sample mean (a), the sample standard deviation (b), and the percent of the estimates that fall at or below four (c).

77. Calculate the median, M, the first quartile, Q_1, and the third quartile Q_3. Then construct a box plot of the data.

78. The middle 50 percent of the data are between _____ and _____.

Use the following information to answer the next three exercises. Seventy fifth and sixth graders were asked their favorite dinner.

	Pizza	Hamburgers	Spaghetti	Fried Shrimp
5th Grader	15	6	9	0
6th Grader	15	7	10	8

Table A5

79. Find the probability that one randomly chosen child is in the 6th grade and prefers fried shrimp.

A. $\frac{32}{70}$
B. $\frac{8}{32}$
C. $\frac{8}{8}$
D. $\frac{8}{70}$

80. Find the probability that a child does not prefer pizza.

A. $\frac{30}{70}$
B. $\frac{30}{40}$
C. $\frac{40}{70}$
D. 1

81. Find the probability a child is in the fifth grade given that the child prefers spaghetti.

A. $\frac{9}{19}$
B. $\frac{9}{70}$
C. $\frac{9}{30}$
D. $\frac{19}{70}$

82. A sample of convenience is a random sample.

A. True
B. False

83. A statistic is a number that is a property of the population.

A. True
B. False

84. You should always throw out any data that are outliers.

A. True
B. False

85. Lee bakes pies for a small restaurant in Felton, CA. She generally bakes 20 pies in a day, on average. Of interest is the number of pies she bakes each day.

A. Define the random variable X.
B. State the distribution for X.
C. Find the probability that Lee bakes more than 25 pies in any given day.

86. Six different brands of Italian salad dressing were randomly selected at a supermarket. The grams of fat per serving are 7, 7, 9, 6, 8, and 5. Assume that the underlying distribution is normal. Calculate a 95 percent confidence interval for the population mean grams of fat per serving of Italian salad dressing sold in supermarkets.

87. Given: uniform, exponential, normal distributions. Match each to a statement below.

A. mean = median ≠ mode
B. mean > median > mode
C. mean = median = mode

Chapter 10

Use the following information to answer the next three exercises. In a survey at Kirkwood Ski Resort the following information was recorded.

	0–10	11–20	21–40	40+
Ski	10	12	30	8
Snowboard	6	17	12	5

Table A6

Suppose that one person from Table A6 was randomly selected.

88. Find the probability that the person was a skier or was age 11–20.

89. Find the probability that the person was a snowboarder given he or she was age 21–40.

90. Explain which of the following are true and which are false.

A. Sport and age are independent events.
B. Ski and age 11–20 are mutually exclusive events.
C. $P(\text{Ski AND age 21–40}) < P(\text{Ski} \mid \text{age 21–40})$
D. $P(\text{Snowboard OR age 0–10}) < P(\text{Snowboard} \mid \text{age 0–10})$

91. The average length of time a person with a broken leg wears a cast is approximately six weeks. The standard deviation is about three weeks. Thirty people who had recently healed from broken legs were interviewed. State the distribution that most accurately reflects total time to heal for the 30 people.

92. The distribution for X is uniform. What can we say for certain about the distribution for $\overline{X}$ when $n = 1$?

A. The distribution for $\overline{X}$ is still uniform with the same mean and standard deviation as the distribution for X.
B. The distribution for $\overline{X}$ is normal with the different mean and a different standard deviation as the distribution for X.
C. The distribution for $\overline{X}$ is normal with the same mean but a larger standard deviation than the distribution for X.
D. The distribution for $\overline{X}$ is normal with the same mean but a smaller standard deviation than the distribution for X.

93. The distribution for X is uniform. What can we say for certain about the distribution for $\sum X$ when $n = 50$?

A. The distribution for $\sum X$ is still uniform with the same mean and standard deviation as the distribution for X.

B. The distribution for $\sum X$ is normal with the same mean but a larger standard deviation as the distribution for *X*.
C. The distribution for $\sum X$ is normal with a larger mean and a larger standard deviation than the distribution for *X*.
D. The distribution for $\sum X$ is normal with the same mean but a smaller standard deviation than the distribution for *X*.

Use the following information to answer the next three exercises. A group of students measured the lengths of all the carrots in a five-pound bag of baby carrots. They calculated the average length of baby carrots to be 2.0 inches with a standard deviation of 0.25 inches. Suppose we randomly survey 16 five-pound bags of baby carrots.

94. State the approximate distribution for $\overline{X}$, the distribution for the average lengths of baby carrots in 16 five-pound bags. $\overline{X}$ ~ _____.

95. Explain why we cannot find the probability that one individual randomly chosen carrot is greater than 2.25 inches.

96. Find the probability that $\overline{x}$ is between 2.0 and 2.25 inches.

Use the following information to answer the next three exercises. At the beginning of the term, the amount of time a student waits in line at the campus store is normally distributed with a mean of five minutes and a standard deviation of two minutes.

97. Find the 90^{th} percentile of waiting time in minutes.

98. Find the median waiting time for one student.

99. Find the probability that the average waiting time for 40 students is at least 4.5 minutes.

Chapter 11

Use the following information to answer the next four exercises. Suppose that the time that owners keep their cars—purchased new—is normally distributed with a mean of seven years and a standard deviation of two years. We are interested in how long an individual keeps his car—purchased new. Our population is people who buy their cars new.

100. Sixty percent of individuals keep their cars *at most* how many years?

101. Suppose that we randomly survey one person. Find the probability that person keeps his or her car *less than* 2.5 years.

102. If we are to pick individuals 10 at a time, find the distribution for the *mean* car length ownership.

103. If we are to pick 10 individuals, find the probability that the *sum* of their ownership time is more than 55 years.

104. For which distribution is the median not equal to the mean?

A. Uniform
B. Exponential
C. Normal
D. Student *t*

105. Compare the standard normal distribution to the Student's *t* distribution, centered at zero. Explain which of the following are true and which are false.

A. As the number surveyed increases, the area to the left of –1 for the Student's *t* distribution approaches the area for the standard normal distribution.
B. As the degrees of freedom decrease, the graph of the Student's *t* distribution looks more like the graph of the standard normal distribution.
C. If the number surveyed is 15, the normal distribution should never be used.

Use the following information to answer the next five exercises. We are interested in the checking account balance of 24-old college students. We randomly survey 16 20-year-old college students. We obtain a sample mean of $640 and a sample standard deviation of $150. Let *X* = checking account balance of an individual 20-year-old college student.

106. Explain why we cannot determine the distribution of *X*.

107. If you were to create a confidence interval or perform a hypothesis test for the population mean checking account balance of 20-year-old college students, what distribution would you use?

108. Find the 95 percent confidence interval for the true mean checking account balance of a 20-year-old college student.

109. What type of data is the balance of the checking account considered to be?

110. What type of data is the number of 20-year-olds considered to be?

111. On average, a busy emergency room gets a patient with a shotgun wound about once per week. We are interested in the number of patients with a shotgun wound the emergency room gets per 28 days.

A. Define the random variable *X*.
B. State the distribution for *X*.
C. Find the probability that the emergency room gets no patients with shotgun wounds in the next 28 days.

Use the following information to answer the next two exercises. The probability that a certain slot machine will pay back money when a quarter is inserted is .30. Assume that each play of the slot machine is independent from each other. A person puts in 15 quarters for 15 plays.

112. Is the expected number of plays of the slot machine that will pay back money greater than, less than, or the same as the median? Explain your answer.

113. Is it likely that exactly eight of the 15 plays would pay back money? Justify your answer numerically.

114. A game is played with the following rules:

- It costs $10 to enter.
- A fair coin is tossed four times.
- If you do not get four heads or four tails, you lose your $10.
- If you get four heads or four tails, you get back your $10, plus $30 more.

Over the long run of playing this game, what are your expected earnings?

115.

- The mean grade on a math exam in Rachel's class was 74, with a standard deviation of five. Rachel earned an 80.
- The mean grade on a math exam in Becca's class was 47, with a standard deviation of two. Becca earned a 51.
- The mean grade on a math exam in Matt's class was 70, with a standard deviation of eight. Matt earned an 83.

Find whose score was the best, compared to his or her own class. Justify your answer numerically.

Use the following information to answer the next two exercises. A random sample of 70 compulsive gamblers were asked the number of days they go to casinos per week. The results are given in the following graph.

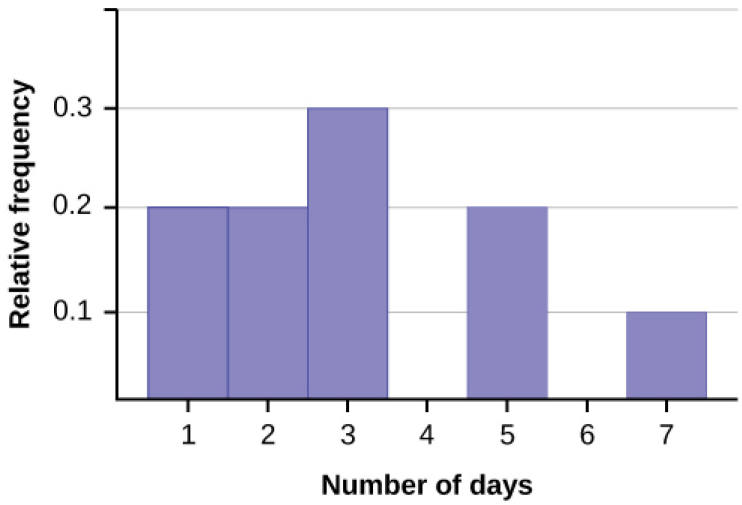


Figure A5

116. Find the number of responses that were five.

117. Find the mean, standard deviation, the median, the first quartile, the third quartile, and the *IQR*.

118. Based upon research at De Anza College, it is believed that about 19 percent of the student population speaks a language other than English at home. Suppose that a study was done this year to see if that percent has decreased. Ninety-eight students were randomly surveyed with the following results: Fourteen said that they speak a language other than English at home.

A. State an appropriate null hypothesis.
B. State an appropriate alternative hypothesis.
C. Define the random variable, P'.
D. Calculate the test statistic.
E. Calculate the *p*-value.
F. At the 5 percent level of decision, what is your decision about the null hypothesis?
G. What is the Type I error?
H. What is the Type II error?

119. Assume that you are an emergency paramedic called in to rescue victims of an accident. You need to help a patient who is bleeding profusely. The patient is also considered to be a high risk for contracting a blood-borne illness. Assume that the null hypothesis is that the patient does *not* have the a blood-borne illness. What is a Type I error?

120. It is often said that Californians are more casual than the rest of Americans. Suppose that a survey was done to see if the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals. Fifty of each was surveyed with the following results: Fifteen Californians wear jeans to work and six non-Californians wear jeans to work.
Let *C* = Californian professional; *NC* = non-Californian professional

A. State appropriate null and alternate hypotheses.
B. Define the random variable.
C. Calculate the test statistic and *p*-value.
D. At the 5 percent significance level, what is your decision?
E. What is the Type I error?
F. What is the Type II error?

Use the following information to answer the next two exercises. A group of statistics students have developed a technique that they feel will lower their anxiety level on statistics exams. They measured their anxiety level at the start of the quarter and again at the end of the quarter. Recorded is the paired data in that order: (1,000, 900); (1,200, 1,050); (600, 700); (1,300, 1,100); (1,000, 900); (900, 900).

121. This is a test of (pick the best answer) —

A. large samples, and independent means
B. small samples, and independent means
C. dependent means

122. State the distribution to use for the test.

Chapter 12

Use the following information to answer the next two exercises. A recent survey of U.S. teenagers was answered by 720 teenagers, age 15–18. Six percent of teenagers surveyed said they are planning on going to college in another country. We are interested in the true proportion of U.S. teens, ages 15–18, who are planning on going to college in another country.

123. Find the 95 percent confidence interval for the true proportion of U.S. teens, ages 15–19, who are planning to go to college in another country.

124. The report also stated that the results of the survey are accurate to within ±3.7 percent at the 95 percent confidence level. Suppose that a new study is to be done. It is desired to be accurate to within 2 percent of the 95 percent confidence level. What is the minimum number that should be surveyed?

125. Given $X \sim Exp\left(\frac{1}{3}\right)$. Sketch the graph that depicts: $P(x > 1)$.

Use the following information to answer the next three exercises. The amount of money a customer spends in one trip to the supermarket is known to have an exponential distribution. Suppose the mean amount of money a customer spends in one trip to the supermarket is $72.

126. Find the probability that one customer spends less than $72 in one trip to the supermarket?

127. Suppose five customers pool their money. How much money altogether would you expect the five customers to spend in one trip to the supermarket in dollars?

128. State the distribution to use if you want to find the probability that the mean amount spent by five customers in one trip to the supermarket is less than $60.

Chapter 13

Use the following information to answer the next two exercises. Suppose that the probability of a drought in any independent year is 20 percent. Out of those years in which a drought occurs, the probability of water rationing is 10 percent. However, in any year, the probability of water rationing is 5 percent.

129. What is the probability of both a drought *and* water rationing occurring?

130. Out of the years with water rationing, find the probability that there is a drought.

Use the following information to answer the next three exercises.

	Apple	Pumpkin	Pecan
Female	40	10	30
Male	20	30	10

Table A7

131. Suppose that one individual is randomly chosen. Find the probability that the person's favorite pie is apple *or* the person is male.

132. Suppose that one male is randomly chosen. Find the probability his favorite pie is pecan.

133. Conduct a hypothesis test to determine if favorite pie type and gender are independent.

Use the following information to answer the next two exercises. Let's say that the probability that an adult watches the news at least once per week is .60.

134. We randomly survey 14 people. On average, how many people do we expect to watch the news at least once per

week?

135. We randomly survey 14 people. Of interest is the number that watch the news at least once per week. State the distribution of X. $X \sim$ _____.

136. The following histogram is most likely to be a result of sampling from which distribution?

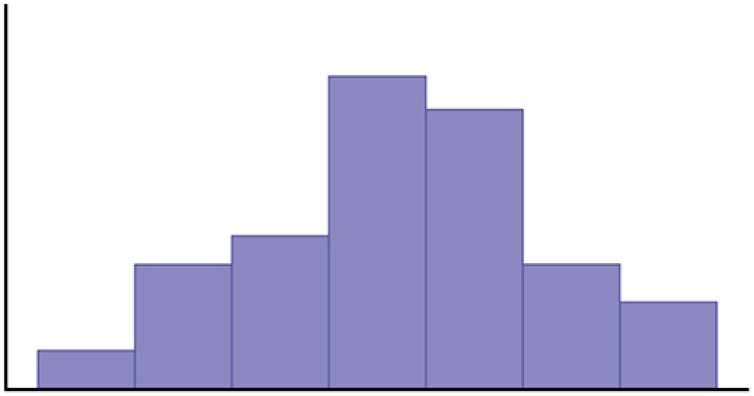


Figure A6

A. Chi-square
B. Geometric
C. Uniform
D. Binomial

137. The ages of De Anza evening students is known to be normally distributed with a population mean of 40 and a population standard deviation of six. A sample of six De Anza evening students reported their ages in years as: 28; 35; 47; 45; 30; 50. Find the probability that the mean of six ages of randomly chosen students is less than 35 years. Hint—Find the sample mean.

138. A math exam was given to all the fifth grade children attending Country School. Two random samples of scores were taken. The null hypothesis is that the mean math scores for boys and girls in fifth grade are the same. Conduct a hypothesis test.

	n	$\bar{x}$	s^2
Boys	55	82	29
Girls	60	86	46

Table A8

139. In a survey of 80 males, 55 had played an organized sport growing up. Of the 70 females surveyed, 25 had played an organized sport growing up. We are interested in whether the proportion for males is higher than the proportion for females. Conduct a hypothesis test.

140. Which of the following is preferable when designing a hypothesis test?

A. Maximize α and minimize β
B. Minimize α and maximize β
C. Maximize α and β
D. Minimize α and β

Use the following information to answer the next three exercises. One hundred twenty people were surveyed as to their favorite beverage. The results are below.

Beverage/Age	0–9	10–19	20–29	30+	Totals
Milk	14	10	6	0	30
Soda	3	8	26	15	52
Juice	7	12	12	7	38
Totals	24	330	44	22	120

Table A9

141. Are the events of milk and 30+—

A. independent events? Justify your answer.
B. mutually exclusive events? Justify your answer.

142. Suppose that one person is randomly chosen. Find the probability that person is 10–19 given that he or she prefers juice.

143. Are Preferred Beverage and Age independent events? Conduct a hypothesis test.

144. Given the following histogram, which distribution is the data most likely to come from?

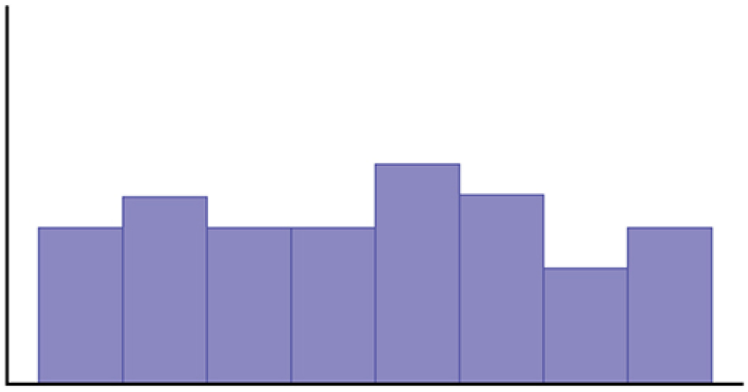


Figure A7

A. Uniform
B. Exponential
C. Normal
D. Chi-square

Solutions

Chapter 3

1. C Parameter

2. A Population

3. B Statistic

4. D Sample

5. E Variable

6. quantitative continuous

7.

A. 2.27
B. 3.04
C. –1, 4, 4

8. Answers will vary.

Chapter 4

9. C (.80)(.30)

10. B No, and they are not mutually exclusive either.

11. A All employed adult women

12. .5773

13. .0522

14. B The middle fifty percent of the members lost from 2 to 8.5 lbs.

15. C All of the data have the same value.

16. C The lowest data value is the median.

17. .279

18. B No, I expect to come out behind in money.

19. X = the number of patients calling in claiming to have the flu, who actually have the flu.
X = 0, 1, 2, ...25

20. B(25, .04)

21. .0165

22. 1

23. C Quantitative discrete

24. all words used by Tom Clancy in his novels

Chapter 5

25.

A. 24 percent
B. 27 percent

26. qualitative

27. .36

28. .7636

29.

A. no
B. no

30. B(10, .76)

31. .9330

32.

A. X = the number of questions posted to the statistics listserv per day.
B. X = 0, 1, 2,...
C. $X \sim P(2)$

D. 0

33. $150

34. Matt

35.

A. False
B. True
C. False
D. False

36. 16

37. first quartile: 2
second quartile: 2
third quartile: 3

38. 0.5

39. $\frac{7}{15}$

40. $\frac{2}{15}$

Chapter 6

41.

A. True
B. True
C. False – the median and the mean are the same for this symmetric distribution.
D. True

42.

A. 8
B. 8
C. $P(x < k) = 0.65 = (k - 3)\left(\frac{1}{10}\right)$. $k = 9.5$

43.

A. False – $\frac{3}{4}$ of the data are at most five.
B. True – each quartile has 25 percent of the data.
C. False – that is unknown.
D. False – 50 percent of the data are four or less.

44. D *G* and *H* are independent events.

45.

A. False – *J* and *K* are independent so they are not mutually exclusive which would imply dependency (meaning *P*(*J* AND *K*) is not 0).
B. False – see answer c.
C. True – *P*(*J* OR *K*) = *P*(*J*) + *P*(*K*) – *P*(*J* AND *K*) = *P*(*J*) + *P*(*K*) – *P*(*J*)*P*(*K*) = .3 + .6 – (.3)(.6) = .72. Note the *P*(*J* AND *K*) = *P*(*J*)*P*(*K*) because *J* and *K* are independent.
D. False – *J* and *K* are independent so *P*(*J*) = *P*(*J* | *K*).

46. A $P(5)$

Chapter 7

47. A $U(0, 4)$

48. B 2 hours

49. A $\frac{1}{4}$

50.

A. .7165
B. 4.16
C. 0

51. C 5 years

52. C exponential

53. .63

54. A $B(14, .20)$

55. A $B(14, .20)$

Chapter 8

56. C The mean amount of weight lost by 15 people on the special weight-loss diet.

57. .9951

58. 12.99

59. C $\frac{1}{2}$

60. B .60

61. C $N(60, 5.477)$

62. .9990

63. A eight days

64. C .7500

65. A 80 percent

66. B 35 percent

67. B no

68. B Quantitative continuous

69. C 150

70. D .06

71. C .44

72. B 0

Chapter 9

73. D Matt is shorter than the average 14 year old boy.

74. Answers will vary.

75.

x	Relative Frequency	Cumulative Relative Frequency
1	.3	.3
2	.2	.2
4	.4	.4
5	.1	.1

Table A10

76.

A. 2.8
B. 1.48
C. 90 percent

77. $M = 3$; $Q_1 = 1$; $Q_3 = 4$

78. 1 and 4

79. D $\frac{8}{70}$

80. C $\frac{40}{70}$

81. A $\frac{9}{19}$

82. B False

83. B False

84. B False

85.

A. X = the number of pies Lee bakes every day.
B. $P(20)$
C. .1122

86. CI: (5.25, 8.48)

87.

A. uniform
B. exponential
C. normal

Chapter 10

88. $\frac{77}{100}$

89. $\frac{12}{42}$

90.

A. False
B. False
C. True
D. False

91. $N(180, 16.43)$

92. A The distribution for $\overline{X}$ is still uniform with the same mean and standard deviation as the distribution for X.

93. C The distribution for $\sum X$ is normal with a larger mean and a larger standard deviation than the distribution for X.

94. $N\left(2, \frac{.25}{\sqrt{16}}\right)$

95. Answers will vary.

96. .5000

97. 7.6

98. 5

99. .9431

Chapter 11

100. 7.5

101. .0122

102. $N(7, .63)$

103. .9911

104. B exponential

105.

A. True
B. False
C. False

106. Answers will vary.

107. Student's t with $df = 15$

108. (560.07, 719.93)

109. quantitative continuous data

110. quantitative discrete data

111.

A. X = the number of patients with a shotgun wound the emergency room gets per 28 days.
B. $P(4)$
C. .0183

112. greater than

113. no; $P(x = 8) = .0348$

114. You will lose $5.

115. Becca

116. 14

117. sample mean = 3.2
sample standard deviation = 1.85
median = 3
$Q_1 = 2$
$Q_3 = 5$

IQR = 3

118. d. $z = -1.19$
e. .1171
f. Do not reject the null hypothesis.

119. We conclude that the patient does have the illness when, in fact, the patient does not.

120. c. $z = 2.21$; $p = .0136$
d. Reject the null hypothesis.
e. We conclude that the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals when, in fact, it is not greater.
f. We cannot conclude that the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals when, in fact, it is greater.

121. C dependent means

122. t_5

Chapter 12

123. (.0424, .0770)

124. 2,401

125. Check student's solution.

126. .6321

127. $360

128. $N\left(72, \frac{72}{\sqrt{5}}\right)$

Chapter 13

129. .02

130. .40

131. $\frac{100}{140}$

132. $\frac{10}{60}$

133. *p*-value = 0; reject the null hypothesis; conclude that they are dependent events

134. 8.4

135. $B(14, .60)$

136. D Binomial

137. .3669

138. *p*-value = .0006; reject the null hypothesis; conclude that the averages are not equal

139. *p*-value = 0; reject the null hypothesis; conclude that the proportion of males is higher

140. minimize α and β

141.

A. no
B. yes, $P(M \text{ AND } 30+) = 0$

142. $\frac{12}{38}$

143. no; *p*-value = 0

144. A uniform

References

Baran, D. (2010). Twenty percent of Americans have never used email. Retrieved from http://www.webguild.org/20080519/20-percent-of-americans-have-never-used-email.

Parade Magazine. (n.d.). Retrieved from https://parade.com/.

San Jose Mercury News. (n.d.). Retrieved from http://www.mercurynews.com/.

B APPENDIX B PRACTICE TESTS (1–4) AND FINAL EXAMS

Practice Test 1

1.1: Definitions of Statistics, Probability, and Key Terms

Use the following information to answer the next three exercises. A grocery store is interested in how much money, on average, their customers spend each visit in the produce department. Using their store records, they draw a sample of 1,000 visits and calculate each customer's average spending on produce.

1. Identify the population, sample, parameter, statistic, variable, and data for this example.

A. population
B. sample
C. parameter
D. statistic
E. variable
F. data

2. What kind of data is *amount of money spent on produce per visit*?

A. Qualitative
B. Quantitative-continuous
C. Quantitative-discrete

3. The study finds that the mean amount spent on produce per visit by the customers in the sample is $12.84. This is an example of a

A. Population
B. Sample
C. Parameter
D. Statistic
E. Variable

1.2: Data, Sampling, and Variation in Data and Sampling

Use the following information to answer the next two exercises. A health club is interested in knowing how many times a typical member uses the club in a week. They decide to ask every tenth customer on a specified day to complete a short survey, including information about how many times they have visited the club in the past week.

4. What kind of a sampling design is this?

A. Cluster
B. Stratified
C. Simple random
D. Systematic

5. *Number of visits per week* is what kind of data?

A. Qualitative
B. Quantitative-continuous
C. Quantitative-discrete

6. Describe a situation in which you would calculate a parameter, rather than a statistic.

7. The U.S. federal government conducts a survey of high school seniors concerning their plans for future education and

employment. One question asks whether they are planning to attend a four-year college or university in the following year. Fifty percent answer yes to this question. That 50 percent is a

A. Parameter
B. Statistic
C. Variable
D. Data

8. Imagine that the U.S. federal government had the means to survey all high school seniors in the United States concerning their plans for future education and employment, and found that 50 percent were planning to attend a four-year college or university in the following year. This 50 percent is an example of a

A. Parameter
B. Statistic
C. Variable
D. Data

Use the following information to answer the next three exercises. A survey of a random sample of 100 nurses working at a large hospital asked how many years they had been working in the profession. Their answers are summarized in the following (incomplete) table.

9. Fill in the blanks in the table and round your answers to two decimal places for the Relative Frequency and Cumulative Relative Frequency cells.

# of years	Frequency	Relative Frequency	Cumulative Relative Frequency
< 5	25		
5–10	30		
> 10	empty		

Table B1

10. What proportion of nurses have five or more years of experience?

11. What proportion of nurses have 10 or fewer years of experience?

12. Describe how you might draw a random sample of 30 students from a lecture class of 200 students.

13. Describe how you might draw a stratified sample of students from a college, where the strata are the students' class standing (freshman, sophomore, junior, or senior).

14. A manager wants to draw a sample, without replacement, of 30 employees from a workforce of 150. Describe how the chance of being selected will change over the course of drawing the sample.

15. The manager of a department store decides to measure employee satisfaction by selecting four departments at random, and conducting interviews with all the employees in those four departments. What type of survey design is this?

A. Cluster
B. Stratified
C. Simple random
D. Systematic

16. A popular American television sports program conducts a poll of viewers to see which team they believe will win the National Football League (NFL) championship this year. Viewers vote by calling a number displayed on the television screen and telling the operator which team they think will win. Do you think that those who participate in this poll are representative of all football fans in America?

17. Two researchers studying vaccination rates independently draw samples of 50 children, aged three–18 months, from

a large urban area, and determine if they are up to date on their vaccinations. One researcher finds that 84 percent of the children in her sample are up to date, and the other finds that 86 percent in his sample are up to date. Assuming both followed proper sampling procedures and did their calculations correctly, what is a likely explanation for this discrepancy?

18. A high school increased the length of the school day from 6.5 to 7.5 hours. Students who wished to attend this high school were required to sign contracts pledging to put forth their best effort on their school work and to obey the school rules; if they did not wish to do so, they could attend another high school in the district. At the end of one year, student performance on statewide tests had increased by 10 percentage points over the previous year. Does this prove that a longer school day improves student achievement?

19. You read a newspaper article reporting that eating almonds leads to increased life satisfaction. The study was conducted by the Almond Growers Association, and was based on a randomized survey asking people about their consumption of various foods, including almonds, and also about their satisfaction with different aspects of their life. Does anything about this poll lead you to question its conclusion?

20. Why is non-response a problem in surveys?

1.3: Frequency, Frequency Tables, and Levels of Measurement

21. Compute the mean of the following numbers, and report your answer using one more decimal place than is present in the original data:
14, 5, 18, 23, 6

1.4: Experimental Design and Ethics

22. A psychologist is interested in whether the size of tableware (bowls, plates, etc.) influences how much college students eat. He randomly assigns 100 college students to one of two groups. The first is served a meal using normal-sized tableware, while the second is served the same meal but using tableware that it 20 percent smaller than normal. He records how much food is consumed by each group. Identify the following components of this study.

A. population
B. sample
C. experimental units
D. explanatory variable
E. treatment
F. response variable

23. A researcher analyzes the results of the Scholastic Aptitude Test (SAT) over a five-year period and finds that male students on average score higher on the math section, and female students on average score higher on the verbal section. She concludes that these observed differences in test performance are due to genetic factors. Explain how lurking variables could offer an alternative explanation for the observed differences in test scores.

24. Explain why it would not be possible to use random assignment to study the health effects of exercise.

25. A professor conducts a telephone survey of a city's population by drawing a sample of numbers from the phone book and having her student assistants call each of the selected numbers once to administer the survey. What are some sources of bias with this survey?

26. A professor offers extra credit to students who take part in her research studies. What is an ethical problem with this method of recruiting subjects?

2.1: Stem-and Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

Use the following information to answer the next four exercises. The midterm grades on a chemistry exam, graded on a scale of 0 to 100, were
62, 64, 65, 65, 68, 70, 72, 72, 74, 75, 75, 75, 76, 78, 78, 81, 83, 83, 84, 85, 87, 88, 92, 95, 98, 98, 100, 100, 740

27. Do you see any outliers in this data? If so, how would you address the situation?

28. Construct a stem plot for this data, using only the values in the range zero–100.

29. Describe the distribution of exam scores.

2.2: Histograms, Frequency Polygons, and Time Series Graphs

30. In a class of 35 students, seven students received scores in the 70–79 range. What is the relative frequency of scores in this range?

Use the following information to answer the next three exercises. You conduct a poll of 30 students to see how many classes they are taking this term. Your results are
1; 1; 1; 1
2; 2; 2; 2; 2
3; 3; 3; 3; 3; 3; 3; 3
4; 4; 4; 4; 4; 4; 4; 4; 4
5; 5; 5; 5

31. You decide to construct a histogram of this data. What will be the range of your first bar, and what will be the central point?

32. What will be the widths and central points of the other bars?

33. Which bar in this histogram will be the tallest, and what will be its height?

34. You get data from the U.S. Census Bureau on the median household income for your city, and decide to display it graphically. Which is the better choice for this data, a bar graph or a histogram?

35. You collect data on the color of cars driven by students in your statistics class, and want to display this information graphically. Which is the better choice for this data, a bar graph or a histogram?

2.3: Measures of the Location of the Data

36. Your daughter brings home test scores showing that she scored in the 80^{th} percentile in math and the 76^{th} percentile in reading for her grade. Interpret these scores.

37. You have to wait 90 minutes in the emergency room of a hospital before you can see a doctor. You learn that your wait time was in the 82^{nd} percentile of all wait times. Explain what this means, and whether you think it is good or bad.

2.4: Box Plots

Use the following information to answer the next three exercises. 1; 1; 2; 3; 4; 4; 5; 5; 6; 7; 7; 8; 9

38. What is the median for this data?

39. What is the first quartile for this data?

40. What is the third quartile for this data?

Use the following information to answer the next four exercises. This box plot represents scores on the final exam for a physics class.

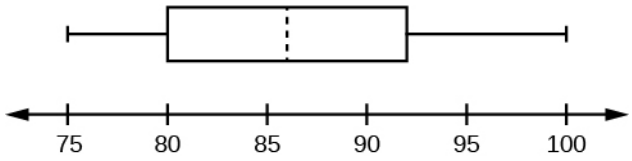


Figure B1

41. What is the median for this data, and how do you know?

42. What are the first and third quartiles for this data, and how do you know?

43. What is the interquartile range for this data?

44. What is the range for this data?

2.5: Measures of the Center of the Data

45. In a marathon, the median finishing time was 3:35:04 (three hours, 35 minutes, and four seconds). You finished in 3:34:10. Interpret the meaning of the median time, and discuss your time in relation to it.

Use the following information to answer the next three exercises. The values, in thousands of dollars, for houses on a block, are 45; 47; 47.5; 51; 53.5; 125.

46. Calculate the mean for this data.

47. Calculate the median for this data.

48. Which do you think better reflects the average value of the homes on this block?

2.6: Skewness and the Mean, Median, and Mode

49. In a left-skewed distribution, which is greater?

A. The mean
B. The media
C. The mode

50. In a right-skewed distribution, which is greater?

A. The mean
B. The median
C. The mode

51. In a symmetrical distribution, what will be the relationship among the mean, median, and mode?

2.7: Measures of the Spread of the Data

Use the following information to answer the next four exercises. 10; 11; 15; 15; 17; 22

52. Compute the mean and standard deviation for this data; use the sample formula for the standard deviation.

53. What number is two standard deviations above the mean of this data?

54. Express the number 13.7 in terms of the mean and standard deviation of this data.

55. In a biology class, the scores on the final exam were normally distributed, with a mean of 85 and a standard deviation of five. Susan got a final exam score of 95. Express her exam result as a *z* score, and interpret its meaning.

3.1: Terminology

Use the following information to answer the next two exercises. You have a jar full of marbles: 50 are red, 25 are blue, and 15 are yellow. Assume you draw one marble at random for each trial and replace it before the next trial.
Let $P(R)$ = the probability of drawing a red marble.
Let $P(B)$ = the probability of drawing a blue marble.
Let $P(Y)$ = the probability of drawing a yellow marble.

56. Find $P(B)$.

57. Which is more likely, drawing a red marble or a yellow marble? Justify your answer numerically.

Use the following information to answer the next two exercises. The following are probabilities describing a group of college students.
Let $P(M)$ = the probability that the student is male
Let $P(F)$ = the probability that the student is female
Let $P(E)$ = the probability the student is majoring in education
Let $P(S)$ = the probability the student is majoring in science

58. Write the symbols for the probability that a student, selected at random, is both female and a science major.

59. Write the symbols for the probability that the student is an education major, given that the student is male.

3.2: Independent and Mutually Exclusive Events

60. Events A and B are independent.
If $P(A) = 0.3$ and $P(B) = 0.5$, find $P(A \text{ AND } B)$.

61. C and D are mutually exclusive events.
If $P(C) = 0.18$ and $P(D) = 0.03$, find $P(C \text{ OR } D)$.

3.3: Two Basic Rules of Probability

62. In a high school graduating class of 300, 200 students are going to college, 40 are planning to work full-time, and 80 are taking a gap year. Are these events mutually exclusive?

Use the following information to answer the next two exercises. An archer hits the center of the target (the bullseye) 70 percent of the time. However, she is a streak shooter, and if she hits the center on one shot, her probability of hitting it on the shot immediately following is 0.85. Written in probability notation
$P(A) = P(B) = P$(hitting the center on one shot) = 0.70
$P(B|A)$ = P(hitting the center on a second shot, given that she hit it on the first) = 0.85

63. Calculate the probability that she will hit the center of the target on two consecutive shots.

64. Are $P(A)$ and $P(B)$ independent in this example?

3.4: Contingency Tables

Use the following information to answer the next three exercises. The following contingency table displays the number of students who report studying at least 15 hours per week, and how many made the honor roll in the past semester.

	Honor Roll	No Honor Roll	Total
Study at least 15 hours/week		200	
Study less than 15 hours/week	125	193	
Total			1,000

Table B2

65. Complete the table.

66. Find P(honor roll|study at least 15 hours per week).

67. What is the probability a student studies less than 15 hours per week?

68. Are the events *study at least 15 hours per week* and *makes the honor roll* independent? Justify your answer numerically.

3.5: Tree and Venn Diagrams

69. At a high school, some students play on the tennis team and some play on the soccer team, but neither plays both tennis and soccer. Draw a Venn diagram illustrating this.

70. At a high school, some students play tennis, some play soccer, and some play both. Draw a Venn diagram illustrating this.

Practice Test 1 Solutions

1.1: Definitions of Statistics, Probability, and Key Terms

1.

A. population: all the shopping visits by all the store's customers
B. sample: the 1,000 visits drawn for the study
C. parameter: the average expenditure on produce per visit by all the store's customers
D. statistic: the average expenditure on produce per visit by the sample of 1,000
E. variable: the expenditure on produce for each visit
F. data: the dollar amounts spent on produce; for instance, $15.40, $11.53, etc.

2. C

3. D

1.2: Data, Sampling, and Variation in Data and Sampling

4. D

5. C

6. Answers will vary.
Sample Answer: Any solution in which you use data from the entire population is acceptable. For instance, a professor

might calculate the average exam score for her class: Because the scores of all members of the class were used in the calculation, the average is a parameter.

7. B

8. A

9.

# of years	Frequency	Relative Frequency	Cumulative Relative Frequency
< 5	25	0.25	0.25
5–10	30	0.30	0.55
> 10	45	0.45	1

Table B3

10. 0.75

11. 0.55

12. Answers will vary.
Sample Answer: One possibility is to obtain the class roster and assign each student a number from 1 to 200. Then, use a random number generator or table of random number to generate 30 numbers between 1 and 200, and select the students matching the random numbers. It would also be acceptable to write each student's name on a card, shuffle them in a box, and draw 30 names at random.

13. One possibility would be to obtain a roster of students enrolled in the college, including the class standing for each student. Then, you would draw a proportionate random sample from within each class. For instance, if 30 percent of the students in the college are freshman, then 30 percent of your sample would be drawn from the freshman class.

14. For the first person picked, the chance of any individual being selected is one in 150. For the second person, it is one in 149, for the third it is one in 148, and so on. For the 30th person selected, the chance of selection is one in 121.

15. A

16. No. There are at least two chances for bias. First, the viewers of this particular program may not be representative of American football fans as a whole. Second, the sample will be self-selected, because people have to make a phone call in order to take part, and those people are probably not representative of the American football fan population as a whole.

17. These results (84 percent in one sample, 86 percent in the other) are probably due to sampling variability. Each researcher drew a different sample of children, and you would not expect them to get exactly the same result, although you would expect the results to be similar, as they are in this case.

18. No. The improvement could also be due to self-selection: Only motivated students were willing to sign the contract, and they would have done well even in a school with 6.5 hour days. Because both changes were implemented at the same time, it is not possible to separate out their influence.

19. At least two aspects of this poll are troublesome. The first is that it was conducted by a group who would benefit by the result—almond sales are likely to increase if people believe that eating almonds will make them happier. The second is that this poll found that almond consumption and life satisfaction are correlated, but it does not establish that eating almonds causes satisfaction. It is equally possible, for instance, that people with higher incomes are more likely to eat almonds and are also more satisfied with their lives.

20. You want the sample of people who take part in a survey to be representative of the population from which they are drawn. People who refuse to take part in a survey often have different views than those who do participate, and so even a random sample may produce biased results if a large percentage of those selected refuse to participate in a survey.

1.3: Frequency, Frequency Tables, and Levels of Measurement

21. 13.2

1.4: Experimental Design and Ethics

22.

A. population: all college students
B. sample: the 100 college students in the study
C. experimental units: each individual college student who participated
D. explanatory variable: the size of the tableware
E. treatment: tableware that is 20 percent smaller than normal
F. response variable: the amount of food eaten

23. There are many lurking variables that could influence the observed differences in test scores. Perhaps the boys, on average, have taken more math courses than the girls, and the girls have taken more English classes than the boys. Perhaps the boys have been encouraged by their families and teachers to prepare for a career in math and science, and thus have put more effort into studying math, while the girls have been encouraged to prepare for fields like communication and psychology that are more focused on language use. A study design would have to control for these and other potential lurking variables (anything that could explain the observed difference in test scores, other than the genetic explanation) in order to draw a scientifically sound conclusion about genetic differences.

24. To use random assignment, you would have to be able to assign people to either exercise or not exercise. Because exercise has many beneficial effects, this would not be an ethical experiment. We will study people who chose to exercise and compare them to people who chose not to exercise, and try to control for the other ways those two groups may differ (lurking variables).

25. Sources of bias include the fact that not everyone has a telephone, that cell phone numbers are often not listed in published directories, and that an individual might not be at home at the time of the phone call; all these factors make it likely that the respondents to the survey will not be representative of the population as a whole.

26. Research subjects should not be coerced into participation, and offering extra credit in exchange for participation could be construed as coercion. In addition, this method will result in a volunteer sample, which cannot be assumed to be representative of the population as a whole.

2.1: Stem-and Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

27. The value 740 is an outlier, because the exams were graded on a scale of zero to 100, and 740 is far outside that range. It may be a data entry error, with the actual score being 74, so the professor should check that exam again to see what the actual score was.

28.

Stem	Leaf
6	2 4 5 5 8
7	0 2 2 4 5 5 5 6 8 8
8	1 3 3 4 5 7 8
9	2 5 8 8
10	0 0

Table B4

29. Most scores on this exam were in the range of 70–89, with a few scoring in the 60–69 range, and a few in the 90–100 range.

2.2: Histograms, Frequency Polygons, and Time Series Graphs

30. $RF = \frac{7}{35} = 0.2$

31. The range will be 0.5–1.5, and the central point will be 1.

32. Range 1.5–2.5, central point 2; range 2.5–3.5, central point 3; range 3.5–4.5, central point 4; range 4.5–5.5, central

point 5.

33. The bar from 3.5 to 4.5, with a central point of 4, will be tallest; its height will be nine, because there are nine students taking four courses.

34. The histogram is a better choice, because income is a continuous variable.

35. A bar graph is the better choice, because this data is categorical rather than continuous.

2.3: Measures of the Location of the Data

36. Your daughter scored better than 80 percent of the students in her grade on math and better than 76 percent of the students in reading. Both scores are very good, and place her in the upper quartile, but her math score is slightly better in relation to her peers than her reading score.

37. You had an unusually long wait time, which is bad: 82 percent of patients had a shorter wait time than you, and only 18 percent had a longer wait time.

2.4: Box Plots

38. 5

39. 3

40. 7

41. The median is 86, as represented by the vertical line in the box.

42. The first quartile is 80, and the third quartile is 92, as represented by the left and right boundaries of the box.

43. $IQR = 92 - 80 = 12$

44. Range = 100 – 75 = 25

2.5: Measures of the Center of the Data

45. Half the runners who finished the marathon ran a time faster than 3:35:04, and half ran a time slower than 3:35:04. Your time is faster than the median time, so you did better than more than half of the runners in this race.

46. 61.5, or $61,500

47. 49.25, or $49,250

48. The median, because the mean is distorted by the high value of one house.

2.6: Skewness and the Mean, Median, and Mode

49. C

50. A

51. They will all be fairly close to one another.

2.7: Measures of the Spread of the Data

52. Mean: 15
Standard deviation: 4.3

$$\mu = \frac{10+11+15+15+17+22}{6} = 15$$

$$s = \sqrt{\frac{\sum (x-\overline{x})^2}{n-1}} = \sqrt{\frac{94}{5}} = 4.3$$

53. 15 + (2)(4.3) = 23.6

54. 13.7 is one standard deviation below the mean of this data, because 15 – 4.3 = 10.7

55. $z = \frac{95-85}{5} = 2.0$
Susan's z score was 2.0, meaning she scored two standard deviations above the class mean for the final exam.

3.1: Terminology

56. $P(B) = \frac{25}{90} = 0.28$

57. Drawing a red marble is more likely.
$P(R) = \frac{50}{80} = 0.62$
$P(Y) = \frac{15}{80} = 0.19$

58. $P(F \text{ AND } S)$

59. $P(E|M)$

3.2: Independent and Mutually Exclusive Events

60. $P(A \text{ AND } B) = (0.3)(0.5) = 0.15$

61. $P(C \text{ OR } D) = 0.18 + 0.03 = 0.21$

3.3: Two Basic Rules of Probability

62. No, they cannot be mutually exclusive, because they add up to more than 300. Therefore, some students must fit into two or more categories (e.g., both going to college and working full time).

63. $P(A \text{ and } B) = (P(B|A))(P(A)) = (0.85)(0.70) = 0.595$

64. No. If they were independent, $P(B)$ would be the same as $P(B|A)$. We know this is not the case, because $P(B) = 0.70$ and $P(B|A) = 0.85$.

3.4: Contingency Tables

65.

	Honor roll	No honor roll	Total
Study at least 15 hours/week	482	200	682
Study less than 15 hours/week	125	193	318
Total	607	393	1,000

Table B5

66. $P(\text{honor roll}|\text{study at least 15 hours word per week}) = \frac{482}{1{,}000} = 0.482$

67. $P(\text{study less than 15 hours word per week}) = \frac{125+193}{1{,}000} = 0.318$

68. Let $P(S)$ = study at least 15 hours per week
Let $P(H)$ = make the honor roll
From the table, $P(S) = 0.682$, $P(H) = 0.607$, and $P(S \text{ AND } H) = 0.482$.
If $P(S)$ and $P(H)$ were independent, then $P(S \text{ AND } H)$ would equal $(P(S))(P(H))$.
However, $(P(S))(P(H)) = (0.682)(0.607) = 0.414$, while $P(S \text{ AND } H) = 0.482$.
Therefore, $P(S)$ and $P(H)$ are not independent.

3.5: Tree and Venn Diagrams

69.

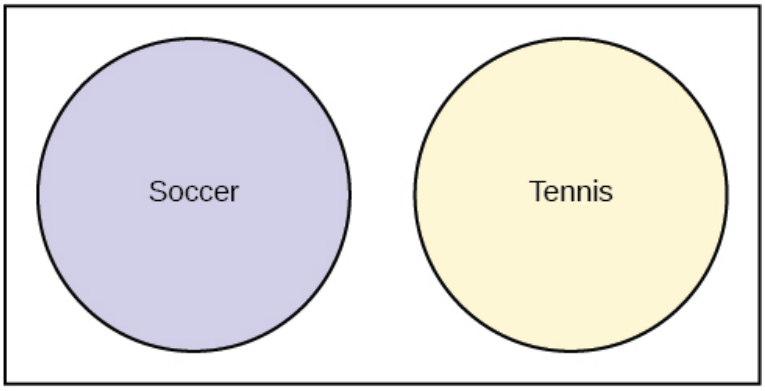


Figure B2

70.

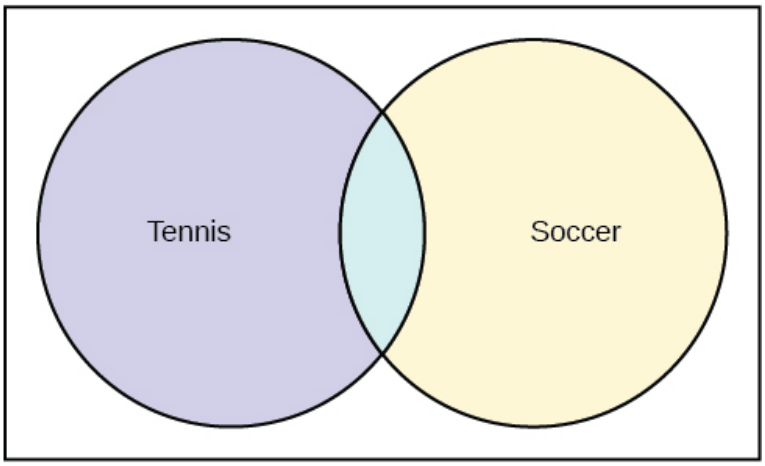


Figure B3

Practice Test 2

4.1: Probability Distribution Function (PDF) for a Discrete Random Variable

Use the following information to answer the next five exercises. You conduct a survey among a random sample of students at a particular university. The data collected includes their major, the number of classes they took the previous semester, and the amount of money they spent on books purchased for classes in the previous semester.

1. If X = student's major, then what is the domain of X?

2. If Y = the number of classes taken in the previous semester, what is the domain of Y?

3. If Z = the amount of money spent on books in the previous semester, what is the domain of Z?

4. Why are X, Y, and Z in the previous example random variables?

5. After collecting data, you find that, for one case, $z = -7$. Is this a possible value for Z?

6. What are the two essential characteristics of a discrete probability distribution?

Use this discrete probability distribution represented in this table to answer the following six questions. The university library records the number of books checked out by each patron over the course of one day, with the following result:

x	$P(x)$
0	0.20
1	0.45

Table B6

x	$P(x)$
2	0.20
3	0.10
4	0.05

Table B6

7. Define the random variable X for this example.

8. What is $P(x > 2)$?

9. What is the probability a patron will check out at least one book?

10. What is the probability a patron will take out no more than three books?

11. If the table listed $P(x)$ as 0.15, how would you know that there was a mistake?

12. What is the average number of books taken out by a patron?

4.2: Mean or Expected Value and Standard Deviation

Use the following information to answer the next four exercises. Three jobs are open in a company: one in the accounting department, one in the human resources department, and one in the sales department. The accounting job receives 30 applicants, and the human resources and sales department 60 applicants.

13. If X = the number of applications for a job, use this information to fill in Table B7.

x	$P(x)$	$xP(x)$

Table B7

14. What is the mean number of applicants?

15. What is the PDF for X?

16. Add a fourth column to the table, for $(x - \mu)^2 P(x)$.

17. What is the standard deviation of X?

4.3: Binomial Distribution

18. In a binomial experiment, if $p = 0.65$, what does q equal?

19. What are the required characteristics of a binomial experiment?

20. Joe conducts an experiment to see how many times he has to flip a coin before he gets four heads in a row. Does this qualify as a binomial experiment?

Use the following information to answer the next three exercises. In a particular community, 65 percent of households include at least one person who has graduated from college. You randomly sample 100 households in this community. Let X = the number of households including at least one college graduate.

21. Describe the probability distribution of X.

22. What is the mean of X?

23. What is the standard deviation of X?

Use the following information to answer the next four exercises. Joe is the star of his school's baseball team. His batting average is 0.400, meaning that for every 10 times he comes to bat (an at-bat), four of those times he gets a hit. You decide to track his batting performance for his next 20 at-bats.

24. Define the random variable X in this experiment.

25. Assuming Joe's probability of getting a hit is independent and identical across all 20 at-bats, describe the distribution of X.

26. Given this information, what number of hits do you predict Joe will get?

27. What is the standard deviation of X?

4.4: Geometric Distribution

28. What are the three major characteristics of a geometric experiment?

29. You decide to conduct a geometric experiment by flipping a coin until it comes up heads. This takes five trials. Represent the outcomes of this trial, using H for heads and T for tails.

30. You are conducting a geometric experiment by drawing cards from a normal 52-card pack, with replacement, until you draw the Queen of Hearts. What is the domain of X for this experiment?

31. You are conducting a geometric experiment by drawing cards from a normal 52-card deck, without replacement, until you draw a red card. What is the domain of X for this experiment?

Use the following information to answer the next three exercises. In a particular university, 27 percent of students are engineering majors. You decide to select students at random until you choose one that is an engineering major. Let X = the number of students you select until you find one that is an engineering major.

32. What is the probability distribution of X?

33. What is the mean of X?

34. What is the standard deviation of X?

4.5: Hypergeometric Distribution

35. You draw a random sample of 10 students to participate in a survey, from a group of 30, consisting of 16 boys and 14 girls. You are interested in the probability that seven of the students chosen will be boys. Does this qualify as a hypergeometric experiment? List the conditions and whether or not they are met.

36. You draw five cards, without replacement, from a normal 52-card deck of playing cards, and are interested in the probability that two of the cards are spades. What are the group of interest, size of the group of interest, and sample size for this example?

4.6: Poisson Distribution

37. What are the key characteristics of the Poisson distribution?

Use the following information to answer the next three exercises. The number of drivers to arrive at a toll booth in an hour can be modeled by the Poisson distribution.

38. If X = the number of drivers, and the average numbers of drivers per hour is four, how would you express this distribution?

39. What is the domain of X?

40. What are the mean and standard deviation of X?

5.1: Continuous Probability Functions

41. You conduct a survey of students to see how many books they purchased the previous semester, the total amount they paid for those books, the number they sold after the semester was over, and the amount of money they received for the books they sold. Which variables in this survey are discrete, and which are continuous?

42. With continuous random variables, we never calculate the probability that X has a particular value, but we always speak in terms of the probability that X has a value within a particular range. Why is this?

43. For a continuous random variable, why are $P(x < c)$ and $P(x \leq c)$ equivalent statements?

44. For a continuous probability function, $P(x < 5) = 0.35$. What is $P(x > 5)$, and how do you know?

45. Describe how you would draw the continuous probability distribution described by the function $f(x) = \frac{1}{10}$ for $0 \leq x \leq 10$. What type of a distribution is this?

46. For the continuous probability distribution described by the function $f(x) = \frac{1}{10}$ for $0 \leq x \leq 10$. what is the $P(0 < x < 4)$?

5.2: The Uniform Distribution

47. For the continuous probability distribution described by the function $f(x) = \frac{1}{10}$ for $0 \leq x \leq 10$, what is the $P(2 < x < 5)$?

Use the following information to answer the next four exercises. The number of minutes that a patient waits at a medical clinic to see a doctor is represented by a uniform distribution between zero and 30 minutes, inclusive.

48. If X equals the number of minutes a person waits, what is the distribution of X?

49. Write the probability density function for this distribution.

50. What is the mean and standard deviation for waiting time?

51. What is the probability that a patient waits less than 10 minutes?

5.3: The Exponential Distribution

52. The distribution of the variable X, representing the average time to failure for an automobile battery, can be written as $X \sim Exp(m)$. Describe this distribution in words.

53. If the value of m for an exponential distribution is 10, what are the mean and standard deviation for the distribution?

54. Write the probability density function for a variable distributed as $X \sim Exp(0.2)$.

6.1: The Standard Normal Distribution

55. Translate this statement about the distribution of a random variable X into words: $X \sim (100, 15)$.

56. If the variable X has the standard normal distribution, express this symbolically.

Use the following information for the next six exercises. According to the World Health Organization, distribution of height in centimeters for girls aged five years and zero months has the distribution $X \sim N(109, 4.5)$.

57. What is the z score for a height of 112 inches?

58. What is the z score for a height of 100 centimeters?

59. Find the z score for a height of 105 centimeters and explain what that means in the context of the population.

60. What height corresponds to a z score of 1.5 in this population?

61. Using the empirical rule, we expect about 68 percent of the values in a normal distribution to lie within one standard deviation above or below the mean. What does this mean, in terms of a specific range of values, for this distribution?

62. Using the empirical rule, about what percentage of heights in this distribution do you expect to be between 95.5 cm and 122.5 cm?

6.2: Using the Normal Distribution

Use the following information to answer the next four exercises. The distributor of raffle tickets claims that 20 percent of the tickets are winners. You draw a sample of 500 tickets to test this proposition.

63. Can you use the normal approximation to the binomial for your calculations? Why or why not.

64. What are the expected mean and standard deviation for your sample, assuming the distributor's claim is true?

65. What is the probability that your sample will have a mean greater than 100?

66. If the z score for your sample result is –2, explain what this means, using the empirical rule.

7.1: The Central Limit Theorem for Sample Means (Averages)

67. What does the central limit theorem state with regard to the distribution of sample means?

68. The distribution of results from flipping a fair coin is uniform: Heads and tails are equally likely on any flip, and over a large number of trials, you expect about the same number of heads and tails. Yet if you conduct a study by flipping 30 coins and recording the number of heads, and repeat this 100 times, the distribution of the mean number of heads will be approximately normal. How is this possible?

69. The mean of a normally-distributed population is 50, and the standard deviation is four. If you draw 100 samples of size 40 from this population, describe what you would expect to see in terms of the sampling distribution of the sample mean.

70. X is a random variable with a mean of 25 and a standard deviation of two. Write the distribution for the sample mean of samples of size 100 drawn from this population.

71. Your friend is doing an experiment drawing samples of size 50 from a population with a mean of 117 and a standard deviation of 16. This sample size is large enough to allow use of the central limit theorem, so he says the standard deviation of the sampling distribution of sample means will also be 16. Explain why this is wrong, and calculate the correct value.

72. You are reading a research article that refers to *the standard error of the mean*. What does this mean, and how is it calculated?

Use the following information to answer the next six exercises. You repeatedly draw samples of $n = 100$ from a population with a mean of 75 and a standard deviation of 4.5.

73. What is the expected distribution of the sample means?

74. One of your friends tries to convince you that the standard error of the mean should be 4.5. Explain what error your friend made.

75. What is the z score for a sample mean of 76?

76. What is the z score for a sample mean of 74.7?

77. What sample mean corresponds to a z score of 1.5?

78. If you decrease the sample size to 50, will the standard error of the mean be smaller or larger? What would be its value?

Use the following information to answer the next two questions. We use the empirical rule to analyze data for samples of size 60 drawn from a population with a mean of 70 and a standard deviation of 9.

79. What range of values would you expect to include 68 percent of the sample means?

80. If you increased the sample size to 100, what range would you expect to contain 68 percent of the sample means, applying the empirical rule?

7.2: The Central Limit Theorem for Sums

81. How does the central limit theorem apply to sums of random variables?

82. Explain how the rules applying the central limit theorem to sample means, and to sums of a random variable, are similar.

83. If you repeatedly draw samples of size 50 from a population with a mean of 80 and a standard deviation of four, and calculate the sum of each sample, what is the expected distribution of these sums?

Use the following information to answer the next four exercises. You draw one sample of size 40 from a population with a mean of 125 and a standard deviation of seven.

84. Compute the sum. What is the probability that the sum for your sample will be less than 5,000?

85. If you drew samples of this size repeatedly, computing the sum each time, what range of values would you expect to contain 95 percent of the sample sums?

86. What value is one standard deviation below the mean?

87. What value corresponds to a z score of 2.2?

7.3: Using the Central Limit Theorem

88. What does the law of large numbers say about the relationship between the sample mean and the population mean?

89. Applying the law of large numbers, which sample mean would you expect to be closer to the population mean: a sample of size 10 or a sample of size 100?

Use this information for the next three questions. A manufacturer makes screws with a mean diameter of 0.15 cm (centimeters) and a range of 0.10 cm to 0.20 cm; within that range, the distribution is uniform.

90. If X = the diameter of one screw, what is the distribution of X?

91. Suppose you repeatedly draw samples of size 100 and calculate their mean. Applying the central limit theorem, what is the distribution of these sample means?

92. Suppose you repeatedly draw samples of 60 and calculate their sum. Applying the central limit theorem, what is the distribution of these sample sums?

Practice Test 2 Solutions

Probability Distribution Function (PDF) for a Discrete Random Variable

1. The domain of X = {English, Mathematics, . . .}, i.e., a list of all the majors offered at the university, plus *undeclared*.

2. The domain of Y = {0, 1, 2, . . .}; i.e., the integers from zero to the upper limit of classes allowed by the university.

3. The domain of Z = any amount of money from zero upwards.

4. Because they can take any value within their domain, and their value for any particular case is not known until the survey is completed.

5. No, because the domain of Z includes only positive numbers (you cannot spend a negative amount of money). Possibly the value –7 is a data entry error, or a special code to indicate that the student did not answer the question.

6. The probabilities must sum to 1.0, and the probabilities of each event must be between 0 and 1, inclusive.

7. Let X = the number of books checked out by a patron.

8. $P(x > 2) = 0.10 + 0.05 = 0.15$

9. $P(x \geq 0) = 1 - 0.20 = 0.80$

10. $P(x \leq 3) = 1 - 0.05 = 0.95$

11. The probabilities would sum to 1.10, and the total probability in a distribution must always equal 1.0.

12. $\overline{x} = 0(0.20) + 1(0.45) + 2(0.20) + 3(0.10) + 4(0.05) = 1.35$

Mean or Expected Value and Standard Deviation

13.

x	$P(x)$	$xP(x)$
30	0.33	9.90
40	0.33	13.20
60	0.33	19.80

Table B8

14. $\overline{x} = 9.90 + 13.20 + 19.80 = 42.90$

15. $P(x = 30) = 0.33$
$P(x = 40) = 0.33$
$P(x = 60) = 0.33$

16.

x	$P(x)$	$xP(x)$	$(x-\mu)^2P(x)$
30	0.33	9.90	$(30-42.90)^2(0.33) = 54.91$
40	0.33	13.20	$(40-42.90)^2(0.33) = 2.78$
60	0.33	19.90	$(60-42.90)^2(0.33) = 96.49$

Table B9

17. $\sigma_x = \sqrt{54.91 + 2.78 + 96.49} = 12.42$

Binomial Distribution

18. $q = 1 - 0.65 = 0.35$

19.

1. There are a fixed number of trials.
2. There are only two possible outcomes, and they add up to one.
3. The trials are independent and conducted under identical conditions.

20. No, because there are not a fixed number of trials

21. $X \sim B(100, 0.65)$

22. $\mu = np = 100(0.65) = 65$

23. $\sigma_x = \sqrt{npq} = \sqrt{100(0.65)(0.35)} = 4.77$

24. X = Joe gets a hit in one at-bat (in one occasion of his coming to bat)

25. $X \sim B(20, 0.4)$

26. $\mu = np = 20(0.4) = 8$

27. $\sigma_x = \sqrt{npq} = \sqrt{20(0.40)(0.60)} = 2.19$

4.4: Geometric Distribution

28.

1. A series of Bernoulli trials are conducted until one is a success, and then the experiment stops.
2. At least one trial is conducted, but there is no upper limit to the number of trials.
3. The probability of success or failure is the same for each trial.

29. *T T T T H*

30. The domain of $X = \{1, 2, 3, 4, 5, \ldots n\}$. Because you are drawing with replacement, there is no upper bound to the number of draws that may be necessary.

31. The domain of $X = \{1, 2, 3, 4, 5, 6, 7, 8., 9, 10, 11, 12, \ldots 27\}$. Because you are drawing without replacement, and 26 of the 52 cards are red, you have to draw a red card within the first 17 draws.

32. $X \sim G(0.24)$

33. $\mu = \frac{1}{p} = \frac{1}{0.27} = 3.70$

34. $\sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-0.27}{0.27^2}} = 3.16$

4.5: Hypergeometric Distribution

35. Yes, because you are sampling from a population composed of two groups (boys and girls), have a group of interest (boys), and are sampling without replacement (hence, the probabilities change with each pick, and you are not performing Bernoulli trials).

36. The group of interest is the cards that are spades, the size of the group of interest is 13, and the sample size is five.

4.6: Poisson Distribution

37. A Poisson distribution models the number of events occurring in a fixed interval of time or space, when the events are independent and the average rate of the events is known.

38. $X \sim P(4)$

39. The domain of $X = \{0, 1, 2, 3, \ldots\}$; i.e., any integer from 0 upwards.

40. $\mu = 4$
$\sigma = \sqrt{4} = 2$

5.1: Continuous Probability Functions

41. The discrete variables are the number of books purchased, and the number of books sold after the end of the semester. The continuous variables are the amount of money spent for the books, and the amount of money received when they were sold.

42. Because for a continuous random variable, $P(x = c) = 0$, where c is any single value. Instead, we calculate $P(c < x < d)$; i.e., the probability that the value of x is between the values c and d.

43. Because $P(x = c) = 0$ for any continuous random variable.

44. $P(x > 5) = 1 - 0.35 = 0.65$, because the total probability of a continuous probability function is always 1.

45. This is a uniform probability distribution. You would draw it as a rectangle with the vertical sides at 0 and 20, and the horizontal sides at $\frac{1}{10}$ and 0.

46. $P(0 < x < 4) = (4 - 0)\left(\frac{1}{10}\right) = 0.4$

5.2: The Uniform Distribution

47. $P(2 < x < 5) = (5 - 2)\left(\frac{1}{10}\right) = 0.3$

48. $X \sim U(0, 15)$

49. $f(x) = \frac{1}{b-a}$ for $(a \leq x \leq b)$ so $f(x) = \frac{1}{30}$ for $(0 \leq x \leq 30)$

50. $\mu = \frac{a+b}{2} = \frac{0+30}{5} = 15.0$

$\sigma = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(30-0)^2}{12}} = 8.66$

51. $P(x < 10) = (10)\left(\frac{1}{30}\right) = 0.33$

5.3: The Exponential Distribution

52. X has an exponential distribution with decay parameter m and mean and standard deviation $\frac{1}{m}$. In this distribution, there will be relatively large numbers of small values, with values becoming less common as they become larger.

53. $\mu = \sigma = \frac{1}{m} = \frac{1}{10} = 0.1$

54. $f(x) = 0.2e^{-0.2x}$ where $x \geq 0$.

6.1: The Standard Normal Distribution

55. The random variable X has a normal distribution with a mean of 100 and a standard deviation of 15.

56. $X \sim N(0,1)$

57. $z = \frac{x-\mu}{\sigma}$ so $z = \frac{112-109}{4.5} = 0.67$

58. $z = \frac{x-\mu}{\sigma}$ so $z = \frac{100-109}{4.5} = -2.00$

59. $z = \frac{105-109}{4.5} = -0.89$
This girl is shorter than average for her age, by 0.89 standard deviations.

60. 109 + (1.5)(4.5) = 115.75 cm

61. We expect about 68 percent of the heights of girls aged five years and zero months to be between 104.5 cm and

113.5 cm.

62. We expect 99.7 percent of the heights in this distribution to be between 95.5 cm and 122.5 cm, because that range represents the values three standard deviations above and below the mean.

6.2: Using the Normal Distribution

63. Yes, because both np and nq are greater than five.
np = (500)(0.20) = 100 and *nq* = 500(0.80) = 400

64. $\mu = np = (500)(0.20) = 100$

$\sigma = \sqrt{npq} = \sqrt{500(0.20)(0.80)} = 8.94$

65. Fifty percent, because in a normal distribution, half the values lie above the mean.

66. The results of our sample were two standard deviations below the mean, suggesting it is unlikely that 20 percent of the raffle tickets are winners, as claimed by the distributor, and that the true percentage of winners is lower. Applying the Empirical Rule, if that claim were true, we would expect to see a result this far below the mean only about 2.5 percent of the time.

7.1: The Central Limit Theorem for Sample Means (Averages)

67. The central limit theorem states that if samples of sufficient size are drawn from a population, the distribution of sample means will be normal, even if the distribution of the population is not normal.

68. The sample size of 30 is sufficiently large in this example to apply the central limit theorem. This theorem states that, for samples of sufficient size drawn from a population, the sampling distribution of the sample mean will approach normality, regardless of the distribution of the population from which the samples were drawn.

69. You would not expect each sample to have a mean of 50, because of sampling variability. However, you would expect the sampling distribution of the sample means to cluster around 50, with an approximately normal distribution, so that values close to 50 are more common than values further removed from 50.

70. $\overline{X} \sim N(25, 0.2)$ because $\overline{X} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$

71. The standard deviation of the sampling distribution of the sample means can be calculated using the formula $\left(\frac{\sigma_x}{\sqrt{n}}\right)$, which in this case is $\left(\frac{16}{\sqrt{50}}\right)$. The correct value for the standard deviation of the sampling distribution of the sample means is therefore 2.26.

72. The standard error of the mean is another name for the standard deviation of the sampling distribution of the sample mean. Given samples of size n drawn from a population with standard deviation σ_x, the standard error of the mean is $\left(\frac{\sigma_x}{\sqrt{n}}\right)$.

73. $X \sim N(75, 0.45)$

74. Your friend forgot to divide the standard deviation by the square root of *n*.

75. $z = \frac{\overline{x} - \mu_x}{\sigma_x} = \frac{76-75}{4.5} = 2.2$

76. $z = \frac{\overline{x} - \mu_x}{\sigma_x} = \frac{74.7-75}{4.5} = -0.67$

77. 75 + (1.5)(0.45) = 75.675

78. The standard error of the mean will be larger, because you will be dividing by a smaller number. The standard error of the mean for samples of size *n* = 50 is
$\left(\frac{\sigma_x}{\sqrt{n}}\right) = \frac{4.5}{\sqrt{50}} = 0.64$

79. You would expect this range to include values up to one standard deviation above or below the mean of the sample means. In this case:
$70 + \frac{9}{\sqrt{60}} = 71.16$ and $70 - \frac{9}{\sqrt{60}} = 68.84$ so you would expect 68 percent of the sample means to be between 68.84 and 71.16.

80. $70 + \frac{9}{\sqrt{100}} = 70.9$ and $70 - \frac{9}{\sqrt{100}} = 69.1$ so you would expect 68 percent of the sample means to be between 69.1 and 70.9. Note that this is a narrower interval due to the increased sample size.

7.2: The Central Limit Theorem for Sums

81. For a random variable *X*, the random variable ΣX will tend to become normally distributed as the size n of the samples used to compute the sum increases.

82. Both rules state that the distribution of a quantity (the mean or the sum) calculated on samples drawn from a population will tend to have a normal distribution as the sample size increases, regardless of the distribution of population from which the samples are drawn.

83. $\Sigma X \sim N\left(n\mu_x, \left(\sqrt{n}\right)(\sigma_x)\right)$ so $\Sigma X \sim N(4{,}000, 28.3)$

84. The probability is 0.50, because 5,000 is the mean of the sampling distribution of sums of size 40 from this population. Sums of random variables computed from a sample of sufficient size are normally distributed, and in a normal distribution, half the values lie below the mean.

85. Using the empirical rule, you would expect 95 percent of the values to be within two standard deviations of the mean. Using the formula for the standard deviation is for a sample sum $\left(\sqrt{n}\right)(\sigma_x) = \left(\sqrt{40}\right)(7) = 44.3$, so you would expect 95 percent of the values to be between 5,000 + (2)(44.3) and 5,000 – (2)(44.3), or between 4,911.4 and 588.6.

86. $\mu - \left(\sqrt{n}\right)(\sigma_x) = 5{,}000 - \left(\sqrt{40}\right)(7) = 4{,}955.7$

87. $5{,}000 + (2.2)\left(\sqrt{40}\right)(7) = 5097.4$

7.3: Using the Central Limit Theorem

88. The law of large numbers says that, as sample size increases, the sample mean tends to get nearer and nearer to the population mean.

89. You would expect the mean from a sample of size 100 to be nearer to the population mean, because the law of large numbers says that, as sample size increases, the sample mean tends to approach the population mean.

90. $X \sim N(0.10, 0.20)$

91. $\overline{X} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$ and the standard deviation of a uniform distribution is $\frac{b-a}{\sqrt{12}}$. In this example, the standard deviation of the distribution is $\frac{b-a}{\sqrt{12}} = \frac{0.10}{\sqrt{12}} = 0.03$
so $\overline{X} \sim N(0.15, 0.003)$

92. $\Sigma X \sim N((n)(\mu_x), (\sqrt{n})(\sigma_x))$ so $\Sigma X \sim N(9.0, 0.23)$

Practice Test 3

8.1: Confidence Interval, Single Population Mean, Population Standard Deviation Known, Normal

Use the following information to answer the next seven exercises. You draw a sample of size 30 from a normally distributed population with a standard deviation of four.

1. What is the standard error of the sample mean in this scenario, rounded to two decimal places?

2. What is the distribution of the sample mean?

3. If you want to construct a two-sided 95 percent confidence interval, how much probability will be in each tail of the distribution?

4. What is the appropriate *z* score and error bound or margin of error (*EBM*) for a 95 percent confidence interval for this data?

5. Rounding to two decimal places, what is the 95 percent confidence interval if the sample mean is 41?

6. What is the 90 percent confidence interval if the sample mean is 41? Round to two decimal places

7. Suppose the sample size in this study had been 50, rather than 30. What would the 95 percent confidence interval be if

the sample mean is 41? Round your answer to two decimal places.

8. For any given data set and sampling situation, which would you expect to be wider: a 95 percent confidence interval or a 99 percent confidence interval?

8.2: Confidence Interval, Single Population Mean, Standard Deviation Unknown, Student's *t*

9. Comparing graphs of the standard normal distribution (*z* distribution) and a *t* distribution with 15 degrees of freedom (*df*), how do they differ?

10. Comparing graphs of the standard normal distribution (*z* distribution) and a *t* distribution with 15 degrees of freedom (*df*), how are they similar?

Use the following information to answer the next five exercises. Body temperature is known to be distributed normally among healthy adults. Because you do not know the population standard deviation, you use the *t* distribution to study body temperature. You collect data from a random sample of 20 healthy adults and find that your sample temperatures have a mean of 98.4 and a sample standard deviation of 0.3 (both in degrees Fahrenheit).

11. What are the degrees of freedom (*df*) for this study?

12. For a two-tailed 95 percent confidence interval, what is the appropriate *t* value to use in the formula?

13. What is the 95 percent confidence interval?

14. What is the 99 percent confidence interval? Round to two decimal places.

15. Suppose your sample size had been 30 rather than 20. What would the 95 percent confidence interval be then? Round to two decimal places

8.3: Confidence Interval for a Population Proportion

Use this information to answer the next four exercises. You conduct a poll of 500 randomly selected city residents, asking them if they own an automobile. Of the respondents, 280 say they own an automobile, and 220 say they do not.

16. Find the sample proportion and sample standard deviation for this data.

17. What is the 95 percent two-sided confidence interval? Round to four decimal places.

18. Calculate the 90 percent confidence interval. Round to four decimal places.

19. Calculate the 99 percent confidence interval. Round to four decimal places.

Use the following information to answer the next three exercises. You are planning to conduct a poll of community members aged 65 and older, to determine how many own mobile phones. You want to produce an estimate whose 95 percent confidence interval will be within four percentage points (plus or minus) of the true population proportion. Use an estimated population proportion of 0.5.

20. What sample size do you need?

21. Suppose you knew from prior research that the population proportion was 0.6. What sample size would you need?

22. Suppose you wanted a 95 percent confidence interval within three percentage points of the population. Assume the population proportion is 0.5. What sample size do you need?

9.1: Null and Alternate Hypotheses

23. In your state, 58 percent of registered voters in a community are registered as republicans. You want to conduct a study to see if this also holds up in your community. State the null and alternative hypotheses to test this.

24. You believe that at least 58 percent of registered voters in a community are registered as republicans. State the null and alternative hypotheses to test this.

25. The mean household value in a city is $268,000. You believe that the mean household value in a particular neighborhood is lower than the city average. Write the null and alternative hypotheses to test this.

26. State the appropriate alternative hypothesis to this null hypothesis: H_0: $\mu = 107$

27. State the appropriate alternative hypothesis to this null hypothesis: H_0: $p < 0.25$

9.2: Outcomes and the Type I and Type II Errors

28. If you reject H_0 when H_0 is correct, what type of error is this?

29. If you fail to reject H_0 when H_0 is false, what type of error is this?

30. What is the relationship between the Type II error and the power of a test?

31. A new blood test is being developed to screen patients for cancer. Positive results are followed up by a more accurate (and expensive) test. It is assumed that the patient does not have cancer. Describe the null hypothesis and the Type I and Type II errors for this situation, and explain which type of error is more serious.

32. Explain in words what it means that a screening test for TB has an α level of 0.10. The null hypothesis is that the patient does not have TB.

33. Explain in words what it means that a screening test for TB has a β level of 0.20. The null hypothesis is that the patient does not have TB.

34. Explain in words what it means that a screening test for TB has a power of 0.80.

9.3: Distribution Needed for Hypothesis Testing

35. If you are conducting a hypothesis test of a single population mean, and you do not know the population variance, what test will you use if the sample size is 10 and the population is normal?

36. If you are conducting a hypothesis test of a single population mean, and you know the population variance, what test will you use?

37. If you are conducting a hypothesis test of a single population proportion, with np and nq greater than or equal to five, what test will you use, and with what parameters?

38. Published information indicates that, on average, college students spend less than 20 hours studying per week. You draw a sample of 25 students from your college and find the sample mean to be 18.5 hours, with a standard deviation of 1.5 hours. What distribution will you use to test whether study habits at your college are the same as the national average, and why?

39. A published study says that 95 percent of American children are vaccinated against a disease, with a standard deviation of 1.5 percent. You draw a sample of 100 children from your community and check their vaccination records to see if the vaccination rate in your community is the same as the national average. What distribution will you use for this test, and why?

9.4: Rare Events, the Sample, Decision, and Conclusion

40. You are conducting a study with an α level of 0.05. If you get a result with a p-value of 0.07, what will be your decision?

41. You are conducting a study with $\alpha = 0.01$. If you get a result with a p-value of 0.006, what will be your decision?

Use the following information to answer the next five exercises. According to the World Health Organization, the average height of a one-year-old child is 29". You believe children with a particular disease are smaller than average, so you draw a sample of 20 children with this disease and find a mean height of 27.5" and a sample standard deviation of 1.5".

42. What are the null and alternative hypotheses for this study?

43. What distribution will you use to test your hypothesis, and why?

44. What is the test statistic and the p-value?

45. Based on your sample results, what is your decision?

46. Suppose the mean for your sample was 25. Redo the calculations and describe what your decision would be.

9.5: Additional Information and Full Hypothesis Test Examples

47. You conduct a study using α = 0.05. What is the level of significance for this study?

48. You conduct a study, based on a sample drawn from a normally distributed population with a known variance, with the following hypotheses:
H_0: $\mu = 35.5$
H_a: $\mu \neq 35.5$

Will you conduct a one-tailed or two-tailed test?

49. You conduct a study, based on a sample drawn from a normally distributed population with a known variance, with the following hypotheses:
H_0: $\mu \geq 35.5$
H_a: $\mu < 35.5$
Will you conduct a one-tailed or two-tailed test?

Use the following information to answer the next three exercises. Nationally, 80 percent of adults own an automobile. You are interested in whether the same proportion in your community own cars. You draw a sample of 100 and find that 75 percent own cars.

50. What are the null and alternative hypotheses for this study?

51. What test will you use, and why?

10.1: Comparing Two Independent Population Means with Unknown Population Standard Deviations

52. You conduct a poll of political opinions, interviewing both members of 50 married couples. Are the groups in this study independent or matched?

53. You are testing a new drug to treat insomnia. You randomly assign 80 volunteer subjects to either the experimental (new drug) or control (standard treatment) conditions. Are the groups in this study independent or matched?

54. You are investigating the effectiveness of a new math textbook for high school students. You administer a pretest to a group of students at the beginning of the semester, and a posttest at the end of a year's instruction using this textbook, and compare the results. Are the groups in this study independent or matched?

Use the following information to answer the next two exercises. You are conducting a study of the difference in time at two colleges for undergraduate degree completion. At College A, students take an average of 4.8 years to complete an undergraduate degree, while at College B, they take an average of 4.2 years. The pooled standard deviation for this data is 1.6 years.

55. Calculate Cohen's *d* and interpret it.

56. Suppose the mean time to earn an undergraduate degree at College A was 5.2 years. Calculate the effect size and interpret it.

57. You conduct an independent-samples *t* test with sample size 10 in each of two groups. If you are conducting a two-tailed hypothesis test with $\alpha = 0.01$, what *p*-values will cause you to reject the null hypothesis?

58. You conduct an independent samples *t* test with sample size 15 in each group, with the following hypotheses:
H_0: $\mu \geq 110$
H_a: $\mu < 110$
If $\alpha = 0.05$, what *t* values will cause you to reject the null hypothesis?

10.2: Comparing Two Independent Population Means with Known Population Standard Deviations

Use the following information to answer the next six exercises. College students in the sciences often complain that they must spend more on textbooks each semester than students in the humanities. To test this, you draw random samples of 50 science and 50 humanities students from your college, and record how much each spent last semester on textbooks. Consider the science students to be group one, and the humanities students to be group two.

59. What is the random variable for this study?

60. What are the null and alternative hypotheses for this study?

61. If the 50 science students spent an average of $530 with a sample standard deviation of $20, and the 50 humanities students spent an average of $380 with a sample standard deviation of $15, would you not reject or reject the null hypothesis? Use an alpha level of 0.05. What is your conclusion?

62. What would be your decision, if you were using $\alpha = 0.01$?

10.3: Comparing Two Independent Population Proportions

Use the information to answer the next six exercises. You want to know if the proportion of homes with cable television

service differs between Community A and Community B. To test this, you draw a random sample of 100 for each and record whether they have cable service.

63. What are the null and alternative hypotheses for this study?

64. If 65 households in Community A have cable service, and 78 households in Community B, what is the pooled proportion?

65. At $\alpha = 0.03$, will you reject the null hypothesis? What is your conclusion? Sixty-five households in Community A have cable service, and 78 households in community B. One hundred households in each community were surveyed.

66. Using an alpha value of 0.01, would you reject the null hypothesis? What is your conclusion? Sixty-five households in Community A have cable service, and 78 households in Community B. One hundred households in each community were surveyed.

10.4: Matched or Paired Samples

Use the following information to answer the next five exercises. You are interested in whether a particular exercise program helps people run a mile faster. You conduct a study in which you weigh the participants at the start of the study, and again at the conclusion, after they have participated in the exercise program for six months. You compare the results using a matched-pairs *t* test, in which the data is {time to run a mile at conclusion, time at start}. You believe that, on average, the participants will be able to run a mile faster after six months on the exercise program.

67. What are the null and alternative hypotheses for this study?

68. Calculate the test statistic, assuming that $\overline{x}_d = -5$, $s_d = 6$, and $n = 30$ (pairs).

69. What are the degrees of freedom for this statistic?

70. Using $\alpha = 0.05$, what is your decision regarding the effectiveness of this program in improving running speed? What is the conclusion?

71. What would it mean if the *t* statistic had been 4.56, and what would have been your decision in that case?

11.1: Facts About the Chi-Square Distribution

72. What is the mean and standard deviation for a chi-square distribution with 20 degrees of freedom?

11.2: Goodness-of-Fit Test

Use the following information to answer the next four exercises. Nationally, about 66 percent of high school graduates enroll in higher education. You perform a chi-square goodness of fit test to see if this same proportion applies to your high school's most recent graduating class of 200. Your null hypothesis is that the national distribution also applies to your high school.

73. What are the expected numbers of students from your high school graduating class enrolled and not enrolled in higher education?

74. Fill out the rest of this table.

	Observed (*O*)	Expected (*E*)	*O* – *E*	(*O* – *E*)2	$\frac{(O-E)^2}{z}$
Enrolled	145				
Not enrolled	55				

Table B10

75. What are the degrees of freedom for this chi-square test?

76. What is the chi-square test statistic and the *p*-value? At the five percent significance level, what do you conclude?

77. For a chi-square distribution with 92 degrees of freedom, the curve _____.

78. For a chi-square distribution with five degrees of freedom, the curve is _____.

11.3: Test of Independence

Use the following information to answer the next four exercises. You are considering conducting a chi-square test of independence for the data in this table, which displays data about cell phone ownership for freshman and seniors at a high school. Your null hypothesis is that cell phone ownership is independent of class standing.

79. Compute the expected values for the cells.

	Cell = Yes	Cell = No
Freshman	100	150
Senior	200	50

Table B11

80. Compute $\frac{(O-E)^2}{z}$ for each cell, where O = observed and E = expected.

81. What is the chi-square statistic and degrees of freedom for this study?

82. At the $\alpha = 0.5$ significance level, what is your decision regarding the null hypothesis?

11.4: Test of Homogeneity

83. You conduct a chi-square test of homogeneity for data in a five-by-two table. What are the degrees of freedom for this test?

11.5: Comparison Summary of the Chi-Square Tests: Goodness-of-Fit, Independence and Homogeneity

84. A 2013 poll in the State of California surveyed people about a tax. The results are presented in the following table, and are classified by ethnic group and response type. Are the poll responses independent of the participants' ethnic group? Conduct a hypothesis test at the five percent significance level.

Ethnic Group/Response Type	Favor	Oppose	No Opinion	Row Total
White/Non-Hispanic	234	433	43	710
Latino	147	106	19	272
African American	24	41	6	71
Asian American	54	48	16	118
Column Total	459	628	84	1171

Table B12

85. In a test of homogeneity, what must be true about the expected value of each cell?

86. Stated in general terms, what are the null and alternative hypotheses for the chi-square test of independence?

87. Stated in general terms, what are the null and alternative hypotheses for the chi-square test of homogeneity?

11.6: Test of a Single Variance

88. A lab test claims to have a variance of no more than five. You believe the variance is greater. What are the null and alternative hypotheses to test this?

Practice Test 3 Solutions

8.1: Confidence Interval, Single Population Mean, Population Standard Deviation Known, Normal

1. $\frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{30}} = 0.73$

2. normal

3. 0.025 or 2.5 percent; A 95 percent confidence interval contains 95 percent of the probability, and excludes 5 percent, and the 5 percent excluded is split evenly between the upper and lower tails of the distribution.

4. *z* score = 1.96; $EBM = z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) = (1.96)(0.73) = 1.4308$

5. 41 ± 1.43 = (39.57, 42.43); using the calculator function ZInterval, answer is (40.74, 41.26). Answers differ due to rounding.

6. The *z*-value for a 90 percent confidence interval is 1.645, so *EBM* = 1.645(0.73) = 1.20085.
The 90 percent confidence interval is 41 ± 1.20 = (39.80, 42.20).
The calculator function ZInterval answer is (40.78, 41.23). Answers differ due to rounding.

7. The standard error of measurement is $\frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{50}} = 0.57$

$EBM = z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) = (1.96)(0.57) = 1.12$

The 95 percent confidence interval is 41 ± 1.12 = (39.88, 42.12).
The calculator function ZInterval answer is (40.84, 41.16). Answers differ due to rounding.

8. The 99 percent confidence interval, because it includes all but one percent of the distribution. The 95 percent confidence interval will be narrower, because it excludes five percent of the distribution.

8.2: Confidence Interval, Single Population Mean, Standard Deviation Unknown, Student's *t*

9. The *t* distribution will have more probability in its tails (*thicker tails*) and less probability near the mean of the distribution (*shorter in the center*).

10. Both distributions are symmetrical and centered at zero.

11. *df* = *n* – 1 = 20 – 1 = 19

12. You can get the *t* value from a probability table or a calculator. In this case, for a *t* distribution with 19 degrees of freedom and a 95 percent two-sided confidence interval, the value is 2.093; i.e., $t_{\frac{\alpha}{2}} = 2.093$. The calculator function is invT(0.975, 19).

13. $EBM = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) = (2.093)\left(\frac{0.3}{\sqrt{20}}\right) = 0.140$

98.4 ± 0.14 = (98.26, 98.54).
The calculator function TInterval answer is (98.26, 98.54).

14. $t_{\frac{\alpha}{2}} = 2.861$. The calculator function is invT(0.995, 19).

$EBM = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) = (2.861)\left(\frac{0.3}{\sqrt{20}}\right) = 0.192$

98.4 ± 0.19 = (98.21, 98.59). The calculator function TInterval answer is (98.21, 98.59).

15. *df* = *n* – 1 = 30 – 1 = 29. $t_{\frac{\alpha}{2}} = 2.045$

$EBM = z_t\left(\frac{s}{\sqrt{n}}\right) = (2.045)\left(\frac{0.3}{\sqrt{30}}\right) = 0.112$

98.4 ± 0.11 = (98.29, 98.51). The calculator function TInterval answer is (98.29, 98.51).

8.3: Confidence Interval for a Population Proportion

16. $p' = \frac{280}{500} = 0.56$

$q' = 1 - p' = 1 - 0.56 = 0.44$

$s = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.56(0.44)}{500}} = 0.0222$

17. Because you are using the normal approximation to the binomial, $z_{\frac{\alpha}{2}} = 1.96$.

Calculate the error bound for the population (*EBP*):

$EBP = z_{\frac{\alpha}{2}}\sqrt{\frac{pq}{n}} = 1.96\,(0.222) = 0.0435$

Calculate the 95 percent confidence interval:
0.56 ± 0.0435 = (0.5165, 0.6035).
The calculator function 1-PropZint answer is (0.5165, 0.6035).

18. $z_{\frac{\alpha}{2}} = 1.64$

$EBP = z_{\frac{\alpha}{2}}\sqrt{\frac{pq}{n}} = 1.64\,(0.0222) = 0.0364$

0.56 ± 0.03 = (0.5236, 0.5964). The calculator function 1-PropZint answer is (0.5235, 0.5965).

19. $z_{\frac{\alpha}{2}} = 2.58$

$EBP = z_{\frac{\alpha}{2}}\sqrt{\frac{pq}{n}} = 2.58\,(0.0222) = 0.0573$

0.56 ± 0.05 = (0.5127, 0.6173).
The calculator function 1-PropZint answer is (0.5028, 0.6172).

20. *EBP* = 0.04 (because 4 percent = 0.04)
$z_{\frac{\alpha}{2}} = 1.96$ for a 95 percent confidence interval.

$n = \frac{z^2 pq}{EBP^2} = \frac{1.96^2(0.5)(0.5)}{0.04^2} = \frac{0.9604}{0.0016} = 600.25$

You need 601 subjects (rounding upward from 600.25).

21. $n = \frac{n^2 pq}{EBP^2} = \frac{1.96^2(0.6)(0.4)}{0.04^2} = \frac{0.9220}{0.0016} = 576.24$

You need 577 subjects (rounding upward from 576.24).

22. $n = \frac{n^2 pq}{EBP^2} = \frac{1.96^2(0.5)(0.5)}{0.03^2} = \frac{0.9604}{0.0009} = 1067.11$

You need 1,068 subjects (rounding upward from 1,067.11).

9.1: Null and Alternate Hypotheses

23. H_0: $p = 0.58$
H_a: $p \neq 0.58$

24. H_0: $p \geq 0.58$
H_a: $p < 0.58$

25. H_0: $\mu \geq \$268{,}000$
H_a: $\mu < \$268{,}000$

26. H_a: $\mu \neq 107$

27. H_a: $p \geq 0.25$

9.2: Outcomes and the Type I and Type II Errors

28. a Type I error

29. a Type II error

30. Power = 1 – β = 1 – *P*(Type II error).

31. The null hypothesis is that the patient does not have cancer. A Type I error would be detecting cancer when it is not present. A Type II error would be not detecting cancer when it is present. A Type II error is more serious, because failure to detect cancer could keep a patient from receiving appropriate treatment.

32. The screening test has a 10 percent probability of a Type I error, meaning that 10 percent of the time, it will detect TB when it is not present.

33. The screening test has a 20 percent probability of a Type II error, meaning that 20 percent of the time, it will fail to detect TB when it is in fact present.

34. Eighty percent of the time, the screening test will detect TB when it is actually present.

9.3: Distribution Needed for Hypothesis Testing

35. The Student's *t* test.

36. The normal distribution or *z* test.

37. The normal distribution with $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$

38. t_{24}. You use the *t* distribution because you do not know the population standard deviation, and the degrees of freedom are 24 because $df = n - 1$.

39. $\overline{X} \sim N\left(0.95, \frac{0.051}{\sqrt{100}}\right)$
Because you know the population standard deviation and have a large sample, you can use the normal distribution.

9.4: Rare Events, the Sample, Decision, and Conclusion

40. Fail to reject the null hypothesis, because $\alpha \leq p$.

41. Reject the null hypothesis, because $\alpha \geq p$.

42. H_0: $\mu \geq 29.0$"
H_a: $\mu < 29.0$"

43. t_{19}. Because you do not know the population standard deviation, use the *t* distribution. The degrees of freedom are 19, because $df = n - 1$.

44. The test statistic is –4.4721 and the *p*-value is 0.00013 using the calculator function TTEST.

45. With $\alpha = 0.05$, reject the null hypothesis.

46. With $\alpha = 0.05$, the *p*-value is almost zero using the calculator function TTEST, so reject the null hypothesis.

9.5: Additional Information and Full Hypothesis Test Examples

47. The level of significance is five percent.

48. two-tailed

49. one-tailed

50. H_0: $p = 0.8$
H_a: $p \neq 0.8$

51. You will use the normal test for a single population proportion because *np* and *nq* are both greater than five.

10.1: Comparing Two Independent Population Means with Unknown Population Standard Deviations

52. They are matched (paired), because you interviewed married couples.

53. They are independent, because participants were assigned at random to the groups.

54. They are matched (paired), because you collected data twice from each individual.

55. $d = \frac{\overline{x}_1 - \overline{x}_2}{s_{pooled}} = \frac{4.8 - 4.2}{1.6} = 0.375$
This is a small effect size, because 0.375 falls between Cohen's small (0.2) and medium (0.5) effect sizes.

56. $d = \frac{\overline{x}_1 - \overline{x}_2}{s_{pooled}} = \frac{5.2 - 4.2}{1.6} = 0.625$
The effect size is 0.625. By Cohen's standard, this is a medium effect size, because it falls between the medium (0.5) and large (0.8) effect sizes.

57. *p*-value < 0.01.

58. You will only reject the null hypothesis if you get a value significantly below the hypothesized mean of 110.

10.2: Comparing Two Independent Population Means with Known Population Standard Deviations

59. $\overline{X}_1 - \overline{X}_2$; i.e., the mean difference in amount spent on textbooks for the two groups.

60. H_0: $\overline{X}_1 - \overline{X}_2 \leq 0$
H_a: $\overline{X}_1 - \overline{X}_2 > 0$
This could also be written as
H_0: $\overline{X}_1 \leq \overline{X}_2$
H_a: $\overline{X}_1 > \overline{X}_2$

61. Using the calculator function 2-SampTTest, reject the null hypothesis. At the five percent significance level, there is sufficient evidence to conclude that the science students spend more on textbooks than the humanities students.

62. Using the calculator function 2-SampTTest, reject the null hypothesis. At the one percent significance level, there is sufficient evidence to conclude that the science students spend more on textbooks than the humanities students.

10.3: Comparing Two Independent Population Proportions

63. H_0: $p_A = p_B$
H_a: $p_A \neq p_B$

64. $p_c = \frac{x_A + x_A}{n_A + n_A} = \frac{65+78}{100+100} = 0.715$

65. Using the calculator function 2-PropZTest, the *p*-value = 0.0417. Reject the null hypothesis. At the three percent significance level, here is sufficient evidence to conclude that there is a difference between the proportions of households in the two communities that have cable service.

66. Using the calculator function 2-PropZTest, the *p*-value = 0.0417. Do not reject the null hypothesis. At the one percent significance level, there is insufficient evidence to conclude that there is a difference between the proportions of households in the two communities that have cable service.

10.4: Matched or Paired Samples

67. H_0: $\overline{x}_d \geq 0$
H_a: $\overline{x}_d < 0$

68. t = –4.5644.

69. df = 30 – 1 = 29.

70. Using the calculator function TTEST, the *p*-value = 0.00004, so reject the null hypothesis. At the five percent level, there is sufficient evidence to conclude that the participants lost weight, on average.

71. A positive t statistic would mean that participants, on average, gained weight over the six months.

11.1: Facts About the Chi-Square Distribution

72. $\mu = df = 20$
$\sigma = \sqrt{2(df)} = \sqrt{40} = 6.32$

11.2: Goodness-of-Fit Test

73. Enrolled = 200(0.66) = 132. Not enrolled = 200(0.34) = 68.

74.

	Observed (O)	Expected (E)	O – E	(O – E)2	$\frac{(O-E)^2}{z}$
Enrolled	145	132	145 – 132 = 13	169	$\frac{169}{132} = 1.280$
Not enrolled	55	68	55 – 68 = –13	169	$\frac{169}{68} = 2.485$

Table B13

75. $df = n - 1 = 2 - 1 = 1$.

76. Using the calculator function Chi-Square GOF Test (in STAT TESTS), the test statistic is 3.7656 and the *p*-value is 0.0523. Do not reject the null hypothesis. At the five percent significance level, there is insufficient evidence to conclude that high school most recent graduating class distribution of enrolled and not enrolled does not fit that of the national distribution.

77. approximates the normal

78. skewed right

11.3: Test of Independence

79.

	Cell = Yes	Cell = No	Total
Freshman	$\frac{250(300)}{500} = 150$	$\frac{250(200)}{500} = 100$	250
Senior	$\frac{250(300)}{500} = 150$	$\frac{250(200)}{500} = 100$	250
Total	300	200	500

Table B14

80. $\frac{(100-150)^2}{150} = 16.67$
$\frac{(150-100)^2}{100} = 25$
$\frac{(200-100)^2}{150} = 16.67$
$\frac{(50-100)^2}{100} = 25$

81. Chi-square = 16.67 + 25 + 16.67 + 25 = 83.34.
df = (*r* – 1)(*c* – 1) = 1.

82. *p*-value = *P*(Chi-square, 83.34) = 0.
Reject the null hypothesis.
You could also use the calculator function STAT TESTS Chi-Square Test.

11.4: Test of Homogeneity

83. The table has five rows and two columns. *df* = (*r* – 1)(*c* – 1) = (4)(1) = 4.

11.5: Comparison Summary of the Chi-Square Tests: Goodness-of-Fit, Independence and Homogeneity

84. Using the calculator function (STAT TESTS) Chi-Square Test, the *p*-value = 0. Reject the null hypothesis. At the five percent significance level, there is sufficient evidence to conclude that the poll responses are independent of the participants' ethnic group.

85. The expected value of each cell must be at least five.

86. H_0: The variables are independent.
H_a: The variables are not independent.

87. H_0: The populations have the same distribution.
H_a: The populations do not have the same distribution.

11.6: Test of a Single Variance

88. H_0: $\sigma^2 \leq 5$
H_a: $\sigma^2 > 5$

Practice Test 4

12.1 Linear Equations

1. Which of the following equations is/are linear?

A. $y = -3x$
B. $y = 0.2 + 0.74x$
C. $y = -9.4 - 2x$
D. A and B
E. A, B, and C

2. To complete a painting job requires four hours setup time, plus one hour per 1,000 square feet. How would you express this information in a linear equation?

3. A statistics instructor is paid a per-class fee of $2,000, plus $100 for each student in the class. How would you express this information in a linear equation?

4. A tutoring school requires students to pay a one-time enrollment fee of $500, plus tuition of $3,000 per year. Express this information in an equation.

12.2: Slope and *y*-intercept of a Linear Equation

Use the following information to answer the next four exercises. For the labor costs of doing repairs, an auto mechanic charges a flat fee of $75 per car, plus an hourly rate of $55.

5. What are the independent and dependent variables for this situation?

6. Write the equation and identify the slope and intercept.

7. What is the labor charge for a job that takes 3.5 hours to complete?

8. One job takes 2.4 hours to complete, while another takes 6.3 hours. What is the difference in labor costs for these two jobs?

12.3: Scatter Plots

9. Describe the pattern in this scatter plot, and decide whether the *X* and *Y* variables would be good candidates for linear regression.

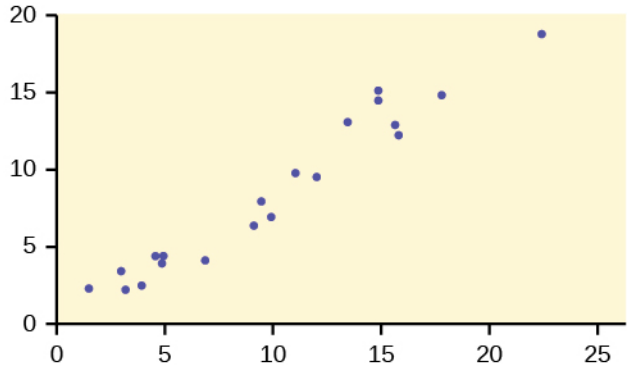


Figure B4

10. Describe the pattern in this scatter plot, and decide whether the *X* and *Y* variables would be good candidates for linear regression.

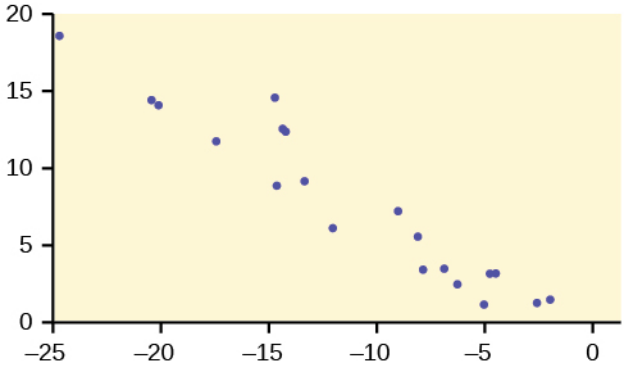


Figure B5

11. Describe the pattern in this scatter plot, and decide whether the *X* and *Y* variables would be good candidates for linear regression.

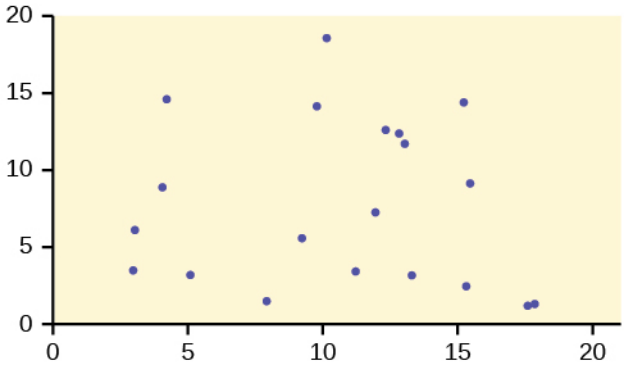


Figure B6

12. Describe the pattern in this scatter plot, and decide whether the *X* and *Y* variables would be good candidates for linear regression.

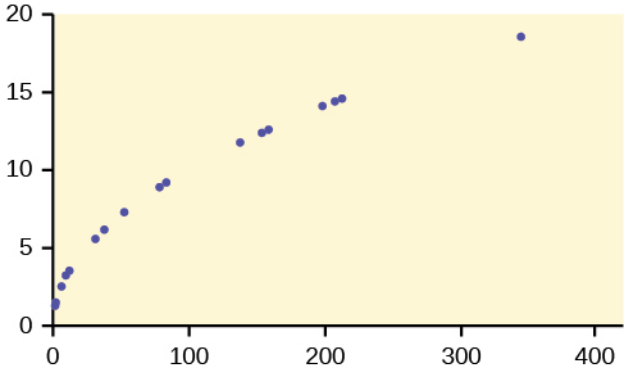


Figure B7

12.4: The Regression Equation

Use the following information to answer the next four exercises. Height (in inches) and weight (in pounds) in a sample of college freshman males have a linear relationship with the following summary statistics:
$\overline{x}$ = 68.4
$\overline{y}$ =141.6
s_x = 4.0
s_y = 9.6
r = 0.73
Let Y = weight and X = height, and write the regression equation in the form
$\hat{y} = a + bx$

13. What is the value of the slope?

14. What is the value of the *y*-intercept?

15. Write the regression equation predicting weight from height in this data set, and calculate the predicted weight for someone 68 inches tall.

12.5: Correlation Coefficient and Coefficient of Determination

16. The correlation between body weight and fuel efficiency (measured as miles per gallon) for a sample of 2,012 model cars is –0.56. Calculate the coefficient of determination for this data and explain what it means.

17. The correlation between high school GPA and freshman college GPA for a sample of 200 university students is 0.32. How much variation in freshman college GPA is not explained by high school GPA?

18. Rounded to two decimal places, what correlation between two variables is necessary to have a coefficient of determination of at least 0.50?

12.6: Testing the Significance of the Correlation Coefficient

19. Write the null and alternative hypotheses for a study to determine if two variables are significantly correlated.

20. In a sample of 30 cases, two variables have a correlation of 0.33. Do a *t* test to see if this result is significant at the α = 0.05 level. Use the formula

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

21. In a sample of 25 cases, two variables have a correlation of 0.45. Do a *t* test to see if this result is significant at the α = 0.05 level. Use the formula

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

12.7: Prediction

Use the following information to answer the next two exercises. A study relating the grams of potassium (*Y*) to the grams of fiber (*X*) per serving in enriched flour products (bread, rolls, etc.) produced the equation

$$\hat{y} = 25 + 16x$$

22. For a product with five grams of fiber per serving, what are the expected grams of potassium per serving?

23. Comparing two products, one with three grams of fiber per serving and one with six grams of fiber per serving, what is the expected difference in grams of potassium per serving?

12.8: Outliers

24. In the context of regression analysis, what is the definition of an outlier, and what is a rule of thumb to evaluate if a given value in a data set is an outlier?

25. In the context of regression analysis, what is the definition of an influential point, and how does an influential point differ from an outlier?

26. The least squares regression line for a data set is $\hat{y} = 5 + 0.3x$ and the standard deviation of the residuals is 0.4. Does a case with the values x = 2, y = 6.2 qualify as an outlier?

27. The least squares regression line for a data set is $\hat{y} = 2.3 - 0.1x$ and the standard deviation of the residuals is 0.13. Does a case with the values x = 4.1, y = 2.34 qualify as an outlier?

13.1: One-Way ANOVA

28. What are the five basic assumptions to be met if you want to do a one-way ANOVA?

29. You are conducting a one-way ANOVA comparing the effectiveness of four drugs in lowering blood pressure in hypertensive patients. What are the null and alternative hypotheses for this study?

30. What is the primary difference between the independent samples *t* test and one-way ANOVA?

31. You are comparing the results of three methods of teaching geometry to high school students. The final exam scores X_1, X_2, X_3, for the samples taught by the different methods have the following distributions:
$X_1 \sim N(85, 3.6)$
$X_1 \sim N(82, 4.8)$
$X_1 \sim N(79, 2.9)$

Each sample includes 100 students, and the final exam scores have a range of zero–100. Assuming the samples are independent and randomly selected, have the requirements for conducting a one-way ANOVA been met? Explain why or why not for each assumption.

32. You conduct a study comparing the effectiveness of four types of fertilizer to increase crop yield on wheat farms. When examining the sample results, you find that two of the samples have an approximately normal distribution, and two have an approximately uniform distribution. Is this a violation of the assumptions for conducting a one-way ANOVA?

13.2: The *F* Distribution

Use the following information to answer the next seven exercises. You are conducting a study of three types of feed supplements for cattle to test their effectiveness in producing weight gain among calves whose feed includes one of the supplements. You have four groups of 30 calves (one is a control group receiving the usual feed, but no supplement). You will conduct a one-way ANOVA after one year to see if there are differences in the mean weight for the four groups.

33. What is SS_{within} in this experiment, and what does it mean?

34. What is $SS_{between}$ in this experiment, and what does it mean?

35. What are k and i for this experiment?

36. If $SS_{within} = 374.5$ and $SS_{total} = 621.4$ for this data, what is $SS_{between}$?

37. What are $MS_{between}$, and MS_{within} for this experiment?

38. What is the F statistic for this data?

39. If there had been 35 calves in each group, instead of 30, with the sums of squares remaining the same, would the F statistic be larger or smaller?

13.3: Facts About the *F* Distribution

40. Which of the following numbers are possible F statistics?

A. 2.47
B. 5.95
C. –3.61
D. 7.28
E. 0.97

41. Histograms $F1$ and $F2$ below display the distribution of cases from samples from two populations, one distributed $F_{3,15}$ and one distributed $F_{5,500}$. Which sample came from which population?

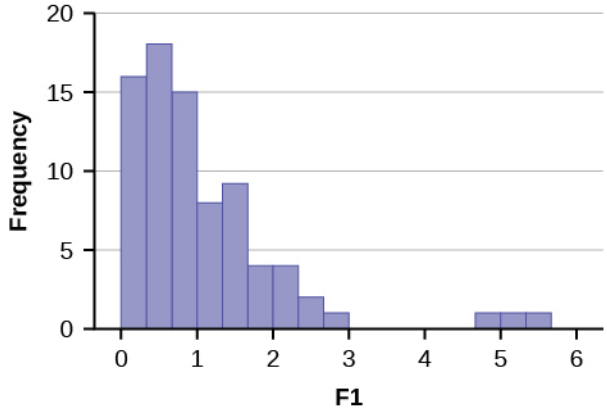


Figure B8

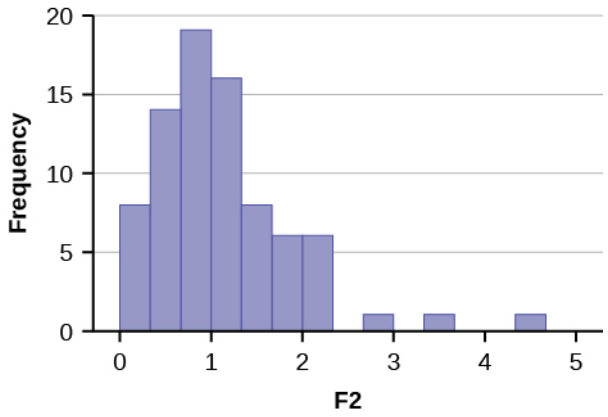


Figure B9

42. The *F* statistic from an experiment with $k = 3$ and $n = 50$ is 3.67. At $\alpha = 0.05$, will you reject the null hypothesis?

43. The *F* statistic from an experiment with $k = 4$ and $n = 100$ is 4.72. At $\alpha = 0.01$, will you reject the null hypothesis?

13.4: Test of Two Variances

44. What assumptions must be met to perform the *F* test of two variances?

45. You believe there is greater variance in grades given by the math department at your university than in the English department. You collect all the grades for undergraduate classes in the two departments for a semester, compute the variance of each, and conduct an *F* test of two variances. What are the null and alternative hypotheses for this study?

Practice Test 4 Solutions

12.1 Linear Equations

1. e. A, B, and C.
All three are linear equations of the form $y = mx + b$.

2. Let y = the total number of hours required, and x the square footage, measured in units of 1,000. The equation is $y = x + 4$

3. Let y = the total payment, and x the number of students in a class. The equation is $y = 100(x) + 2{,}000$

4. Let y = the total cost of attendance, and x the number of years enrolled. The equation is $y = 3{,}000(x) + 500$

12.2: Slope and *y*-intercept of a Linear Equation

5. The independent variable is the hours worked on a car. The dependent variable is the total labor charges to fix a car.

6. Let y = the total charge, and x the number of hours required. The equation is $y = 55x + 75$
The slope is 55 and the intercept is 75.

7. $y = 55(3.5) + 75 = 267.50$

8. Because the intercept is included in both equations, while you are only interested in the difference in costs, you do not need to include the intercept in the solution. The difference in number of hours required is 6.3 – 2.4 = 3.9.
Multiply this difference by the cost per hour: 55(3.9) = 214.5.
The difference in cost between the two jobs is $214.50.

12.3: Scatter Plots

9. The *X* and *Y* variables have a strong linear relationship. These variables would be good candidates for analysis with linear regression.

10. The *X* and *Y* variables have a strong negative linear relationship. These variables would be good candidates for analysis with linear regression.

11. There is no clear linear relationship between the *X* and *Y* variables, so they are not good candidates for linear regression.

12. The *X* and *Y* variables have a strong positive relationship, but it is curvilinear rather than linear. These variables are not good candidates for linear regression.

12.4: The Regression Equation

13. $r\left(\frac{s_y}{s_x}\right) = 0.73\left(\frac{9.6}{4.0}\right) = 1.752 \approx 1.75$

14. $a = \overline{y} - b\overline{x} = 141.6 - 1.752(68.4) = 21.7632 \approx 21.76$

15. $\hat{y} = 21.76 + 1.75(68) = 140.76$

12.5: Correlation Coefficient and Coefficient of Determination

16. The coefficient of determination is the square of the correlation, or r^2.
For this data, r^2 = (–0.56)2 = 0.3136 ≈ 0.31 or 31 percent. This means that 31 percent of the variation in fuel efficiency can be explained by the bodyweight of the automobile.

17. The coefficient of determination = 0.32^2 = 0.1024. This is the amount of variation in freshman college GPA that can be explained by high school GPA. The amount that cannot be explained is 1 – 0.1024 = 0.8976 ≈ 0.90. So, about 90 percent of variance in freshman college GPA in this data is not explained by high school GPA.

18. $r = \sqrt{r^2}$
$\sqrt{0.5} = 0.707106781 \approx 0.71$
You need a correlation of 0.71 or higher to have a coefficient of determination of at least 0.5.

12.6: Testing the Significance of the Correlation Coefficient

19. H_0: $\rho = 0$
H_a: $\rho \neq 0$

20. $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.33\sqrt{30-2}}{\sqrt{1-0.33^2}} = 1.85$
The critical value for $\alpha = 0.05$ for a two-tailed test using the t_{29} distribution is 2.045. Your value is less than this, so you fail to reject the null hypothesis and conclude that the study produced no evidence that the variables are significantly correlated.
Using the calculator function tcdf, the p-value is 2tcdf(1.85, 10^99, 29) = 0.0373. Do not reject the null hypothesis and conclude that the study produced no evidence that the variables are significantly correlated.

21. $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.45\sqrt{25-2}}{\sqrt{1-0.45^2}} = 2.417$
The critical value for $\alpha = 0.05$ for a two-tailed test using the t_{24} distribution is 2.064. Your value is greater than this, so you reject the null hypothesis and conclude that the study produced evidence that the variables are significantly correlated.
Using the calculator function tcdf, the p-value is 2tcdf(2.417, 10^99, 24) = 0.0118. Reject the null hypothesis and conclude that the study produced evidence that the variables are significantly correlated.

12.7: Prediction

22. $\hat{y} = 25 + 16(5) = 105$

23. Because the intercept appears in both predicted values, you can ignore it in calculating a predicted difference score. The difference in grams of fiber per serving is 6 – 3 = 3, and the predicted difference in grams of potassium per serving is (16)(3) = 48.

12.8: Outliers

24. An outlier is an observed value that is far from the least squares regression line. A rule of thumb is that a point more than two standard deviations of the residuals from its predicted value on the least squares regression line is an outlier.

25. An influential point is an observed value in a data set that is far from other points in the data set, in a horizontal direction. Unlike an outlier, an influential point is determined by its relationship with other values in the data set, not by its relationship to the regression line.

26. The predicted value for y is $\hat{y} = 5 + 0.3x = 5.6$. The value of 6.2 is less than two standard deviations from the predicted value, so it does not qualify as an outlier.
Residual for (2, 6.2): 6.2 – 5.6 = 0.6 (0.6 < 2(0.4))

27. The predicted value for y is $\hat{y}$ = 2.3 – 0.1(4.1) = 1.89. The value of 2.32 is more than two standard deviations from the predicted value, so it qualifies as an outlier.

Residual for (4.1, 2.34): 2.32 – 1.89 = 0.43 (0.43 > 2(0.13))

13.1: One-Way ANOVA

28.

1. Each sample is drawn from a normally distributed population.
2. All samples are independent and randomly selected.
3. The populations from which the samples are drawn have equal standard deviations.
4. The factor is a categorical variable.
5. The response is a numerical variable.

29. H_0: $\mu 1 = \mu 2 = \mu 3 = \mu 4$
H_a: At least two of the group means $\mu 1$, $\mu 2$, $\mu 3$, $\mu 4$ are not equal.

30. The independent samples *t* test can only compare means from two groups, while one-way ANOVA can compare means of more than two groups.

31. Each sample appears to have been drawn from normally distributed populations, the factor is a categorical variable (method), the outcome is a numerical variable (test score), and you were told the samples were independent and randomly selected, so those requirements are met. However, each sample has a different standard deviation, and this suggests that the populations from which they were drawn also have different standard deviations, which is a violation of an assumption for one-way ANOVA. Further statistical testing will be necessary to test the assumption of equal variance before proceeding with the analysis.

32. One of the assumptions for a one-way ANOVA is that the samples are drawn from normally distributed populations. Since two of your samples have an approximately uniform distribution, this casts doubt on whether this assumption has been met. Further statistical testing will be necessary to determine if you can proceed with the analysis.

13.2: The *F* Distribution

33. SS_{within} is the sum of squares within groups, representing the variation in outcome that cannot be attributed to the different feed supplements but due to individual or chance factors among the calves in each group.

34. $SS_{between}$ is the sum of squares between groups, representing the variation in outcome that can be attributed to the different feed supplements.

35. k = the number of groups = 4
n_1 = the number of cases in group 1 = 30
n = the total number of cases = 4(30) = 120

36. $SS_{total} = SS_{within} + SS_{between}$, so $SS_{between} = SS_{total} - SS_{within}$
621.4 – 374.5 = 246.9

37. The mean squares in an ANOVA are found by dividing each sum of squares by its respective degrees of freedom (*df*).
For SS_{total}, $df = n - 1 = 120 - 1 = 119$.
For $SS_{between}$, df = k – 1 = 4 – 1 = 3.
For SS_{within}, $df = 120 - 4 = 116$.
$MS_{between} = \frac{246.9}{3} = 82.3$
$MS_{within} = \frac{374.5}{116} = 3.23$

38. $F = \frac{MS_{between}}{MS_{within}} = \frac{82.3}{3.23} = 25.48$

39. It would be larger, because you would be dividing by a smaller number. The value of $MS_{between}$ would not change with a change of sample size, but the value of MS_{within} would be smaller, because you would be dividing by a larger number (df_{within} would be 136, not 116). Dividing a constant by a smaller number produces a larger result.

13.3: Facts About the *F* Distribution

40. All but choice c, –3.61. *F* Statistics are always greater than or equal to 0.

41. As the degrees of freedom increase in an *F* distribution, the distribution becomes more nearly normal. Histogram *F*2 is closer to a normal distribution than histogram *F*1, so the sample displayed in histogram *F*1 was drawn from the $F_{3,15}$ population, and the sample displayed in histogram *F*2 was drawn from the $F_{5,500}$ population.

42. Using the calculator function Fcdf, *p*-value = Fcdf(3.67, 1E, 3, 50) = 0.0182. Reject the null hypothesis.

43. Using the calculator function Fcdf, *p*-value = Fcdf(4.72, 1E, 4, 100) = 0.0016 Reject the null hypothesis.

13.4: Test of Two Variances

44. The samples must be drawn from populations that are normally distributed, and must be drawn from independent populations.

45. Let σ_M^2 = variance in math grades, and σ_E^2 = variance in English grades.
H_0: $\sigma_M^2 \leq \sigma_E^2$
H_a: $\sigma_M^2 > \sigma_E^2$

Practice Final Exam 1

Use the following information to answer the next two exercises. An experiment consists of tossing two, 12-sided dice (the numbers 1–12 are printed on the sides of each die).

- Let Event *A* = both dice show an even number.
- Let Event *B* = both dice show a number greater than eight

1. Events *A* and *B* are

A. Mutually exclusive
B. Independent
C. Mutually exclusive and independent
D. Neither mutually exclusive nor independent

2. Find $P(A|B)$.

A. $\frac{2}{4}$
B. $\frac{16}{144}$
C. $\frac{4}{16}$
D. $\frac{2}{144}$

3. Which of the following are TRUE when we perform a hypothesis test on matched or paired samples?

A. Sample sizes are almost never small.
B. Two measurements are drawn from the same pair of individuals or objects.
C. Two sample means are compared to each other.
D. Answer choices b and c are both true.

Use the following information to answer the next two exercises. One hundred eighteen students were asked what type of color their bedrooms were painted: light colors, dark colors, or vibrant colors. The results were tabulated according to gender.

	Light colors	Dark colors	Vibrant colors
Female	20	22	28
Male	10	30	8

Table B15

4. Find the probability that a randomly chosen student is male or has a bedroom painted with light colors.

A. $\frac{10}{118}$
B. $\frac{68}{118}$
C. $\frac{48}{118}$
D. $\frac{10}{48}$

5. Find the probability that a randomly chosen student is male given the student's bedroom is painted with dark colors.

A. $\frac{30}{118}$

B. $\frac{30}{48}$
C. $\frac{22}{118}$
D. $\frac{30}{52}$

Use the following information to answer the next two exercises. We are interested in the number of times a teenager must be reminded to do his or her chores each week. A survey of 40 mothers was conducted. Table B16 shows the results of the survey.

x	$P(x)$
0	$\frac{2}{40}$
1	$\frac{5}{40}$
2	
3	$\frac{14}{40}$
4	$\frac{7}{40}$
5	$\frac{4}{40}$

Table B16

6. Find the probability that a teenager is reminded two times.

A. 8
B. $\frac{8}{40}$
C. $\frac{6}{40}$
D. 2

7. Find the expected number of times a teenager is reminded to do his or her chores.

A. 15
B. 2.78
C. 1.0
D. 3.13

Use the following information to answer the next two exercises. On any given day, approximately 37.5 percent of the cars parked in the De Anza parking garage are parked crookedly. We randomly survey 22 cars. We are interested in the number of cars that are parked crookedly.

8. For every 22 cars, how many would you expect to be parked crookedly, on average?

A. 8.25
B. 11
C. 18
D. 7.5

9. What is the probability that at least 10 of the 22 cars are parked crookedly?

A. 0.1263
B. 0.1607
C. 0.2870
D. 0.8393

10. Using a sample of 15 Stanford-Binet IQ scores, we wish to conduct a hypothesis test. Our claim is that the mean IQ score on the Stanford-Binet IQ test is more than 100. It is known that the standard deviation of all Stanford-Binet IQ

scores is 15 points. Which of the following is the correct distribution to use for the hypothesis test?

A. Binomial
B. Student's *t*
C. Normal
D. Uniform

Use the following information to answer the next three exercises. De Anza College keeps statistics on the pass rate of students who enroll in math classes. In a sample of 1,795 students enrolled in Math 1A (1st quarter calculus), 1,428 passed the course. In a sample of 856 students enrolled in Math 1B (2nd quarter calculus), 662 passed. In general, are the pass rates of Math 1A and Math 1B statistically the same? Let A = the subscript for Math 1A and B = the subscript for Math 1B.

11. If you were to conduct an appropriate hypothesis test, the alternate hypothesis would be

A. H_a: $p_A = p_B$
B. H_a: $p_A > p_B$
C. H_o: $p_A = p_B$
D. H_a: $p_A \neq p_B$

12. The Type I error is to

A. conclude that the pass rate for Math 1A is the same as the pass rate for Math 1B when, in fact, the pass rates are different.
B. conclude that the pass rate for Math 1A is different than the pass rate for Math 1B when, in fact, the pass rates are the same.
C. conclude that the pass rate for Math 1A is greater than the pass rate for Math 1B when, in fact, the pass rate for Math 1A is less than the pass rate for Math 1B.
D. conclude that the pass rate for Math 1A is the same as the pass rate for Math 1B when, in fact, they are the same.

13. The correct decision is to

A. reject H_0.
B. not reject H_0.
C. There is not enough information given to conduct the hypothesis test.

Kia, Alejandra, and Iris are runners on the track teams at three different schools. Their running times, in minutes, and the statistics for the track teams at their respective schools, for a one mile run, are given in the table below:

	Running Time	School Average Running Time	School Standard Deviation
Kia	4.9	5.2	0.15
Alejandra	4.2	4.6	0.25
Iris	4.5	4.9	0.12

Table B17

14. Which student is the BEST when compared to the other runners at her school?

A. Kia
B. Alejandra
C. Iris
D. Impossible to determine

Use the following information to answer the next two exercises. The following adult ski sweater prices are from the Gorsuch Ltd. Winter catalog: $212, $292, $278, $199, $280, $236.

Assume the underlying sweater price population is approximately normal. The null hypothesis is that the mean price of adult ski sweaters from Gorsuch Ltd. is at least $275.

15. Which of the following is the correct distribution to use for the hypothesis test?

A. Normal
B. Binomial
C. Student's *t*
D. Exponential

16. The hypothesis test

A. is two-tailed.
B. is left-tailed.
C. is right-tailed.
D. has no tails.

17. Sara, a statistics student, wanted to determine the mean number of books that college professors have in their office. She randomly selected two buildings on campus and asked each professor in the selected buildings how many books are in his or her office. Sara surveyed 25 professors. The type of sampling selected is

A. simple random sampling.
B. systematic sampling.
C. cluster sampling.
D. stratified sampling.

18. A clothing store would use which measure of the center of data when placing orders for the typical *middle* customer?

A. Mean
B. Median
C. Mode
D. IQR

19. In a hypothesis test, the *p*-value is

A. the probability that an outcome of the data will happen purely by chance when the null hypothesis is true.
B. called the preconceived alpha.
C. compared to beta to decide whether to reject or not reject the null hypothesis.
D. Answer choices A and B are both true.

Use the following information to answer the next three exercises. A community college offers classes six days a week: Monday through Saturday. Maria conducted a study of the students in her classes to determine how many days per week the students who are in her classes come to campus for classes. In each of her five classes she randomly selected 10 students and asked them how many days they come to campus for classes. Each of her classes are the same size. The results of her survey are summarized in Table B18.

Number of Days on Campus	Frequency	Relative Frequency	Cumulative Relative Frequency
1	2		
2	12	.24	
3	10	.20	
4			.98
5	0		
6	1	.02	1

Table B18

20. Combined with convenience sampling, what other sampling technique did Maria use?

A. Simple random
B. Systematic

C. Cluster
D. Stratified

21. How many students come to campus for classes four days a week?

A. 49
B. 25
C. 30
D. 13

22. What is the 60^{th} percentile for this data?

A. 2
B. 3
C. 4
D. 5

Use the following information to answer the next two exercises. The following data are the results of a random survey of 110 reservists called to active duty to increase security at California airports.

Number of Dependents	Frequency
0	11
1	27
2	33
3	20
4	19

Table B19

23. Construct a 95 percent confidence interval for the true population mean number of dependents of reservists called to active duty to increase security at California airports.

A. (1.85, 2.32)
B. (1.80, 2.36)
C. (1.97, 2.46)
D. (1.92, 2.50)

24. The 95 percent confidence interval above means:

A. Five percent of confidence intervals constructed this way will not contain the true population aveage number of dependents.
B. We are 95 percent confident the true population mean number of dependents falls in the interval.
C. Both of the above answer choices are correct.
D. None of the above.

25. $X \sim U(4, 10)$. Find the 30^{th} percentile.

A. 0.3000
B. 3
C. 5.8
D. 6.1

26. If $X \sim Exp(0.8)$, then $P(x < \mu) =$ —

A. 0.3679
B. 0.4727
C. 0.6321
D. cannot be determined

27. The lifetime of a computer circuit board is normally distributed with a mean of 2,500 hours and a standard deviation of 60 hours. What is the probability that a randomly chosen board will last at most 2,560 hours?

A. 0.8413
B. 0.1587
C. 0.3461
D. 0.6539

28. A survey of 123 reservists called to active duty as a result of the September 11, 2001, attacks was conducted to determine the proportion that were married. Eighty-six reported being married. Construct a 98 percent confidence interval for the true population proportion of reservists called to active duty that are married.

A. (0.6030, 0.7954)
B. (0.6181, 0.7802)
C. (0.5927, 0.8057)
D. (0.6312, 0.7672)

29. Winning times in 26 mile marathons run by world class runners average 145 minutes with a standard deviation of 14 minutes. A sample of the last 10 marathon winning times is collected. Let x = mean winning times for 10 marathons. The distribution for x is

A. $N\left(145, \frac{14}{\sqrt{10}}\right)$
B. $N(145,14)$
C. t_9
D. t_{10}

30. Suppose that Phi Beta Kappa honors the top 1 percent of college and university seniors. Assume that grade point means (GPA) at a certain college are normally distributed with a 2.5 mean and a standard deviation of 0.5. What would be the minimum GPA needed to become a member of Phi Beta Kappa at that college?

A. 3.99
B. 1.34
C. 3.00
D. 3.66

The number of people living on American farms has declined steadily during the 20th century. Here are data on the farm population (in millions of persons) from 1935 to 1980.

Year	1935	1940	1945	1950	1955	1960	1965	1970	1975	1980
Population	32.1	30.5	24.4	23	19.1	15.6	12.4	9.7	8.9	7.2

Table B20

31. The linear regression equation is $\hat{y}$ = 1166.93 – 0.5868x. What was the expected farm population in millions of persons for 1980?

A. 7.2
B. 5.1
C. 6
D. 8

32. In linear regression, which is the best possible *SSE*?

A. 13.46
B. 18.22
C. 24.05
D. 16.33

33. In regression analysis, if the correlation coefficient is close to one, what can be said about the best fit line?

A. It is a horizontal line. Therefore, we cannot use it.
B. There is a strong linear pattern. Therefore, it is most likely a good model to be used.

C. The coefficient correlation is close to the limit. Therefore, it is hard to make a decision.
D. We do not have the equation. Therefore, we cannot say anything about it.

Use the following information to answer the next three exercises. A study of the career plans of young women and men sent questionnaires to all 722 members of the senior class in the College of Business Administration at the University of Illinois. One question asked which major within the business program the student had chosen. Here are the data from the students who responded.

	Female	Male
Accounting	68	56
Administration	91	40
Economics	5	6
Finance	61	59

Table B21 Does the data suggest that there is a relationship between the gender of students and their choice of major?

34. The distribution for the test is

A. $\text{Chi}^2{}_8$.
B. $\text{Chi}^2{}_3$.
C. t_{721}.
D. $N(0, 1)$.

35. The expected number of females who choose finance is

A. 37.
B. 61.
C. 60.
D. 70.

36. The *p*-value is 0.0127 and the level of significance is 0.05. The conclusion to the test is:

A. there is insufficient evidence to conclude that the choice of major and the gender of the student are not independent of each other.
B. there is sufficient evidence to conclude that the choice of major and the gender of the student are not independent of each other.
C. there is sufficient evidence to conclude that students find economics very hard.
D. there is in sufficient evidence to conclude that more females prefer administration than males.

37. An agency reported that the work force nationwide is composed of 10 percent professional, 10 percent clerical, 30 percent skilled, 15 percent service, and 35 percent semiskilled laborers. A random sample of 100 San Jose residents indicated 15 professional, 15 clerical, 40 skilled, 10 service, and 20 semiskilled laborers. At $\alpha = 0.10$, does the work force in San Jose appear to be consistent with the agency report for the nation? Which kind of test is it?

A. Chi^2 goodness of fit
B. Chi^2 test of independence
C. Independent groups proportions
D. Unable to determine

Practice Final Exam 1 Solutions

Solutions

1. B independent

2. C $\frac{4}{16}$

3. B Two measurements are drawn from the same pair of individuals or objects.

4. B $\frac{68}{118}$

5. D $\frac{30}{52}$

6. B $\frac{8}{40}$

7. B 2.78

8. A 8.25

9. C 0.2870

10. C Normal

11. D H_a: $p_A \neq p_B$

12. B conclude that the pass rate for Math 1A is different than the pass rate for Math 1B when, in fact, the pass rates are the same.

13. B not reject H_0

14. C Iris

15. C Student's *t*

16. B is left-tailed.

17. C cluster sampling

18. B median

19. A the probability that an outcome of the data will happen purely by chance when the null hypothesis is true.

20. D stratified

21. B 25

22. C 4

23. A (1.85, 2.32)

24. C Both above are correct.

25. C 5.8

26. C 0.6321

27. A 0.8413

28. A (0.6030, 0.7954)

29. A $N\left(145, \frac{14}{\sqrt{10}}\right)$

30. D 3.66

31. B 5.1

32. A 13.46

33. B There is a strong linear pattern. Therefore, it is most likely a good model to be used.

34. B $\text{Chi}^2{}_3$.

35. D 70

36. B There is sufficient evidence to conclude that the choice of major and the gender of the student are not independent of each other.

37. A Chi^2 goodness-of-fit

Practice Final Exam 2

1. A study was done to determine the proportion of teenagers that own a car. The population proportion of teenagers that own a car is the

A. statistic.
B. parameter.
C. population.
D. variable.

Use the following information to answer the next two exercises.

value	frequency
0	1
1	4
2	7
3	9
6	4

Table B22

2. The box plot for the data is

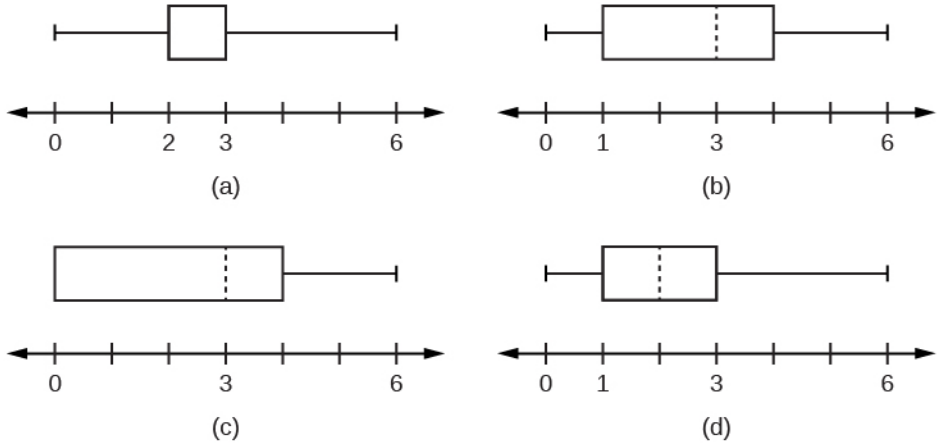


Figure B10

3. If six were added to each value of the data in the table, the 15th percentile of the new list of values is would be

A. six
B. one
C. seven
D. eight

Use the following information to answer the next two exercises. Suppose that the probability of a drought in any independent year is 20 percent. Out of those years in which a drought occurs, the probability of water rationing is 10 percent. However, in any year, the probability of water rationing is 5 percent.

4. What is the probability of both a drought and water rationing occurring?

A. 0.05
B. 0.01

C. 0.02
D. 0.30

5. Which of the following is true?

A. Drought and water rationing are independent events.
B. Drought and water rationing are mutually exclusive events.
C. None of the above.

Use the following information to answer the next two exercises. Suppose that a survey yielded the following data:

gender	apple	pumpkin	pecan
female	40	10	30
male	20	30	10

Table B23 Favorite Pie

6. Suppose that one individual is randomly chosen. The probability that the person's favorite pie is apple or the person is male is —

A. $\frac{40}{60}$
B. $\frac{60}{140}$
C. $\frac{120}{140}$
D. $\frac{100}{140}$

7. Suppose H_0 is favorite pie and gender are independent. The *p*-value is —

A. ≈ 0
B. 1
C. 0.05
D. Cannot be determined

Use the following information to answer the next two exercises. Let's say that the probability that an adult watches the news at least once per week is 0.60. We randomly survey 14 people. Of interest is the number of people who watch the news at least once per week.

8. Which of the following statements is FALSE?

A. $X \sim B(14\ 0.60)$
B. The values for x are {1, 2, 3, . . . 14}.
C. $\mu = 8.4$
D. $P(X = 5) = 0.0408$

9. Find the probability that at least six adults watch the news at least once per week.

A. $\frac{6}{14}$
B. 0.8499
C. 0.9417
D. 0.6429

10. The following histogram is most likely to be a result of sampling from which distribution?

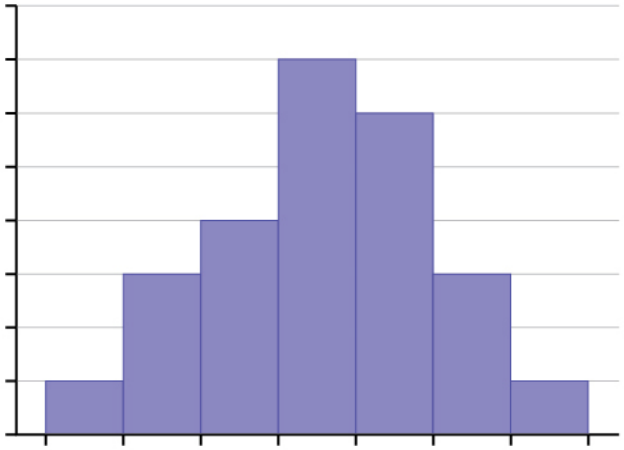


Figure B11

A. Chi-square with $df = 6$
B. Exponential
C. Uniform
D. Binomial

11. The ages of campus day and evening students is known to be normally distributed. A sample of six campus day and evening students reported their ages (in years) as {18, 35, 27, 45, 20, 20}. What is the error bound for the 90 percent confidence interval of the true average age?

A. 11.2
B. 22.3
C. 17.5
D. 8.7

12. If a normally distributed random variable has $\mu = 0$ and $\sigma = 1$, then 97.5 percent of the population values lie above

A. –1.96
B. 1.96
C. 1
D. –1

Use the following information to answer the next three exercises. The amount of money a customer spends in one trip to the supermarket is known to have an exponential distribution. Suppose the average amount of money a customer spends in one trip to the supermarket is $72.

13. What is the probability that one customer spends less than $72 in one trip to the supermarket?

A. 0.6321
B. 0.5000
C. 0.3714
D. 1

14. How much money altogether would you expect the next five customers to spend in one trip to the supermarket (in dollars)?

A. 72
B. $\frac{72^2}{5}$
C. 5184
D. 360

15. If you want to find the probability that the mean amount of money 50 customers spend in one trip to the supermarket is less than $60, the distribution to use is

A. $N(72, 72)$
B. $N\left(72, \frac{72}{\sqrt{50}}\right)$

C. *Exp*(72)
D. $Exp\left(\frac{1}{72}\right)$

Use the following information to answer the next three exercises. The amount of time it takes a fourth grader to carry out the trash is uniformly distributed in the interval from one to 10 minutes.

16. What is the probability that a randomly chosen fourth grader takes more than seven minutes to take out the trash?

A. $\frac{3}{9}$
B. $\frac{7}{9}$
C. $\frac{3}{10}$
D. $\frac{7}{10}$

17. Which graph best shows the probability that a randomly chosen fourth grader takes more than six minutes to take out the trash, given that he or she has already taken more than three minutes?

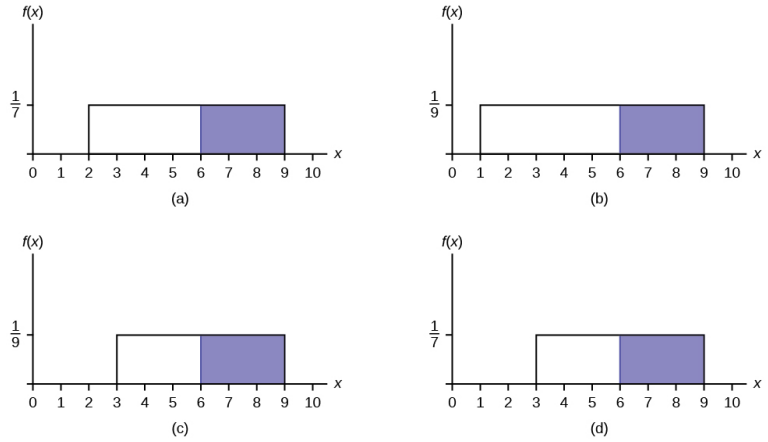


Figure B12

18. We should expect a fourth grader to take how many minutes to take out the trash?

A. 4.5
B. 5.5
C. 5
D. 10

Use the following information to answer the next three exercises. At the beginning of the quarter, the amount of time a student waits in line at the campus cafeteria is normally distributed with a mean of five minutes and a standard deviation of 1.5 minutes.

19. What is the 90th percentile of waiting times in minutes?

A. 1.28
B. 90
C. 7.47
D. 6.92

20. The median waiting time in minutes for one student is

A. 5
B. 50
C. 2.5
D. 1.5

21. Find the probability that the average wait time for ten students is at most 5.5 minutes.

A. 0.6301
B. 0.8541

C. 0.3694
D. 0.1459

22. A sample of 80 software engineers in Silicon Valley is taken, and it is found that 20 percent of them earn approximately $50,000 per year. A point estimate for the true proportion of engineers in Silicon Valley who earn $50,000 per year is

A. 16
B. 0.2
C. 1
D. 0.95

23. If $P(Z < z_\alpha) = 0.1587$ where $Z \sim N(0, 1)$, then α is equal to

A. –1
B. 0.1587
C. 0.8413
D. 1

24. A professor tested 35 students to determine their entering skills. At the end of the term, after completing the course, the same test was administered to the same 35 students to study their improvement. This would be a test of

A. independent groups
B. two proportions
C. matched pairs, dependent groups
D. exclusive groups

A math exam was given to all the third-grade children attending ABC School. Two random samples of scores were taken.

	n	$\overline{x}$	*s*
Boys	55	82	5
Girls	60	86	7

Table B24

25. Which of the following correctly describes the results of a hypothesis test of the claim, "There is a difference between the mean scores obtained by third-grade girls and boys at the 5 percent level of significance"?

A. Do not reject H_0. There is insufficient evidence to conclude that there is a difference in the mean scores.
B. Do not reject H_0. There is sufficient evidence to conclude that there is a difference in the mean scores.
C. Reject H_0. There is insufficient evidence to conclude that there is no difference in the mean scores.
D. Reject H_0. There is sufficient evidence to conclude that there is a difference in the mean scores.

26. In a survey of 80 males, 45 had played an organized sport growing up. Of the 70 females surveyed, 25 had played an organized sport growing up. We are interested in whether the proportion for males is higher than the proportion for females. The correct conclusion is that

A. There is insufficient information to conclude that the proportion for males is the same as the proportion for females.
B. There is insufficient information to conclude that the proportion for males is not the same as the proportion for females.
C. There is sufficient evidence to conclude that the proportion for males is higher than the proportion for females.
D. There is not enough information to make a conclusion.

27. From past experience, a statistics teacher has found that the average score on a midterm is 81, with a standard deviation of 5.2. This term, a class of 49 students had a standard deviation of 5 on the midterm. Do the data indicate that we should reject the teacher's claim that the standard deviation is 5.2? Use $\alpha = 0.05$.

A. Yes
B. No
C. Not enough information given to solve the problem

28. Three loading machines are being compared. Ten samples were taken for each machine. Machine I took an average of 31 minutes to load packages, with a standard deviation of two minutes. Machine II took an average of 28 minutes to load packages, with a standard deviation of 1.5 minutes. Machine III took an average of 29 minutes to load packages, with a standard deviation of one minute. Find the *p*-value when testing that the average loading times are the same.

A. *p*-value is close to zero
B. *p*-value is close to one
C. Not enough information given to solve the problem

Use the following information to answer the next three exercises. A corporation has offices in different parts of the country. It has gathered the following information concerning the number of bathrooms and the number of employees at seven sites:

Number of employees *x*	650	730	810	900	102	107	1150
Number of bathrooms *y*	40	50	54	61	82	110	121

Table B25

29. Is the correlation between the number of employees and the number of bathrooms significant?

A. Yes
B. No
C. Not enough information to answer question

30. The linear regression equation is

A. $\hat{y} = 0.0094 - 79.96x$
B. $\hat{y} = 79.96 + 0.0094x$
C. $\hat{y} = 79.96 - 0.0094x$
D. $\hat{y} = -0.0094 + 79.96x$

31. If a site has 1,150 employees, approximately how many bathrooms should it have?

A. 69
B. 91
C. 91,954
D. We should not be estimating here.

32. Suppose that a sample of size 10 was collected, with $\overline{x} = 4.4$ and $s = 1.4$. H_0: $\sigma^2 = 1.6$ vs. H_a: $\sigma^2 \neq 1.6$. Which graph best describes the results of the test?

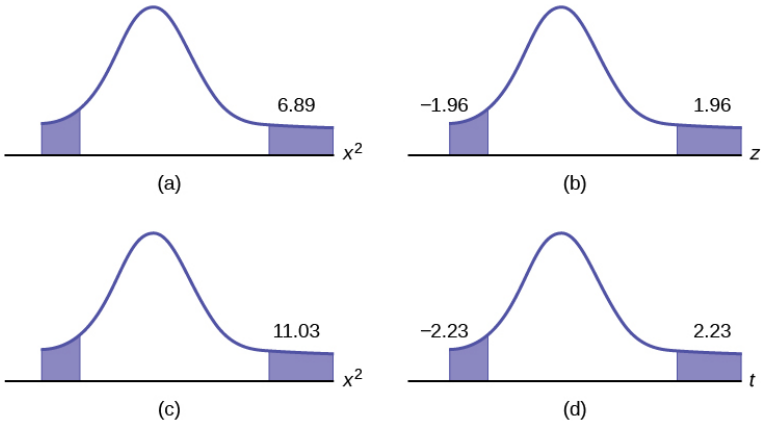


Figure B13

Sixty-four backpackers were asked the number of days since their latest backpacking trip. The number of days is given in Table B26.

# of days	1	2	3	4	5	6	7	8
Frequency	5	9	6	12	7	10	5	10

Table B26

33. Conduct an appropriate test to determine if the distribution is uniform.

A. The *p*-value is > 0.10. There is insufficient information to conclude that the distribution is not uniform.
B. The *p*-value is < 0.01. There is sufficient information to conclude the distribution is not uniform.
C. The *p*-value is between 0.01 and 0.10, but without alpha (α) there is not enough information.
D. There is no such test that can be conducted.

34. Which of the following statements is true when using one-way ANOVA?

A. The populations from which the samples are selected have different distributions.
B. The sample sizes are large.
C. The test is to determine if the different groups have the same means.
D. There is a correlation between the factors of the experiment.

Practice Final Exam 2 Solutions

Solutions

1. B parameter.

2. A

3. C seven

4. C 0.02

5. C none of the above

6. D $\frac{100}{140}$

7. A ≈ 0

8. B The values for *x* are: {1, 2, 3, . . . 14}

9. C 0.9417.

10. D binomial

11. D 8.7

12. A –1.96

13. A 0.6321

14. D 360

15. B $N\left(72, \frac{72}{\sqrt{50}}\right)$

16. A $\frac{3}{9}$

17. D

18. B 5.5

19. D 6.92

20. A 5

21. B 0.8541

22. B 0.2

23. A –1.

24. C matched pairs, dependent groups.

25. D Reject H_0. There is sufficient evidence to conclude that there is a difference in the mean scores.

26. C there is sufficient evidence to conclude that the proportion for males is higher than the proportion for females.

27. B no

28. B *p*-value is close to 1.

29. B No

30. C $\hat{y} = 79.96x - 0.0094$

31. D We should not be estimating here.

32. A

33. A The *p*-value is > 0.10. There is insufficient information to conclude that the distribution is not uniform.

34. C The test is to determine if the different groups have the same means.

C DATA SETS

Lap Times

The following tables provide lap times from Terri Vogel's log book. Times are recorded in seconds for 2.5-mile laps completed in a series of races and practice runs.

	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Race 1	135	130	131	132	130	131	133
Race 2	134	131	131	129	128	128	129
Race 3	129	128	127	127	130	127	129
Race 4	125	125	126	125	124	125	125
Race 5	133	132	132	132	131	130	132
Race 6	130	130	130	129	129	130	129
Race 7	132	131	133	131	134	134	131
Race 8	127	128	127	130	128	126	128
Race 9	132	130	127	128	126	127	124
Race 10	135	131	131	132	130	131	130
Race 11	132	131	132	131	130	129	129
Race 12	134	130	130	130	131	130	130
Race 13	128	127	128	128	128	129	128
Race 14	132	131	131	131	132	130	130
Race 15	136	129	129	129	129	129	129
Race 16	129	129	129	128	128	129	129
Race 17	134	131	132	131	132	132	132
Race 18	129	129	130	130	133	133	127
Race 19	130	129	129	129	129	129	128
Race 20	131	128	130	128	129	130	130

Table C1 Race Lap Times (in seconds)

	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Practice 1	142	143	180	137	134	134	172
Practice 2	140	135	134	133	128	128	131
Practice 3	130	133	130	128	135	133	133
Practice 4	141	136	137	136	136	136	145
Practice 5	140	138	136	137	135	134	134
Practice 6	142	142	139	138	129	129	127
Practice 7	139	137	135	135	137	134	135
Practice 8	143	136	134	133	134	133	132
Practice 9	135	134	133	133	132	132	133
Practice 10	131	130	128	129	127	128	127
Practice 11	143	139	139	138	138	137	138
Practice 12	132	133	131	129	128	127	126
Practice 13	149	144	144	139	138	138	137
Practice 14	133	132	137	133	134	130	131
Practice 15	138	136	133	133	132	131	131

Table C2 Practice Lap Times (in seconds)

Stock Prices

The following table lists initial public offering (IPO) stock prices for all 1999 stocks that at least doubled in value during the first day of trading.

$17.00	$23.00	$14.00	$16.00	$12.00	$26.00
$20.00	$22.00	$14.00	$15.00	$22.00	$18.00
$18.00	$21.00	$21.00	$19.00	$15.00	$21.00
$18.00	$17.00	$15.00	$25.00	$14.00	$30.00
$16.00	$10.00	$20.00	$12.00	$16.00	$17.44
$16.00	$14.00	$15.00	$20.00	$20.00	$16.00
$17.00	$16.00	$15.00	$15.00	$19.00	$48.00

Table C3 IPO Offer Prices

$16.00	$18.00	$9.00	$18.00	$18.00	$20.00
$8.00	$20.00	$17.00	$14.00	$11.00	$16.00
$19.00	$15.00	$21.00	$12.00	$8.00	$16.00
$13.00	$14.00	$15.00	$14.00	$13.41	$28.00
$21.00	$17.00	$28.00	$17.00	$19.00	$16.00
$17.00	$19.00	$18.00	$17.00	$15.00	
$14.00	$21.00	$12.00	$18.00	$24.00	
$15.00	$23.00	$14.00	$16.00	$12.00	
$24.00	$20.00	$14.00	$14.00	$15.00	
$14.00	$19.00	$16.00	$38.00	$20.00	
$24.00	$16.00	$8.00	$18.00	$17.00	
$16.00	$15.00	$7.00	$19.00	$12.00	
$8.00	$23.00	$12.00	$18.00	$20.00	
$21.00	$34.00	$16.00	$26.00	$14.00	

Table C3 IPO Offer Prices

References

Data compiled by Jay R. Ritter of University of Florida using data from *Securities Data Co.* and *Bloomberg*.

D GROUP AND PARTNER PROJECTS

Univariate Data

Student Learning Objectives

- The student will design and carry out a survey.
- The student will analyze and graphically display the results of the survey.

Instructions

As you complete each task below, check it off. Answer all questions in your summary.
____ Decide what data you are going to study.

> Here are two examples, but you may **NOT** use them: number of M&M's per bag, number of pencils students have in their backpacks.

____ Are your data discrete or continuous? How do you know?
____ Decide how you are going to collect the data (for instance, buy 30 bags of M&M's; collect data from the World Wide Web).
____ Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random (using a random number generator) sampling. Do not use convenience sampling. Which method did you use? Why did you pick that method?
____ Conduct your survey. **Your data size must be at least 30.**
____ Summarize your data in a chart with columns showing **data value, frequency, relative frequency and cumulative relative frequency.**
Answer the following (rounded to two decimal places):

a. $\overline{x}$ = _____
b. s = _____
c. First quartile = _____
d. Median = _____
e. 70^{th} percentile = _____

____ What value is 1.5 standard deviations below the mean?
____ Construct a histogram displaying your data.
____ In complete sentences, describe the shape of your graph.
____ Do you notice any potential outliers? If so, what values are they? Show your work in how you used the potential outlier formula to determine whether or not the values might be outliers.
____ Construct a box plot displaying your data.
____ Does the middle 50% of the data appear to be concentrated together or spread apart? Explain how you determined this.
____ Looking at both the histogram and the box plot, discuss the distribution of your data.

Assignment Checklist

You need to turn in the following typed and stapled packet, with pages in the following order:

Cover sheet: name, class time, and name of your study
Summary page: This should contain paragraphs written with complete sentences. It should include answers to all the questions above. It should also include statements describing the population under study, the sample, a parameter or parameters being studied, and the statistic or statistics produced.
URL for data, if your data are from the World Wide Web
Chart of data, frequency, relative frequency, and cumulative relative frequency
Page(s) of graphs: histogram and box plot

Continuous Distributions and Central Limit Theorem

Student Learning Objectives

- The student will collect a sample of continuous data.

- The student will attempt to fit the data sample to various distribution models.
- The student will validate the central limit theorem.

Instructions

As you complete each task below, check it off. Answer all questions in your summary.

Part I: Sampling

____ Decide what **continuous** data you are going to study. (Here are two examples, but you may NOT use them: the amount of money a student spent on college supplies this term, or the length of time distance telephone call lasts.)
____ Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random (using a random number generator) sampling. Do not use convenience sampling. What method did you use? Why did you pick that method?
____ Conduct your survey. Gather **at least 150 pieces of continuous, quantitative data**.
____ Define (in words) the random variable for your data. X = _____
____ Create two lists of your data: (1) unordered data, (2) in order of smallest to largest.
____ Find the sample mean and the sample standard deviation (rounded to two decimal places).

a. $\overline{x}$ = _____
b. s = _____

Part II: Possible Distributions

____ Suppose that X followed the following theoretical distributions. Set up each distribution using the appropriate information from your data.
____ Uniform: $X \sim U$ _____ Use the lowest and highest values as a and b.
____ Normal: $X \sim N$ _____ Use $\overline{x}$ to estimate for μ and s to estimate for σ.
____ **Must** your data fit one of the above distributions? Explain why or why not.
____ **Could** the data fit two or three of the previous distributions (at the same time)? Explain.
____ Calculate the value k(an X value) that is 1.75 standard deviations above the sample mean. k = _____ (rounded to two decimal places) Note: $k = \overline{x} + (1.75)s$
____ Determine the relative frequencies (RF) rounded to four decimal places.

Note

$$RF = \frac{\text{frequency}}{\text{total number surveyed}}$$

a. $RF(X < k)$ = _____
b. $RF(X > k)$ = _____
c. $RF(X = k)$ = _____

Note

You should have one page for the uniform distribution, one page for the exponential distribution, and one page for the normal distribution.

____ State the distribution: $X \sim$ _____
____ Draw a graph for each of the three theoretical distributions. Label the axes and mark them appropriately.
____ Find the following theoretical probabilities (rounded to four decimal places).

a. $P(X < k)$ = _____
b. $P(X > k)$ = _____
c. $P(X = k)$ = _____

____ Compare the relative frequencies to the corresponding probabilities. Are the values close?
____ Does it appear that the data fit the distribution well? Justify your answer by comparing the probabilities to the relative frequencies, and the histograms to the theoretical graphs.

Part III: CLT Experiments

_____ From your original data (before ordering), use a random number generator to pick 40 samples of size five. For each sample, calculate the average.
_____ On a separate page, attached to the summary, include the 40 samples of size five, along with the 40 sample averages.
_____ List the 40 averages in order from smallest to largest.
_____ Define the random variable, $\overline{X}$, in words. $\overline{X}$ = _____
_____ State the approximate theoretical distribution of $\overline{X}$. $\overline{X}$ ~ _____
_____ Base this on the mean and standard deviation from your original data.
_____ Construct a histogram displaying your data. Use five to six intervals of equal width. Label and scale it.
Calculate the value $\overline{k}$ (an $\overline{X}$ value) that is 1.75 standard deviations above the sample mean. $\overline{k}$ = _____ (rounded to two decimal places)
Determine the relative frequencies (*RF*) rounded to four decimal places.

a. $RF(\overline{X} < \overline{k})$ = _____
b. $RF(\overline{X} > \overline{k})$ = _____
c. $RF(\overline{X} = \overline{k})$ = _____

a. $P(\overline{X} < \overline{k})$ = _____
b. $P(\overline{X} > \overline{k})$ = _____
c. $P(\overline{X} = \overline{k})$ = _____

_____ Draw the graph of the theoretical distribution of X.
_____ Compare the relative frequencies to the probabilities. Are the values close?
_____ Does it appear that the data of averages fit the distribution of $\overline{X}$ well? Justify your answer by comparing the probabilities to the relative frequencies, and the histogram to the theoretical graph.
In three to five complete sentences for each, answer the following questions. Give thoughtful explanations.
_____ In summary, do your original data seem to fit the uniform, exponential, or normal distributions? Answer why or why not for each distribution. If the data do not fit any of those distributions, explain why.
_____ What happened to the shape and distribution when you averaged your data? **In theory,** what should have happened? In theory, would "it" always happen? Why or why not?
_____ Were the relative frequencies compared to the theoretical probabilities closer when comparing the X or $\overline{X}$ distributions? Explain your answer.

Assignment Checklist

You need to turn in the following typed and stapled packet, with pages in the following order:
____ **Cover sheet**: name, class time, and name of your study
____ **Summary pages**: These should contain several paragraphs written with complete sentences that describe the experiment, including what you studied and your sampling technique, as well as answers to all of the questions previously asked questions
____ **URL** for data, if your data are from the World Wide Web
____ **Pages, one for each theoretical distribution**, with the distribution stated, the graph, and the probability questions answered
____ **Pages of the data requested**
____ **All graphs required**

Hypothesis Testing-Article

Student Learning Objectives

- The student will identify a hypothesis testing problem in print.
- The student will conduct a survey to verify or dispute the results of the hypothesis test.
- The student will summarize the article, analysis, and conclusions in a report.

Instructions

As you complete each task, check it off. Answer all questions in your summary.
____**Find an article** in a newspaper, magazine, or on the internet which makes a claim about **ONE** population mean or **ONE** population proportion. The claim may be based upon a survey that the article was reporting on. Decide whether this claim is the null or alternate hypothesis.
____**Copy or print out the article** and include a copy in your project, along with the source.

____**State how you will collect your data.** (Convenience sampling is not acceptable.)
____**Conduct your survey. You must have more than 50 responses in your sample.** When you hand in your final project, attach the tally sheet or the packet of questionnaires that you used to collect data. Your data must be real.
____**State the statistics** that are a result of your data collection: sample size, sample mean, and sample standard deviation, OR sample size and number of successes.
____**Make two copies of the appropriate solution sheet.**
____**Record the hypothesis test** on the solution sheet, based on your experiment. **Do a DRAFT solution** first on one of the solution sheets and check it over carefully. Have a classmate check your solution to see if it is done correctly. Make your decision using a 5% level of significance. Include the 95% confidence interval on the solution sheet.
____**Create a graph that illustrates your data.** This may be a pie or bar graph or may be a histogram or box plot, depending on the nature of your data. Produce a graph that makes sense for your data and gives useful visual information about your data. You may need to look at several types of graphs before you decide which is the most appropriate for the type of data in your project.
____**Write your summary** (in complete sentences and paragraphs, with proper grammar and correct spelling) that describes the project. The summary **MUST** include:

a. Brief discussion of the article, including the source
b. Statement of the claim made in the article (one of the hypotheses).
c. Detailed description of how, where, and when you collected the data, including the sampling technique; did you use cluster, stratified, systematic, or simple random sampling (using a random number generator)? As previously mentioned, convenience sampling is not acceptable.
d. Conclusion about the article claim in light of your hypothesis test; this is the conclusion of your hypothesis test, stated in words, in the context of the situation in your project in sentence form, as if you were writing this conclusion for a non-statistician.
e. Sentence interpreting your confidence interval in the context of the situation in your project

Assignment Checklist

Turn in the following typed (12 point) and stapled packet for your final project:
____**Cover sheet** containing your name(s), class time, and the name of your study
____**Summary**, which includes all items listed on summary checklist
____**Solution sheet** neatly and completely filled out. The solution sheet does not need to be typed.
____**Graphic representation of your data**, created following the guidelines previously discussed; include only graphs which are appropriate and useful.
____**Raw data collected AND a table summarizing the sample data** (n, $\overline{x}$ and s; or x, n, and p', as appropriate for your hypotheses); the raw data does not need to be typed, but the summary does. Hand in the data as you collected it. (Either attach your tally sheet or an envelope containing your questionnaires.)

Bivariate Data, Linear Regression, and Univariate Data

Student Learning Objectives

- The students will collect a bivariate data sample through the use of appropriate sampling techniques.
- The student will attempt to fit the data to a linear model.
- The student will determine the appropriateness of linear fit of the model.
- The student will analyze and graph univariate data.

Instructions

1. As you complete each task below, check it off. Answer all questions in your introduction or summary.
2. Check your course calendar for intermediate and final due dates.
3. Graphs may be constructed by hand or by computer, unless your instructor informs you otherwise. All graphs must be neat and accurate.
4. All other responses must be done on the computer.
5. Neatness and quality of explanations are used to determine your final grade.

Part I: Bivariate Data

Introduction
____State the bivariate data your group is going to study.

Here are two examples, but you may **NOT** use them: height vs. weight and age vs. running distance.

____Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random sampling (using a random number generator) sampling. Convenience sampling is **NOT** acceptable.
____Conduct your survey. Your number of pairs must be at least 30.
____Print out a copy of your data.

Analysis
____On a separate sheet of paper construct a scatter plot of the data. Label and scale both axes.
____State the least squares line and the correlation coefficient.
____On your scatter plot, in a different color, construct the least squares line.
____Is the correlation coefficient significant? Explain and show how you determined this.
____Interpret the slope of the linear regression line in the context of the data in your project. Relate the explanation to your data, and quantify what the slope tells you.
____Does the regression line seem to fit the data? Why or why not? If the data does not seem to be linear, explain if any other model seems to fit the data better.
____Are there any outliers? If so, what are they? Show your work in how you used the potential outlier formula in the Linear Regression and Correlation chapter (since you have bivariate data) to determine whether or not any pairs might be outliers.

Part II: Univariate Data

In this section, you will use the data for **ONE** variable only. Pick the variable that is more interesting to analyze. For example: if your independent variable is sequential data such as year with 30 years and one piece of data per year, your *x*-values might be 1971, 1972, 1973, 1974, ..., 2000. This would not be interesting to analyze. In that case, choose to use the dependent variable to analyze for this part of the project.
_____Summarize your data in a chart with columns showing data value, frequency, relative frequency, and cumulative relative frequency.
_____Answer the following question, rounded to two decimal places:

a. Sample mean = _____
b. Sample standard deviation = _____
c. First quartile = _____
d. Third quartile = _____
e. Median = _____
f. 70th percentile = _____
g. Value that is 2 standard deviations above the mean = _____
h. Value that is 1.5 standard deviations below the mean = _____

_____Construct a histogram displaying your data. Group your data into six to ten intervals of equal width. Pick regularly spaced intervals that make sense in relation to your data. For example, do NOT group data by age as 20-26,27-33,34-40,41-47,48-54,55-61 . . . Instead, maybe use age groups 19.5-24.5, 24.5-29.5, . . . or 19.5-29.5, 29.5-39.5, 39.5-49.5, . . .
_____In complete sentences, describe the shape of your histogram.
_____Are there any potential outliers? Which values are they? Show your work and calculations as to how you used the potential outlier formula in Descriptive Statistics (since you are now using univariate data) to determine which values might be outliers.
_____Construct a box plot of your data.
_____Does the middle 50% of your data appear to be concentrated together or spread out? Explain how you determined this.
_____Looking at both the histogram AND the box plot, discuss the distribution of your data. For example: how does the spread of the middle 50% of your data compare to the spread of the rest of the data represented in the box plot; how does this correspond to your description of the shape of the histogram; how does the graphical display show any outliers you may have found; does the histogram show any gaps in the data that are not visible in the box plot; are there any interesting features of your data that you should point out.

Due Dates

- Part I, Intro: _____ (keep a copy for your records)
- Part I, Analysis: _____ (keep a copy for your records)
- Entire Project, typed and stapled: _____

  ____ Cover sheet: names, class time, and name of your study

___ Part I: label the sections â€œIntroâ€◈ and â€œAnalysis.â€◈

___ Part II:

___ Summary page containing several paragraphs written in complete sentences describing the experiment, including what you studied and how you collected your data. The summary page should also include answers to ALL the questions asked above.

___ All graphs requested in the project

___ All calculations requested to support questions in data

___ Description: what you learned by doing this project, what challenges you had, how you overcame the challenges

Note

Include answers to ALL questions asked, even if not explicitly repeated in the items above.

E SOLUTION SHEETS

Hypothesis Testing With One Sample

Class Time: _____
Name: _____

a. H_0: _____
b. H_a: _____
c. In words, clearly state what your random variable $\overline{X}$ or P' represents.
d. State the distribution to use for the test.
e. What is the test statistic?
f. What is the *p*-value? In one or two *complete sentences*, explain what the *p*-value means for this problem.
g. Use the previous information to sketch a picture of this situation. *Clearly*, label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.

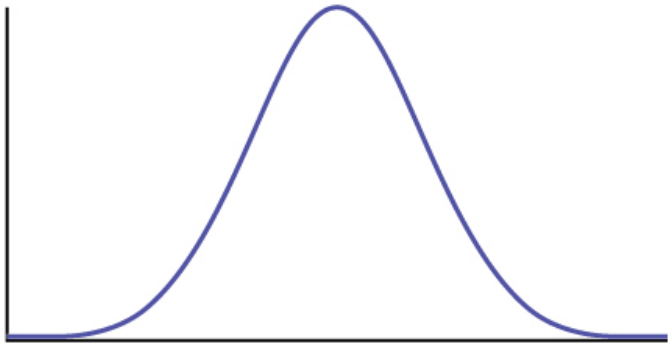


Figure E1

h. Indicate the correct decision (*reject* or *do not reject* the null hypothesis), the reason for it and write appropriate conclusions using *complete sentences.*"
   i. Alpha: _____
   ii. Decision: _____
   iii. Reason for decision: _____
   iv. Conclusion: _____

i. Construct a 95 percent confidence interval for the true mean or proportion. Sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the confidence interval.

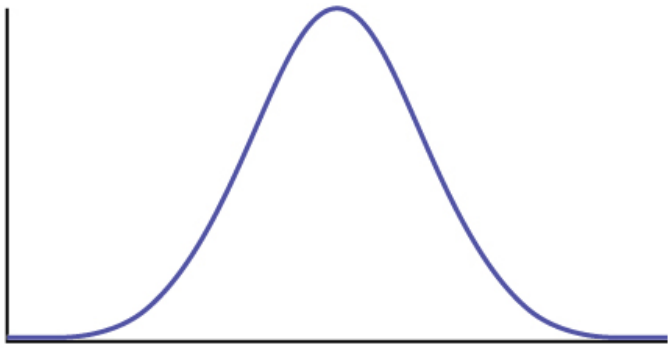


Figure E2

Hypothesis Testing With Two Samples

Class Time: _____
Name: _____

a. H_0: _____
b. H_a: _____

c. In words, *clearly* state what your random variable $\overline{X}_1 - \overline{X}_2$, $P'_1 - P'_2$ or $\overline{X}_d$ represents.
d. State the distribution to use for the test.
e. What is the test statistic?
f. What is the *p*-value? In one to two *complete sentences,* explain what the *p*-value means for this problem.
g. Use the previous information to sketch a picture of this situation. *Clearly* label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.

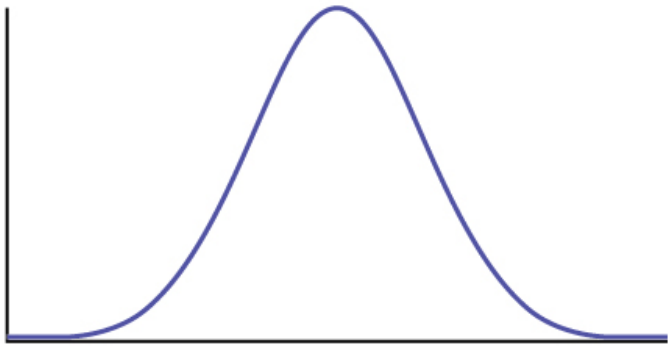


Figure E3

h. Indicate the correct decision (*reject* or *do not reject* the null hypothesis), and write appropriate conclusions using *complete sentences.*
   i. Alpha: _____
   ii. Decision: _____
   iii. Reason for decision: _____
   iv. Conclusion: _____

i. In *complete sentences,* explain how you determined which distribution to use.

The Chi-Square Distribution

Class Time: _____
Name: _____

a. H_0: _____
b. H_a: _____
c. What are the degrees of freedom?
d. State the distribution to use for the test.
e. What is the test statistic?
f. What is the *p*-value? In one to two *complete sentences,* explain what the *p*-value means for this problem.
g. Use the previous information to sketch a picture of this situation. *Clearly* label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.

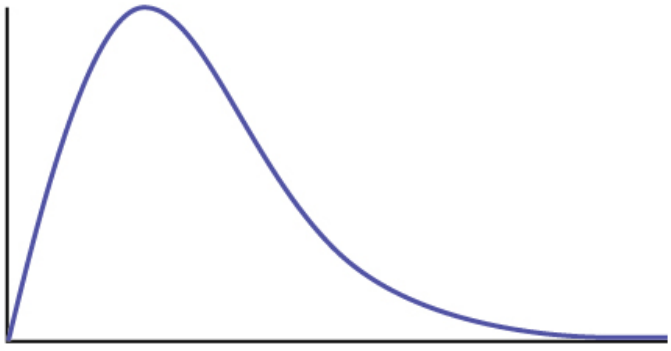


Figure E4

h. Indicate the correct decision (*reject* or *do not reject* the null hypothesis) and write appropriate conclusions, using *complete sentences.*
   i. Alpha: _____

ii. Decision: _____
iii. Reason for decision: _____
iv. Conclusion: _____

F Distribution and One-Way ANOVA

Class Time: _____
Name: _____

a. H_0: _____
b. H_a: _____
c. $df(n)$ = _____ $df(d)$ = _____
d. State the distribution to use for the test.
e. What is the test statistic?
f. What is the *p*-value?
g. Use the previous information to sketch a picture of this situation. *Clearly* label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.

Figure E5

h. Indicate the correct decision (*reject* or *do not reject* the null hypothesis) and write appropriate conclusions, using *complete sentences*.
   a. Alpha: _____
   b. Decision: _____
   c. Reason for decision: _____
   d. Conclusion: _____

F MATHEMATICAL PHRASES, SYMBOLS, AND FORMULAS

English Phrases Written Mathematically

When the English says:	Interpret this as:
X is at least 4.	$X \geq 4$
The minimum of X is 4.	$X \geq 4$
X is no less than 4.	$X \geq 4$
X is greater than or equal to 4.	$X \geq 4$
X is at most 4.	$X \leq 4$
The maximum of X is 4.	$X \leq 4$
X is no more than 4.	$X \leq 4$
X is less than or equal to 4.	$X \leq 4$
X does not exceed 4.	$X \leq 4$
X is greater than 4.	$X > 4$
X is more than 4.	$X > 4$
X exceeds 4.	$X > 4$
X is less than 4.	$X < 4$
There are fewer X than 4.	$X < 4$
X is 4.	$X = 4$
X is equal to 4.	$X = 4$
X is the same as 4.	$X = 4$
X is not 4.	$X \neq 4$
X is not equal to 4.	$X \neq 4$
X is not the same as 4.	$X \neq 4$
X is different than 4.	$X \neq 4$

Table F1

Formulas

Formula 1: Factorial

$n! = n(n-1)(n-2)\ldots(1)$

$0! = 1$

Formula 2: Combinations

$\binom{n}{r} = \frac{n!}{(n-r)!r!}$

Formula 3: Binomial Distribution

$X \sim B(n, p)$

$P(X = x) = \binom{n}{x} p^x q^{n-x}$, for $x = 0, 1, 2, \ldots, n$

Formula 4: Geometric Distribution

$X \sim G(p)$

$P(X = x) = q^{x-1} p$, for $x = 1, 2, 3, \ldots$

Formula 5: Hypergeometric Distribution

$X \sim H(r, b, n)$

$P(X = x) = \left(\frac{\binom{r}{x}\binom{b}{n-x}}{\binom{r+b}{n}} \right)$

Formula 6: Poisson Distribution

$X \sim P(\mu)$

$P(X = x) = \frac{\mu^x e^{-\mu}}{x!}$

Formula 7: Uniform Distribution

$X \sim U(a, b)$

$f(X) = \frac{1}{b-a}$, $a < x < b$

Formula 8: Exponential Distribution

$X \sim Exp(m)$

$f(x) = me^{-mx}\, m > 0, x \geq 0$

Formula 9: Normal Distribution

$X \sim N(\mu, \sigma^2)$

$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$

Formula 10: Gamma Function

$\Gamma(z) = \int_{\infty}^{0} x^{z-1} e^{-x} dx \; z > 0$

$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

$\Gamma(m+1) = m!$ for m, a nonnegative integer

otherwise: $\Gamma(a+1) = a\Gamma(a)$

Formula 11: Student's *t*-distribution

$X \sim t_{df}$

$$f(x) = \frac{\left(1+\frac{x^2}{n}\right)^{\frac{-(n+1)}{2}} \Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}$$

$$X = \frac{Z}{\sqrt{\frac{Y}{n}}}$$

$Z \sim N(0,1)$, $Y \sim X^2_{df}$, n = degrees of freedom

Formula 12: Chi-Square Distribution

$X \sim X^2_{df}$

$f(x) = \frac{x^{\frac{n-2}{2}} e^{\frac{-x}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}$, $x > 0$, n = positive integer and degrees of freedom

Formula 13: F Distribution

$X \sim F_{df(n),df(d)}$

$df(n)$ =degrees of freedom for the numerator

$df(d)$ =degrees of freedom for the denominator

$$f(x) = \frac{\Gamma(\frac{u+v}{2})}{\Gamma(\frac{u}{2})\Gamma(\frac{v}{2})}(\frac{u}{v})^{\frac{u}{2}} x^{(\frac{u}{2}-1)}[1+(\frac{u}{v})x^{-0.5(u+v)}]$$

$X = \frac{Y_u}{W_v}$, Y, W are chi-square

Symbols and Their Meanings

Chapter (1st used)	Symbol	Spoken	Meaning
Sampling and Data	$\sqrt{\ }$	The square root of	same
Sampling and Data	π	Pi	3.14159... (a specific number)
Descriptive Statistics	Q_1	Quartile one	the first quartile
Descriptive Statistics	Q_2	Quartile two	the second quartile
Descriptive Statistics	Q_3	Quartile three	the third quartile
Descriptive Statistics	IQR	interquartile range	$Q_3 - Q_1 = IQR$
Descriptive Statistics	$\overline{x}$	x-bar	sample mean

Table F2 Symbols and their Meanings

Chapter (1st used)	Symbol	Spoken	Meaning
Descriptive Statistics	μ	mu	population mean
Descriptive Statistics	**s** s_x *sx*	s	sample standard deviation
Descriptive Statistics	s^2 s_x^2	s squared	sample variance
Descriptive Statistics	σ σ_x σx	sigma	population standard deviation
Descriptive Statistics	σ^2 σ_x^2	sigma squared	population variance
Descriptive Statistics	Σ	capital sigma	sum
Probability Topics	{ }	brackets	set notation
Probability Topics	S	S	sample space
Probability Topics	A	Event A	event A
Probability Topics	$P(A)$	probability of A	probability of A occurring
Probability Topics	$P(A\|B)$	probability of A given B	prob. of A occurring given B has occurred
Probability Topics	$P(A \text{ OR } B)$	prob. of A or B	prob. of A or B or both occurring
Probability Topics	$P(A \text{ AND } B)$	prob. of A and B	prob. of both A and B occurring (same time)
Probability Topics	A'	A-prime, complement of A	complement of A, not A
Probability Topics	$P(A')$	prob. of complement of A	same
Probability Topics	G_1	green on first pick	same
Probability Topics	$P(G_1)$	prob. of green on first pick	same
Discrete Random Variables	*PDF*	prob. distribution function	same
Discrete Random Variables	*X*	X	the random variable X
Discrete Random Variables	*X* ~	the distribution of X	same
Discrete Random Variables	*B*	binomial distribution	same
Discrete Random Variables	*G*	geometric distribution	same
Discrete Random Variables	*H*	hypergeometric dist.	same
Discrete Random Variables	*P*	Poisson dist.	same

Table F2 Symbols and their Meanings

Chapter (1st used)	Symbol	Spoken	Meaning
Discrete Random Variables	λ	Lambda	average of Poisson distribution
Discrete Random Variables	$\geq$	greater than or equal to	same
Discrete Random Variables	$\leq$	less than or equal to	same
Discrete Random Variables	=	equal to	same
Discrete Random Variables	$\neq$	not equal to	same
Continuous Random Variables	$f(x)$	f of x	function of x
Continuous Random Variables	*pdf*	prob. density function	same
Continuous Random Variables	*U*	uniform distribution	same
Continuous Random Variables	*Exp*	exponential distribution	same
Continuous Random Variables	*k*	*k*	critical value
Continuous Random Variables	$f(x) =$	f of x equals	same
Continuous Random Variables	*m*	*m*	decay rate (for exp. dist.)
The Normal Distribution	*N*	normal distribution	same
The Normal Distribution	*z*	*z*-score	same
The Normal Distribution	*Z*	standard normal dist.	same
The Central Limit Theorem	*CLT*	Central Limit Theorem	same
The Central Limit Theorem	$\overline{X}$	*X*-bar	the random variable *X*-bar
The Central Limit Theorem	μ_X	mean of *X*	the average of *X*
The Central Limit Theorem	$\mu_{\overline{X}}$	mean of *X*-bar	the average of *X*-bar
The Central Limit Theorem	σ_X	standard deviation of *X*	same
The Central Limit Theorem	$\sigma_{\overline{X}}$	standard deviation of *X*-bar	same
The Central Limit Theorem	ΣX	sum of *X*	same

Table F2 Symbols and their Meanings

Chapter (1st used)	Symbol	Spoken	Meaning
The Central Limit Theorem	Σx	sum of *x*	same
Confidence Intervals	*CL*	confidence level	same
Confidence Intervals	*CI*	confidence interval	same
Confidence Intervals	*EBM*	error bound for a mean	same
Confidence Intervals	*EBP*	error bound for a proportion	same
Confidence Intervals	t	Student's *t*-distribution	same
Confidence Intervals	*df*	degrees of freedom	same
Confidence Intervals	$t_{\frac{\alpha}{2}}$	student t with *a*/2 area in right tail	same
Confidence Intervals	p'; $\hat{p}$	*p*-prime; *p*-hat	sample proportion of success
Confidence Intervals	q'; $\hat{q}$	*q*-prime; *q*-hat	sample proportion of failure
Hypothesis Testing	H_0	*H*-naught, *H*-sub 0	null hypothesis
Hypothesis Testing	H_a	*H*-*a*, *H*-sub *a*	alternate hypothesis
Hypothesis Testing	H_1	*H*-1, *H*-sub 1	alternate hypothesis
Hypothesis Testing	α	alpha	probability of Type I error
Hypothesis Testing	β	beta	probability of Type II error
Hypothesis Testing	$\overline{X1} - \overline{X2}$	*X*1-bar minus *X*2-bar	difference in sample means
Hypothesis Testing	$\mu_1 - \mu_2$	*mu*-1 minus *mu*-2	difference in population means
Hypothesis Testing	$P'_1 - P'_2$	*P*1-prime minus *P*2-prime	difference in sample proportions
Hypothesis Testing	$p_1 - p_2$	*p*1 minus *p*2	difference in population proportions
Chi-Square Distribution	X^2	*Ky*-square	Chi-square
Chi-Square Distribution	O	Observed	Observed frequency
Chi-Square Distribution	E	Expected	Expected frequency
Linear Regression and Correlation	$y = a + bx$	*y* equals a plus *b*-*x*	equation of a line

Table F2 Symbols and their Meanings

Chapter (1st used)	Symbol	Spoken	Meaning
Linear Regression and Correlation	$\hat{y}$	*y*-hat	estimated value of *y*
Linear Regression and Correlation	r	correlation coefficient	same
Linear Regression and Correlation	ε	error	same
Linear Regression and Correlation	SSE	Sum of Squared Errors	same
Linear Regression and Correlation	$1.9s$	1.9 times *s*	cut-off value for outliers
F-Distribution and ANOVA	F	*F*-ratio	*F*-ratio

Table F2 Symbols and their Meanings

G NOTES FOR THE TI-83, 83+, 84, 84+ CALCULATORS

Quick Tips

Legend

- [ ] represents a button press
- [ ] represents yellow command or green letter behind a key
- < > represents items on the screen

To adjust the contrast

Press 2nd, then hold ▲ to increase the contrast or ▼ to decrease the contrast.

To capitalize letters and words

Press ALPHA to get one capital letter, or press 2nd, then ALPHA to set all button presses to capital letters.

You can return to the top-level button values by pressing ALPHA again.

To correct a mistake

If you hit a wrong button, press CLEAR and start again.

To write in scientific notation

Numbers in scientific notation are expressed on the TI-83, 83+, 84, and 84+ using E notation, such that...

- 4.321 E 4 = 4.321×10^4
- 4.321 E –4 = 4.321×10^{-4}

To transfer programs or equations from one calculator to another

Both calculators: Insert your respective end of the link cable cable and press 2nd, then [LINK].

Calculator receiving information

1. Use the arrows to navigate to and select <RECEIVE>.
2. Press ENTER.

Calculator sending information

1. Press the appropriate number or letter.
2. Use the up and down arrows to access the appropriate item.
3. Press ENTER to select the item to transfer.
4. Press the right arrow to navigate to and select <TRANSMIT>.
5. Press ENTER.

Note

ERROR 35 LINK generally means that the cables have not been inserted far enough.

Both calculators—Insert your respective end of the link cable, press 2nd, then [QUIT] to exit when done.

Manipulating One-Variable Statistics

Note

These directions are for entering data using the built-in statistical program.

Data	Frequency
–2	10
–1	3
0	4
1	5
3	8

Table G1 Sample Data We are manipulating one-variable statistics.

To begin

1. Turn on the calculator.
   ON
2. Access statistics mode.
   STAT
3. Select <4:ClrList> to clear data from lists, if desired.
   4, then ENTER.
4. Enter the list [L1] to be cleared.
   2nd, [L1], ENTER.
5. Display the last instruction.
   2nd, [ENTRY].
6. Continue clearing any remaining lists in the same fashion, if desired.
   ◀, 2nd, [L2], ENTER
7. Access statistics mode.
   STAT
8. Select <1:Edit . . .>.
   ENTER
9. Enter data. Data values go into [L1]. (You may need to arrow over to [L1]).
   - Type in a data value and enter it. For negative numbers, use the negate – key at the bottom of the keypad.
     (–), 9, ENTER.
   - Continue in the same manner until all data values are entered.
10. In [L2], enter the frequencies for each data value in [L1].
    - Type in a frequency and enter it. If a data value appears only once, the frequency is *1*.
      4, ENTER.
    - Continue in the same manner until all data values are entered.

11. Access statistics mode.
    STAT
12. Navigate to <CALC>.
13. Access <1:1-var Stats>.
    ENTER
14. Indicate that the data is in [L1]...
    2nd , [L1] , , ,
15. ...and indicate that the frequencies are in [L2].
    2nd , [L2] , ENTER .
16. The statistics should be displayed. You may arrow down to get remaining statistics. Repeat as necessary.

Drawing Histograms

Note

We will assume that the data are already entered.

We will construct two histograms with the built-in [STAT PLOT] application. In the first method, we will use the default ZOOM. The second method will involve customizing a new graph.

1. Access graphing mode.
   2nd , [STAT PLOT].
2. Select <1:plot 1> to access plotting - first graph.
   ENTER
3. Use the arrows to navigate to <ON> to turn on Plot 1.
   <ON> , ENTER .
4. Use the arrows to go to the histogram picture and select the histogram. ENTER
5. Use the arrows to navigate to <Xlist>.
6. If [L1] is not selected, select it.
   2nd , [L1] , ENTER .
7. Use the arrows to navigate to <Freq>.
8. Assign the frequencies to [L2].
   2nd , [L2] , ENTER .
9. Go back to access other graphs.
   2nd , [STAT PLOT].
10. Use the arrows to turn off the remaining plots.
11. **Be sure to deselect or clear all equations before graphing.**

To deselect equations

1. Access the list of equations.
   Y=
2. Select each equal sign (=).
   ▼ ▶ ENTER .

3. Continue until all equations are deselected.

To clear equations

1. Access the list of equations.
   Y=
2. Use the arrow keys to navigate to the right of each equal sign (=) and clear them.
   ▼ ▶ CLEAR .
3. Repeat until all equations are deleted.

To draw default histogram

1. Access the ZOOM menu.
   ZOOM
2. Select <9:ZoomStat>.
   9
3. The histogram will display with a window automatically set.

To draw a custom histogram

1. Access window mode to set the graph parameters.
   WINDOW
2. - $X_{\min} = -2.5$
   - $X_{\max} = 3.5$
   - $X_{scl} = 1$ (width of bars)
   - $Y_{\min} = 0$
   - $Y_{\max} = 10$
   - $Y_{scl} = 1$ (spacing of tick marks on *y*-axis)
   - $X_{res} = 1$
3. Access graphing mode to see the histogram.
   GRAPH

To draw box plots

1. Access graphing mode.
   2nd , [STAT PLOT].
2. Select <1:Plot 1> to access the first graph.
   ENTER
3. Use the arrows to select <ON> and turn on Plot 1.
   ENTER
4. Use the arrows to select the box plot picture and enable it.
   ENTER
5. Use the arrows to navigate to <Xlist>.
6. If [L1] is not selected, select it.
   2nd , [L1] , ENTER .
7. Use the arrows to navigate to <Freq>.
8. Indicate that the frequencies are in [L2].
   2nd , [L2] , ENTER .

9. Go back to access other graphs.
   2nd, [STAT PLOT].
10. **Be sure to deselect or clear all equations before graphing** using the method mentioned above.
11. View the box plot.
    GRAPH, [STAT PLOT].

Linear Regression

Sample Data

The following data are real. The percent of declared ethnic minority students at De Anza College for selected years from 1970–1995 is indicated in the following table.

Year	Student Ethnic Minority Percentage
1970	14.13%
1973	12.27%
1976	14.08%
1979	18.16%
1982	27.64%
1983	28.72%
1986	31.86%
1989	33.14%
1992	45.37%
1995	53.1%

Table G2 The independent variable is *Year,* while the independent variable is *Student Ethnic Minority Percentage*.

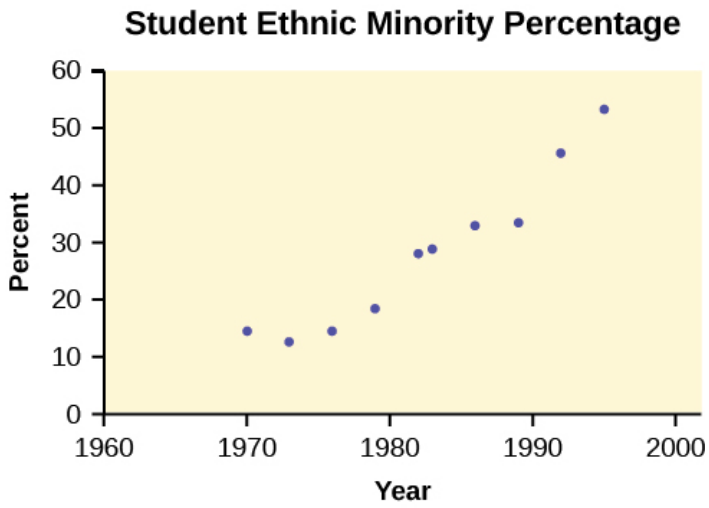


Figure G1 Student Ethnic Minority Percentage By hand, verify the scatterplot above.

Note

The TI-83 has a built-in linear regression feature, which allows the data to be edited. The *x*-values will be in [L1]; the *y*-values in [L2].

To enter data and perform linear regression

1. ON Turns calculator on.
   ON
2. Before accessing this program, be sure to turn off all plots.
   - Access graphing mode.
     2nd, [STAT PLOT].
   - Turn off all plots.
     4, ENTER.
3. Round to three decimal places.
   - Access the mode menu.
     MODE, [STAT PLOT].
   - Navigate to <Float> and then to the right until you reach <3>.
     ▼ ▶.
   - All numbers will be rounded to three decimal places until changed.
     ENTER
4. Enter statistics mode and clear lists [L1] and [L2], as described previously.
   STAT, 4.
5. Enter editing mode to insert values for *x* and *y*.
   STAT, ENTER.
6. Enter each value. Press ENTER to continue.

To display the correlation coefficient

1. Access the catalog.
   2nd, [CATALOG].
2. Arrow down and select <DiagnosticOn>.
   ▼ ..., ENTER, ENTER.
3. *r* and r^2 will be displayed during regression calculations.
4. Access linear regression.
   STAT ▶.
5. Select the form of $y = a + bx$.
   8, ENTER.

The display will show the following information

LinReg

- $y = a + bx$

- $a = -3176.909$
- $b = 1.617$
- $r^2 = 0.924$
- $r = 0.961$

This means the Line of Best Fit (Least Squares Line) is:

- $y = -3176.909 + 1.617x$
- % = –3176.909 + 1.617 (year #)

The correlation coefficient is $r = 0.961$.

To see the scatter plot

1. Access graphing mode.
   2nd, [STAT PLOT].
2. Select <1:Plot 1> To access plotting - first graph.
   ENTER
3. Navigate and select <ON> to turn on <1:Plot 1>.
   <ON> ENTER.
4. Navigate to the first picture.
5. Select the scatter plot.
   ENTER
6. Navigate to <Xlist>.
7. If [L1] is not selected, press 2nd, then [L1] to select it.
8. Confirm that the data values are in [L1].
   <ON>, ENTER.
9. Navigate to <Ylist>.
10. Select that the frequencies are in [L2].
    2nd, [L2], ENTER
11. Go back to access other graphs.
    2nd, [STAT PLOT]
12. Use the arrows to turn off the remaining plots.
13. Access window mode to set the graph parameters.
    WINDOW
    - $X_{\min} = 1970$
    - $X_{\max} = 2000$
    - $X_{scl} = 10$ (spacing of tick marks on *x*-axis)
    - $Y_{\min} = -0.05$
    - $Y_{\max} = 60$
    - $Y_{scl} = 10$ (spacing of tick marks on *y*-axis)
    - $X_{res} = 1$
14. Be sure to deselect or clear all equations before graphing, using the instructions above.
15. Press the graph button to see the scatter plot. GRAPH

To see the regression graph

1. Access the equation menu. The regression equation will be put into Y1.
   Y=

2. Access the vars menu and navigate to <5: Statistics>.
   VARS, 5.
3. Navigate to <EQ>.
4. <1: RegEQ> contains the regression equation which will be entered in Y1.
   ENTER
5. Press the graphing mode button. The regression line will be superimposed over the scatter plot.
   GRAPH

To see the residuals and use them to calculate the critical point for an outlier

1. Access the list. <RESID> will be an item on the menu. Navigate to it.
   2nd, [LIST], then <RESID>.
2. Press enter twice to view the list of residuals. Use the arrows to select them.
   ENTER, ENTER.
3. The critical point for an outlier is $1.9V\frac{\text{SSE}}{n-2}$, where
   - n = number of pairs of data
   - SSE = sum of the squared errors
   - $\sum \text{residual}^2$
4. Store the residuals in [L3].
   STO▶, 2nd, [L3], ENTER.
5. Calculate the $\frac{(\text{residual})^2}{n-2}$. Note that $n-2=8$.
   2nd, [L3], x^2, ÷, then 8.
6. Store this value in [L4].
   STO▶, 2nd, [L4], ENTER.
7. Calculate the critical value using the equation above.
   1, ., 9, ×, 2nd, [V], 2nd, [LIST], ▶, ▶, 5, 2nd, [L4],),), then ENTER.
8. Verify that the calculator displays 7.642669563. This is the critical value.
9. Compare the absolute value of each residual value in [L3] to 7.64. If the absolute value is greater than 7.64, then the (*x*, *y*) corresponding point is an outlier. In this case, none of the points is an outlier.

To obtain estimates of *y* for various *x*-values

There are various ways to determine estimates for "*y*." One way is to substitute values for "*x*" in the equation. Another way is to use the TRACE on the graph of the regression line.

TI-83, 83+, 84, 84+ instructions for distributions and tests

Distributions

Access DISTR for *Distributions*.

For technical assistance, visit the Texas Instruments website at http://www.ti.com and enter your calculator model into the *search* box.

Binomial Distribution

- binompdf(*n*, *p*, *x*) corresponds to $P(X = x)$
- binomcdf(*n*, *p*, *x*) corresponds to $P(X \leq x)$
- To see a list of all probabilities for *x*: 0, 1, . . . , *n*, leave off the "*x*" parameter.

Poisson Distribution

- `poissonpdf(λ,x)` corresponds to $P(X = x)$
- `poissoncdf(λ,x)` corresponds to $P(X \leq x)$

Continuous Distributions (general)

- $-\infty$ uses the value –1EE99 for left bound
- ∞ uses the value 1EE99 for right bound

Normal Distribution

- `normalpdf(x,μ,σ)` yields a probability density function value, only useful to plot the normal curve, in which case "x" is the variable
- `normalcdf(left bound, right bound, μ, σ)` corresponds to P(left bound < X < right bound)
- `normalcdf(left bound, right bound)` corresponds to P(left bound < Z < right bound) – standard normal
- `invNorm(p,μ,σ)` yields the critical value, k: $P(X < k) = p$
- `invNorm(p)` yields the critical value, k: $P(Z < k) = p$ for the standard normal

Student's *t*-Distribution

- `tpdf(x,df)` yields the probability density function value, only useful to plot the student-t curve, in which case "x" is the variable)
- `tcdf(left bound, right bound, df)` corresponds to P(left bound < t < right bound)

Chi-square Distribution

- X^2`pdf(x,df)` yields the probability density function value, only useful to plot the chi^2 curve, in which case "x" is the variable
- X^2`cdf(left bound, right bound, df)` corresponds to P(left bound < X^2 < right bound)

F Distribution

- `Fpdf(x,dfnum,dfdenom)` yields the probability density function value, only useful to plot the F curve, in which case "x" is the variable
- `Fcdf(left bound,right bound,dfnum,dfdenom)` corresponds to P(left bound < F < right bound)

Tests and Confidence Intervals

Access STAT and TESTS.

For the confidence intervals and hypothesis tests, you may enter the data into the appropriate lists and press DATA to have the calculator find the sample means and standard deviations. Or, you may enter the sample means and sample standard deviations directly by pressing STAT once in the appropriate tests.

Confidence Intervals

- `ZInterval` is the confidence interval for mean when σ is known.
- `TInterval` is the confidence interval for mean when σ is unknown; s estimates σ.
- `1-PropZInt` is the confidence interval for proportion.

> **Note**
>
> The confidence levels should be given as percents (e.g., enter "95" or "`.95`" for a 95 percent confidence level).

Hypothesis Tests

- `Z-Test` is the hypothesis test for single mean when σ is known.
- `T-Test` is the hypothesis test for single mean when σ is unknown; s estimates σ.
- `2-SampZTest` is the hypothesis test for two independent means when both σs are known.
- `2-SampTTest` is the hypothesis test for two independent means when both σs are unknown.
- `1-PropZTest` is the hypothesis test for a single proportion.
- `2-PropZTest` is the hypothesis test for two proportions.

- X^2-Test is the hypothesis test for independence.
- X^2GOF-Test is the hypothesis test for goodness-of-fit (TI-84+ only).
- LinRegTTEST is the hypothesis test for Linear Regression (TI-84+ only).

Note

Input the null hypothesis value in the row below "Inpt." For a test of a single mean, "μ⌀" represents the null hypothesis. For a test of a single proportion, "p⌀" represents the null hypothesis. Enter the alternate hypothesis on the bottom row.

H TABLES

The module contains links to government site tables used in statistics.

Note

When you are finished with the table link, use the back button on your browser to return here.

Tables (NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, January 3, 2009)

- Student *t* table (http://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm)
- Normal table (http://www.itl.nist.gov/div898/handbook/eda/section3/eda3671.htm)
- Chi-Square table (http://www.itl.nist.gov/div898/handbook/eda/section3/eda3674.htm)
- F-table (http://www.itl.nist.gov/div898/handbook/eda/section3/eda3673.htm)
- All four tables (http://www.itl.nist.gov/div898/handbook/eda/section3/eda367.htm) can be accessed by going to http://www.itl.nist.gov/div898/handbook/eda/section3/eda367.htm

95% Critical Values of the Sample Correlation Coefficient Table

- 95% Critical Values of the Sample Correlation Coefficient

Index

A
alternative hypothesis 482
and 216, 216, 217

B
binomial distribution 396, 487
binomial probability distribution 244
bivariate 641
Blinding 36
Box plots 95
box-and-whisker plots 95
box-whisker plots 95

C
categorical data 10
Categorical variables 7
central limit theorem 381, 382, 390
central limit theorem for means 385
central limit theorem for sums 387
chi-square distribution 586
coefficient of determination 652
Cohen's *d* 539
complement 172
conditional probability 172, 311
confidence interval 424, 435
confidence intervals 439, 481
confidence level 425, 439
contingency table 188, 596
continuity correction factor 396
continuous 10
continuous random variable 314
control group 36
critical value 358
cumulative distribution function 305
cumulative distribution function (CDF) 315
Cumulative relative frequency 29

D
data 5, 7
degrees of freedom (*df*) 534
descriptive statistics 6
discrete 10
double-blind experiment 36

E
Empirical Rule 352, 424
error bound 439
error bound for a population mean 425
expected value 237
expected values 587
Experimental Probability of Event A 171
exponential distribution 314, 393

F
F distribution 704
F ratio 705
first quartile 87
frequency 28, 77

G
geometric distribution 252
geometric experiment 250

H
histogram 77
hypergeometric experiment 272
hypergeometric probability 255
hypotheses 482
hypothesis test 489, 508
hypothesis testing. 481

I
independent 183
independent events 175
inferential statistics 6, 423
informed consent 38
Institutional Review Boards (IRB) 38
interquartile range 88
interval scale 28

L
law of large numbers 390
level of measurement 27
level of significance of the test 488
lurking variables 36

M
margin of error 424
mathematical models 6
mean 7, 100, 237, 382, 385, 390
median 88, 100
mode 102
multivariate 641
mutually exclusive 178, 184

N
nominal scale 27
normal distribution 435, 486
normally distributed 382, 387, 486
null hypothesis 482, 488
Numerical variables 7

O
observational studies 36
observed values 587
or 216, 217, 217
ordinal scale 27
outlier 69, 89

P
p-value 487, 489
paired data set 85
parameter 423
Pearson 7
percentiles 87
placebo 36
point estimate 424
Poisson probability distribution 257, 272
pooled proportion 544
population 7, 26
population variance 605
potential outlier 662
Probability 6, 170
probability density function 302
probability distribution function 234
proportion 7

Q
Qualitative data 10
quantitative continuous data 10
Quantitative data 10
quantitative discrete data 10
quartiles 87, 88

R
random assignment 36
Random variable 535, 542
random variables 234
ratio scale 28
relative frequency 28, 77
replacement 175
representative sample 7

S
sample 7
sample mean 383
sample size 383
sample space 183, 195
samples 26
sampling 7
sampling distribution 103
sampling variability of a statistic 111
simple random sample 486
standard deviation 108, 435, 486, 486, 487, 534

standard deviation of a discrete probability distribution 238
standard error 534
standard error of the mean 383
statistics 5
Student's *t*-distribution 435, 486, 486
sum of squared errors (SSE) 648
surveys 36

T

test for homogeneity 600
test of a single variance 605
test of independence 596
Theoretical Probability of Event A 171
tree diagram 194
two-way table 188
Type I error 484, 488
Type II error 484

U

unfair 172
uniform distribution 390
Use the following information to answer the next three exercises 216

V

variable 7
variance 110
variances 704
Variation 25
Venn diagram 201

Z

z-score 435