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.

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

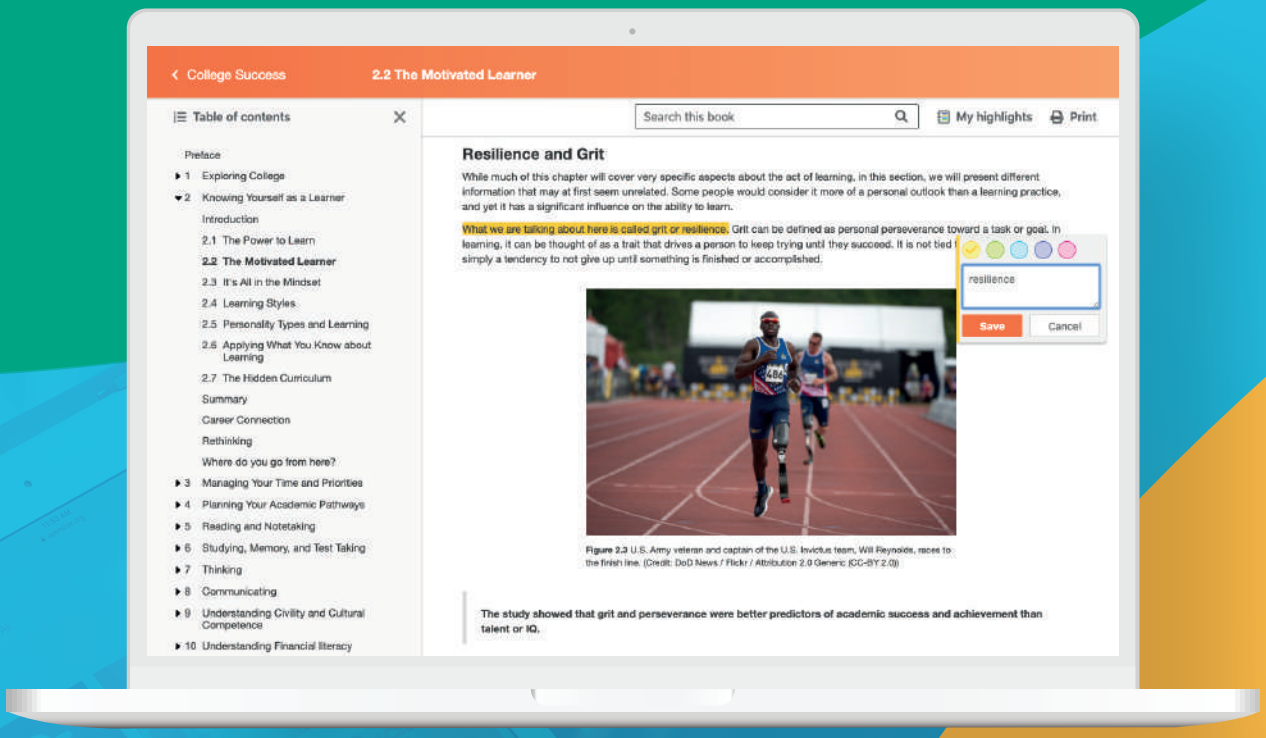
Tammy and Guillermo Treviño

openstax™

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

When you access College Success 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

Table of Contents

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.

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 Sandovl, 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)

Introduction

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

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:

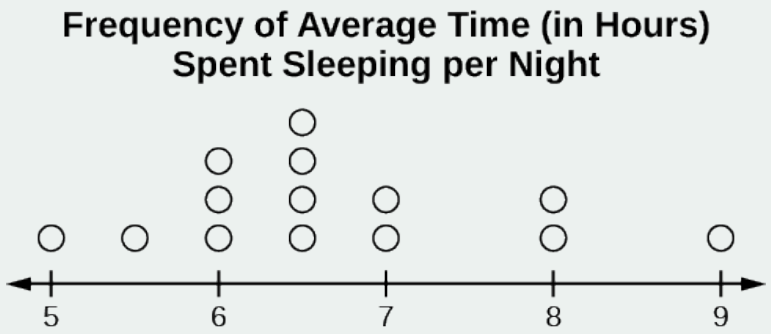


Figure 1.2

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

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 1.1

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

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 1.2

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

Example 1.3

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 1.3

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 = the number of driver dummies—if they had been real people—who would have suffered head injuries.

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

Example 1.4

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 1.4

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 = the number of medical doctors who have been involved in one or more malpractice suits.

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 Σ

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).

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

Solution 1.7

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 Σ

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

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 1.9

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

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 1.2**. What type of data does this graph show?

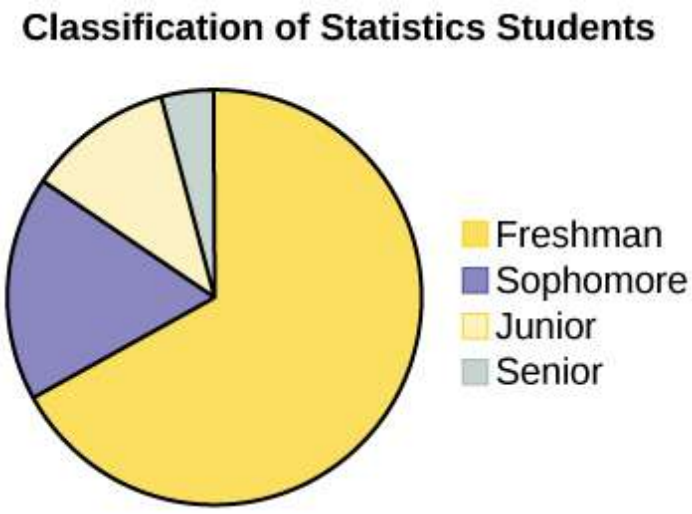


Figure 1.3

Solution 1.10
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 number of students who receive test scores 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.

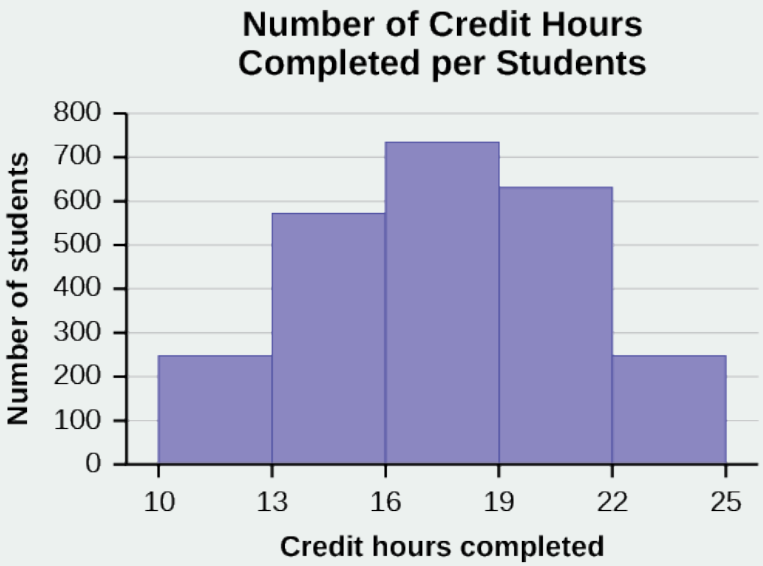


Figure 1.4

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%

Table 1.2 Fall Term 2007 (Census day)

De Anza College				Foothill College		
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 tim

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.

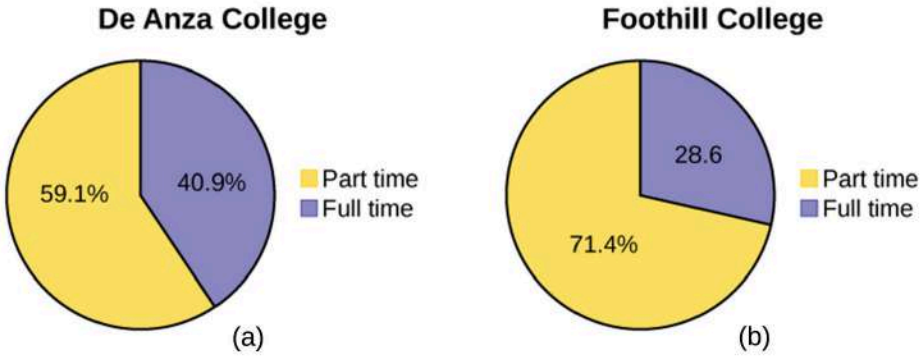


Figure 1.5

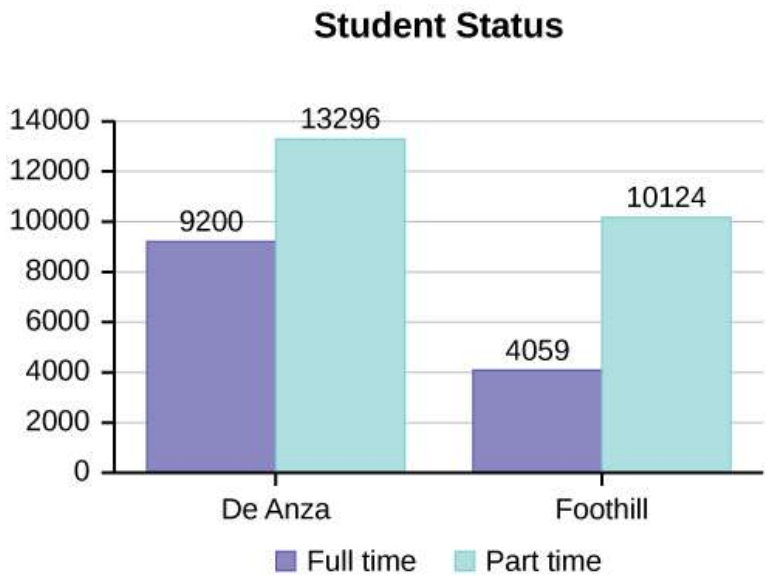


Figure 1.6

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

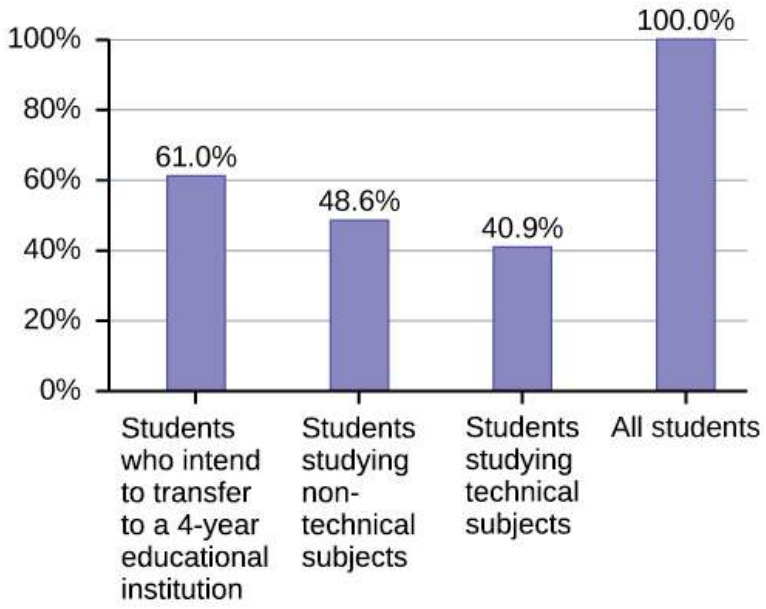


Figure 1.7

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)

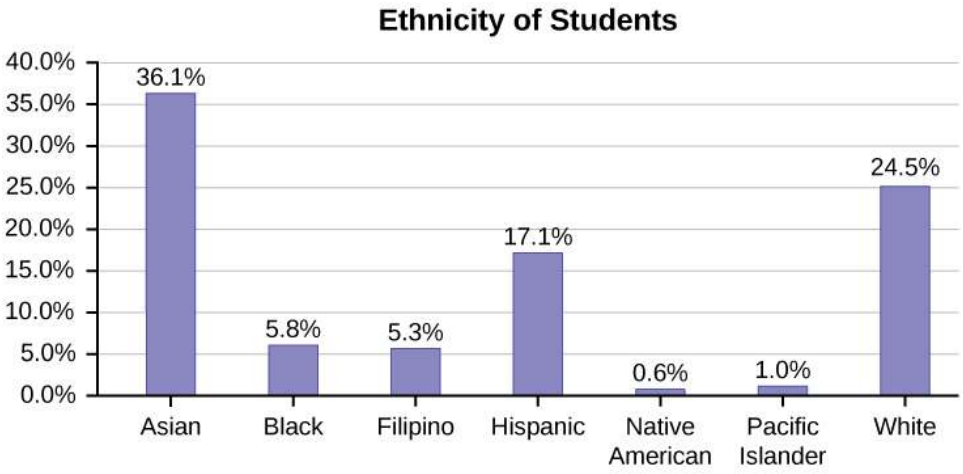


Figure 1.8

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.

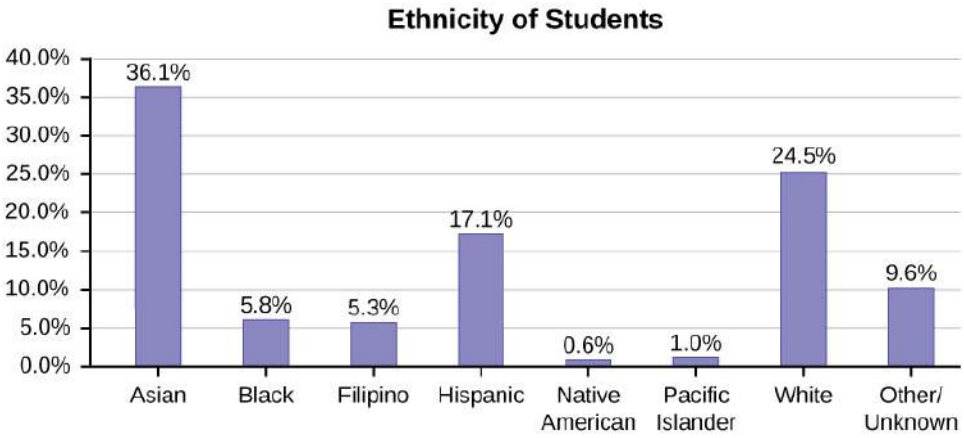


Figure 1.9 Bar Graph with Other/Unknown Category

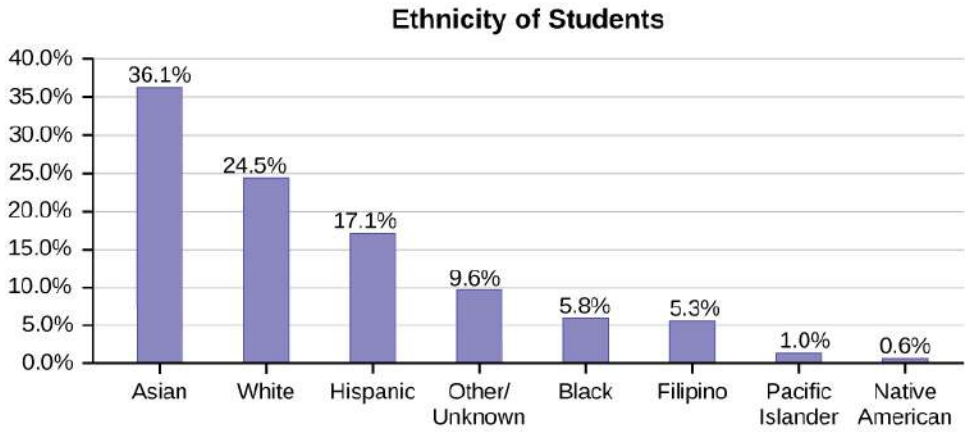


Figure 1.10 Pareto Chart With Bars Sorted by Size

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 1.11**b is organized by the size of each wedge, which makes it a more visually informative graph than the unsorted, alphabetical graph in **Figure 1.11**a.

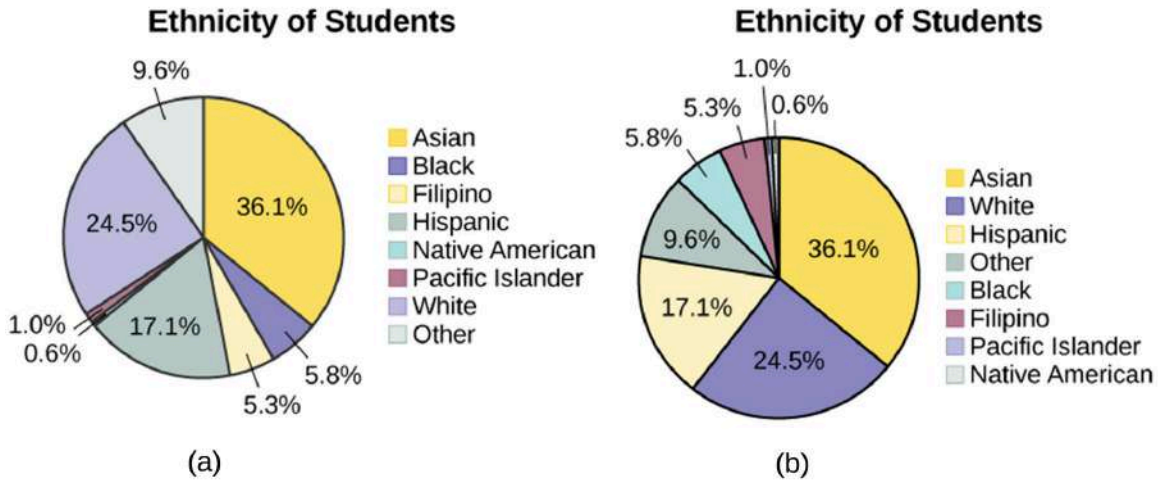


Figure 1.11

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

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 1.11

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**. Each method has pros and cons. 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.

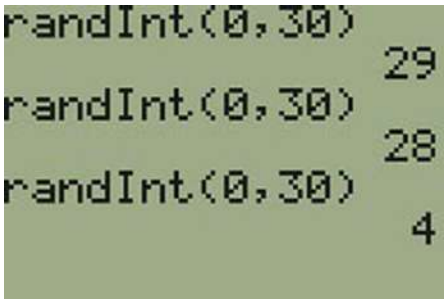


Figure 1.12

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.

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

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 1.12

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. Record the number the calculator displays into the first column. Then, press ENTER.
    d. Record the next number the calculator displays into the second column.
    e. Repeat steps (c) and (d) nine more times until there are a total of 20 quiz scores for the cluster sample.
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

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 1.13

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 high school has 10,000 upperclassman (junior and senior level) students (the population). We are interested in the average amount of money a 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 seniors 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.

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 1.14

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 senior citizens 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 parttime student. Both samples are biased. Also, in both cases, not all students have a chance to be in either sample.

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

Solution 1.14

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.

c. Is the sample biased?

Solution 1.14

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 Σ

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.

Data that is measured using the **interval scale** is similar to ordinal level data because it has a definite ordering but there is a difference between data. The differences between interval scale data can be measured though the data does not have a starting point.

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 because, in both scales, 0 is not the absolute lowest temperature. Temperatures like –10 °F and –15 °C exist and are colder than 0.

Interval level data can be used in calculations, but one type of comparison cannot be done. 80 °C is not four times as hot as 20 °C (nor is 80 °F four times as hot as 20 °F). There is no meaning to the ratio of 80 to 20 (or four to one).

Data that is measured using the **ratio scale** takes care of the ratio problem and gives you the most information. Ratio scale data is like interval scale data, but it has a 0 point 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 smallest 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

Table 1.12 Frequency Table of Student Work Hours

DATA VALUE	FREQUENCY
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

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

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
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.

Table 1.15 represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

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

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

Solution 1.15

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 Σ

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

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

Solution 1.16
Add the relative frequencies in the second and third rows: .03 + .15 = .18 or 18 percent.

Try It Σ

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

Example 1.17

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 1.17

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

Table 1.17 Frequency of Commuting Distances

DATA	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
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

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 1.18

a. No. The frequency column sums to 18, not 19. Not all cumulative relative frequencies are correct.

b. False. The frequency for three miles should be one; for two miles (left out), two. The cumulative relative frequency column should read 1052, 01579, 02105, 03684, 04737, 06316, 07368, 07895, 08421, 09474, 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

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 1.19

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 Σ

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

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 1.20

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

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*.

Example 1.21

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 1.21

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

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 1.22

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

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.

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 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

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 1.23

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

1.1 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

1.2 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

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.

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.

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
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?
c. What percent of students take one or 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.

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

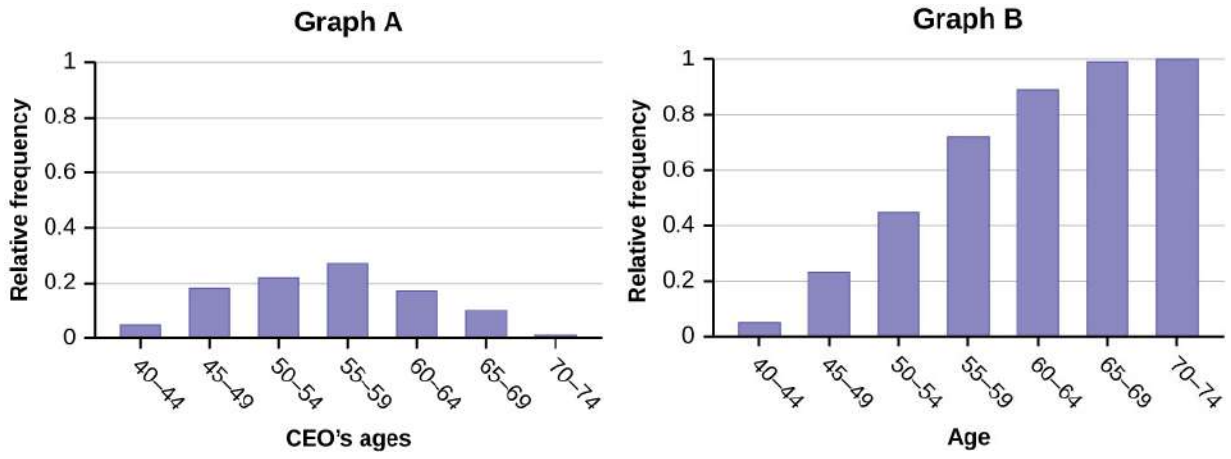
Table 1.40

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?



(a) (b)

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?

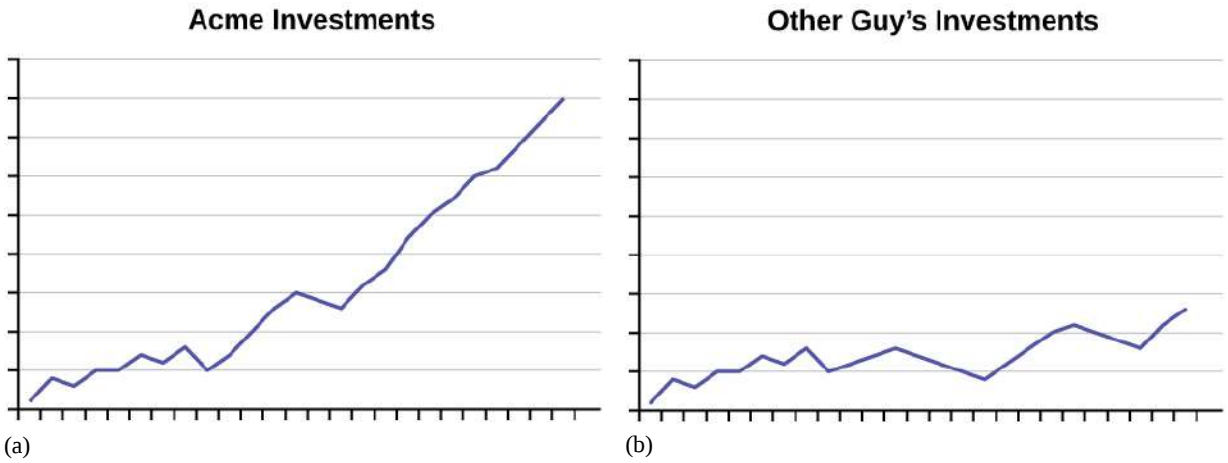


Figure 1.14 As the graphs show, Acme consistently outperforms the Other Guys!

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?

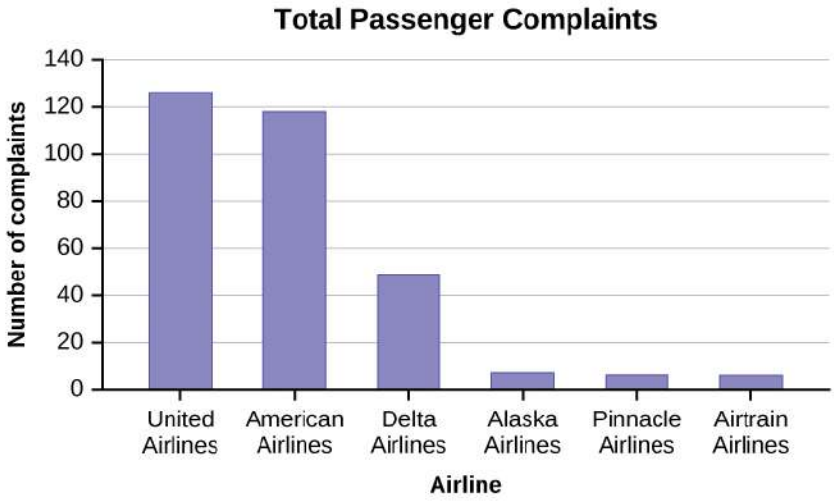


Figure 1.15

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%
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
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)

Introduction

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

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

Do the data seem to have any concentration of values?

The leaves are to the right of the decimal.

Solution 2.2

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

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		

Table 2.4 Presidential Age at Death

President	Age	President	Age	President	Age
Fillmore	74	Wilson	67		
Pierce	64	Harding	57		
Buchanan	77	Coolidge	60		

Table 2.4 Presidential Age at Death

Solution 2.3

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 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 Σ

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

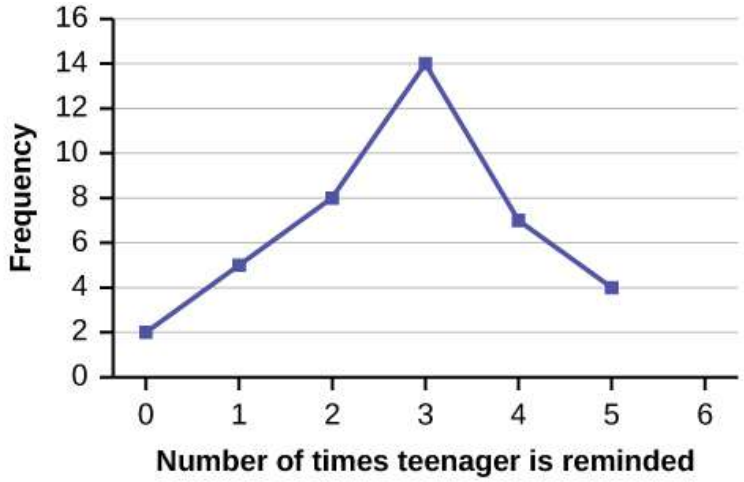


Figure 2.2

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

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 2.5

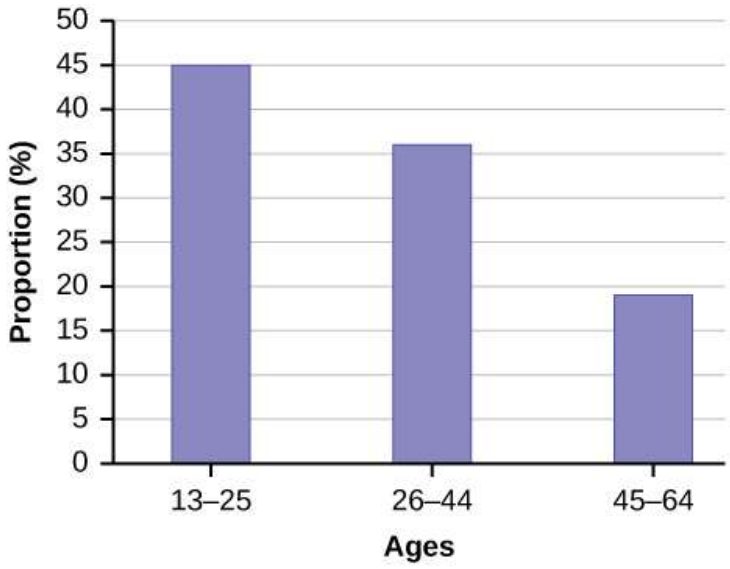


Figure 2.3

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

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 2.6

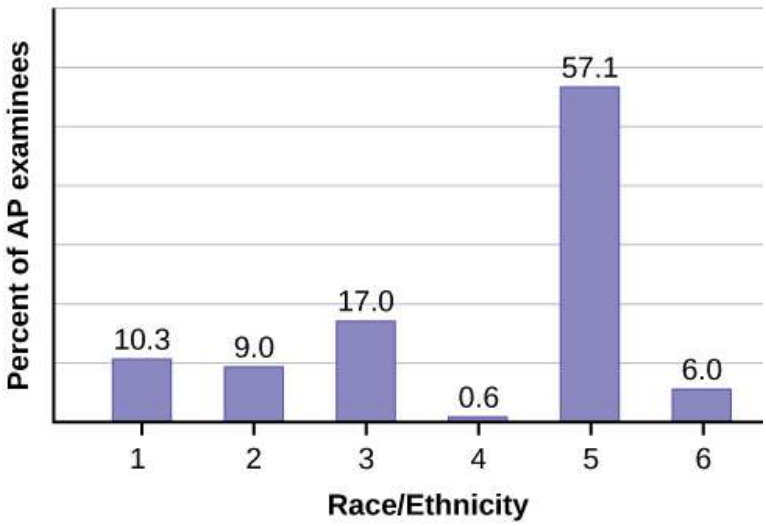


Figure 2.4

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

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 2.7

Dogs = 8/20 = 0.4

Cats = 8/20 = 0.4

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

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 2.8

Men who own dogs = 4/8 = 0.5

Men who own cats = 2/8 = 0.25

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, 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.

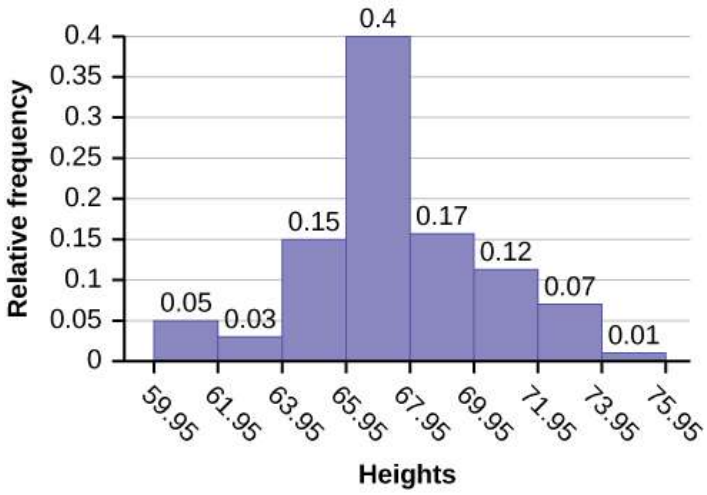


Figure 2.5

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

Table 2.15

Interval	Frequency	Relative Frequency
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, 11.5, 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.

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

Solution 2.10

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.

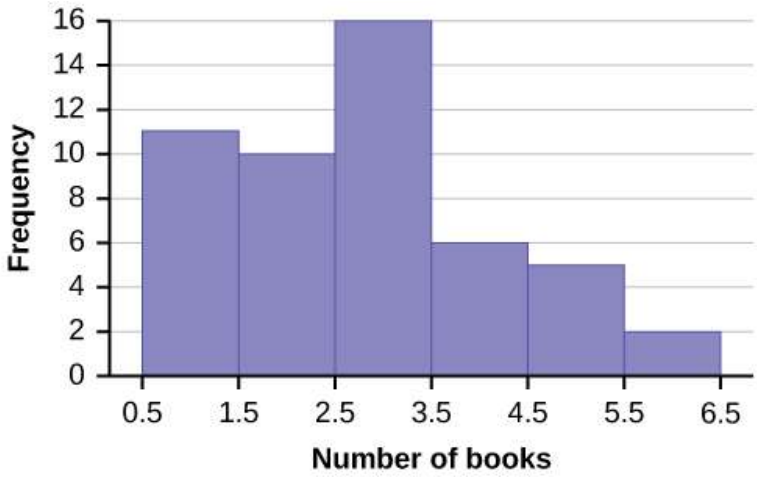


Figure 2.6

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, 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

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 2.11

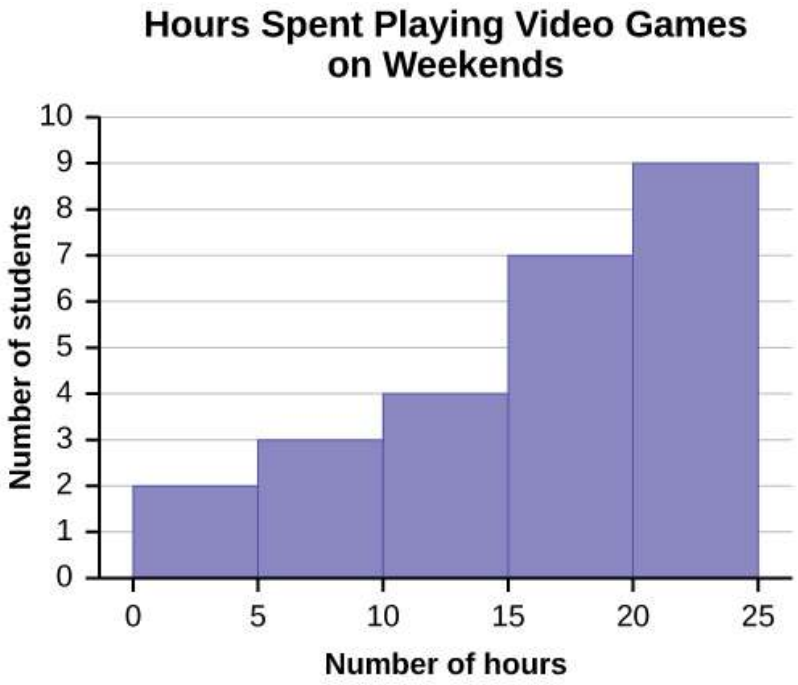


Figure 2.7

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

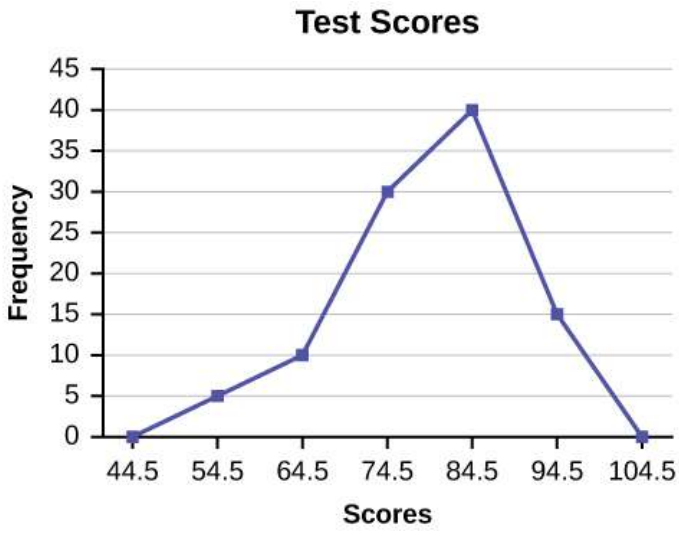


Figure 2.8

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
79.5	89.5	45	95
89.5	99.5	5	100

Table 2.20

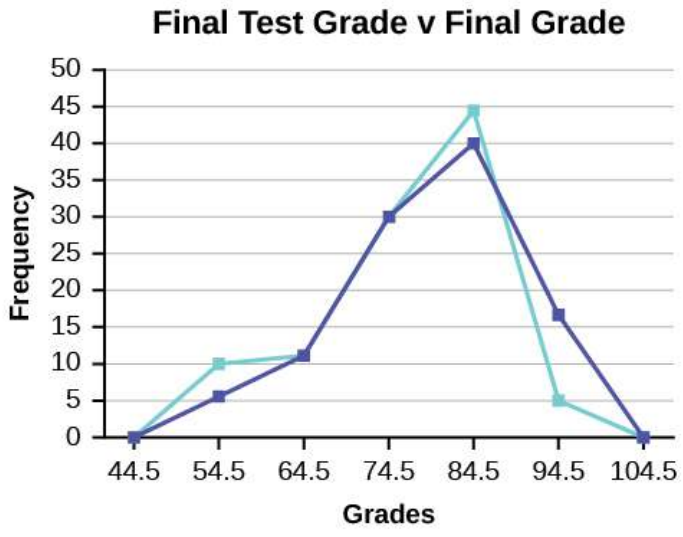


Figure 2.9

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

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 2.14

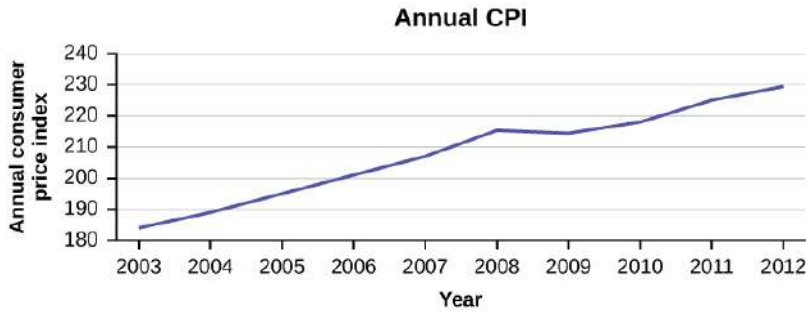


Figure 2.10 The annual amounts are plotted for each year. Then, consecutive points are connected with a line.

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.

$$\frac{6.8 + 7.2}{2} = 7$$

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:

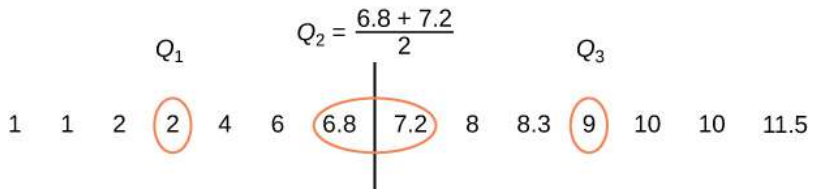


Figure 2.11

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

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 2.15

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 28th 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 28th percentile. They include the two 4s, the five 5s, and the seven 6s. The 28th percentile is between the last six and the first seven. **The 28th percentile is 6.5.**

Find the median. Look again at the Cumulative Relative Frequency column and find .52. The median is the 50th 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 50th percentile is between the 25th, or seven, and 26th, or seven, values. **The median is seven.**

Find the third quartile. The third quartile is the same as the 75th 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 75th 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 38th 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 65th percentile.

Amount of Time Spent on Route (Hours)	Frequency	Relative Frequency	Cumulative Relative Frequency
2	12	.30	.30
3	14	.35	.65
4	10	.25	.90
5	4	.10	1.00

Table 2.25

Example 2.17

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 2.17

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

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 70th percentile.

b. Find the 83rd percentile.

Solution 2.18

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 21st 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 24th position is 71, and the age in the 25th position is 72. Average 71 and 72. The 83rd 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 20th percentile and the 55th 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

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 2.19

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

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 2.20

- 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

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

Solution 2.21

- 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 80^{th} percentile for the number of points earned was 49. Interpret the 80^{th} percentile in the context of this situation.

Example 2.22

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

Solution 2.22

- 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:

0, 0, 10, 20, 30, 30, 30, 40, 45, 60, 60, 60, 90, 120, 300

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.

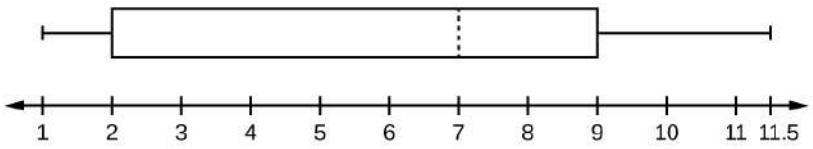


Figure 2.13

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, 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

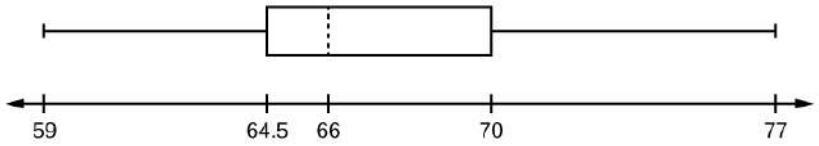


Figure 2.14

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.

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 2.25

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: 75 percent. Ms. Park's class: There are six data values ranging from 25.5 to 78: 27 percent. There are five data values ranging from 78 to the first instance of 81: 23 percent. There are six data values ranging from the second instance of 81 to 89: 27 percent. There are five data values ranging from 90 to 98: 23 percent. There are 17 values between the first quartile, 78, and the largest value, 98: 77 percent.

d.

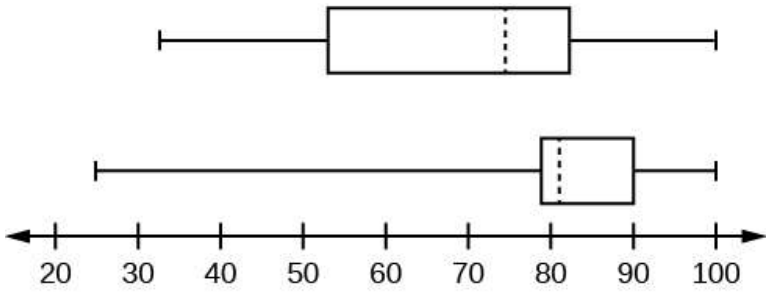


Figure 2.16

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.

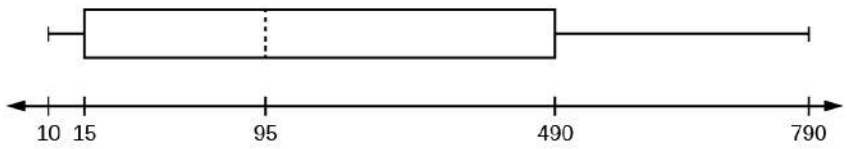


Figure 2.17

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 49th 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 50th and 51st 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

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 2.27

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

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 2.28

$$\bar{x} = \frac{5,\ 000,\ 000 + 49(30,\ 000)}{50} = 129{,}400$$

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

Find the mode.

Solution 2.29

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 $\bar{x}$ 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 $\bar{x}$ 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

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 2.31

- 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.

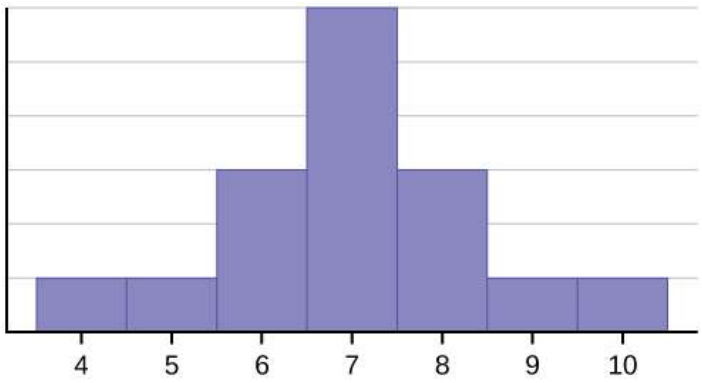


Figure 2.18

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.

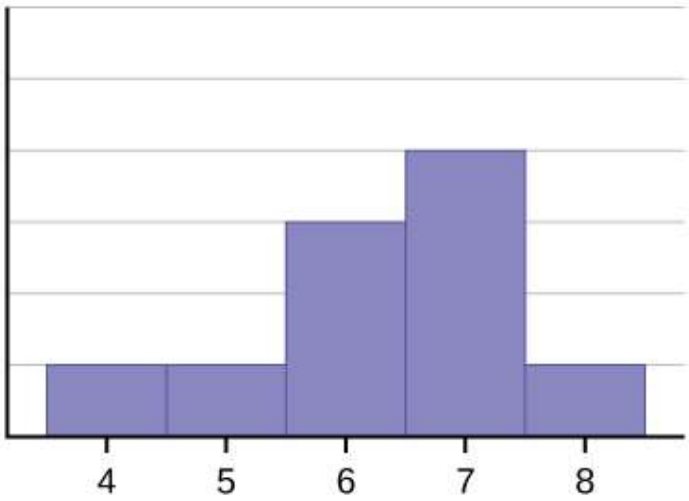


Figure 2.19

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.

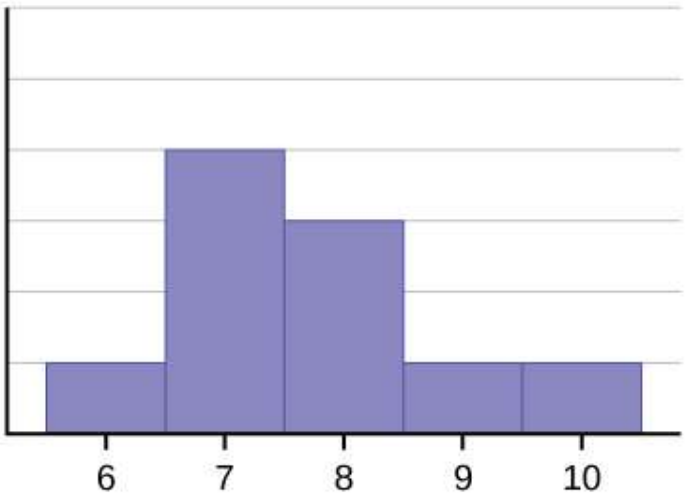


Figure 2.20

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

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 2.32

a.

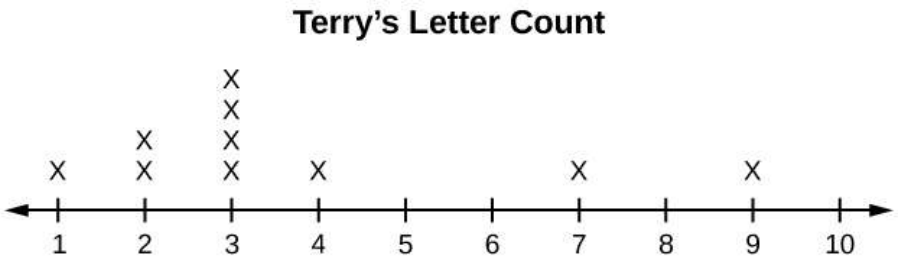


Figure 2.21 Terry's distribution has a right (positive) skew.

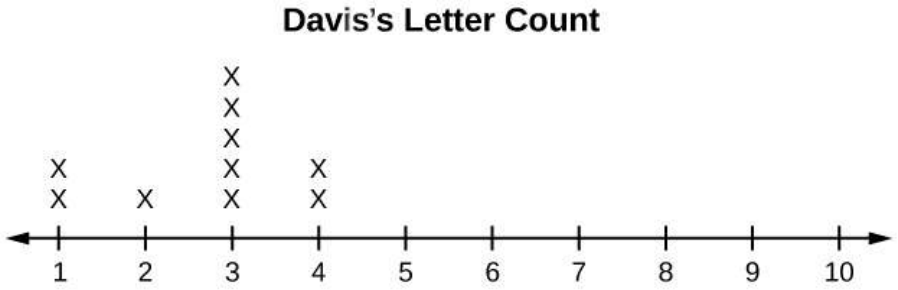


Figure 2.22 Davis's distribution has a left (negative) skew.

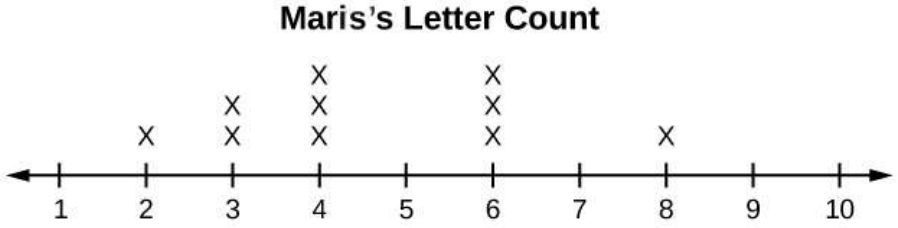


Figure 2.23 Maris's distribution is symmetrically shaped.

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

$$z = \frac{x - \mu}{\sigma} = \frac{7 - 5}{2} = 1.$$

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

$$z = \frac{x - \mu}{\sigma} = \frac{1 - 5}{2} = -2.$$

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:** $x = \bar{x} + (\#ofSTDEV)(s)$
- **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 $\bar{x}$ 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 $x - \bar{x}$.

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 $x - \bar{x}$ 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 $\boldsymbol{n-1}$, one less than the number of items in the sample.

Formulas for the Sample Standard Deviation

- $s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$ or $s = \sqrt{\frac{\Sigma f(x-\bar{x})^2}{n-1}}$
- For the sample standard deviation, the denominator is $\boldsymbol{n-}$; 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

$$\bar{x} = \frac{9 + 9.5(2) + 10(4) + 10.5(4) + 11(6) + 11.5(3)}{20} = 10.525$$

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	$(x - \bar{x})$	$(x - \bar{x})^2$	$(f)(x - \bar{x})^2$
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):

$$s^2 = \frac{9.7375}{20 - 1} = .5125$$

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.

- 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: $x = \bar{x} + (\#\text{ofSTDEVs})(s)$
- For a population: $x = \mu + (\#\text{ofSTDEVs})(\sigma)$
- For this example, use $x = \bar{x} + (\#\text{ofSTDEVs})(s)$ 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 $(\bar{x} + 1s)$.

c. Find the value that is two standard deviations below the mean. Find $(\bar{x} - 2s)$.

d. Find the values that are 1.5 standard deviations **from** (below and above) the mean.

Solution 2.33

a.
- 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.
- $\bar{x}$ = 10.525.
- Use Sx because this is sample data (not a population): Sx=.715891.

b. $(\bar{x} + 1s) = 10.53 + (1)(.72) = 11.25$

c. $(\bar{x} - 2s) = 10.53 - (2)(.72) = 9.09$

d.
- $(\bar{x} - 1.5s) = 10.53 - (1.5)(.72) = 9.45$
- $(\bar{x} + 1.5s) = 10.53 + (1.5)(.72) = 11.61$

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

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 2.34

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.

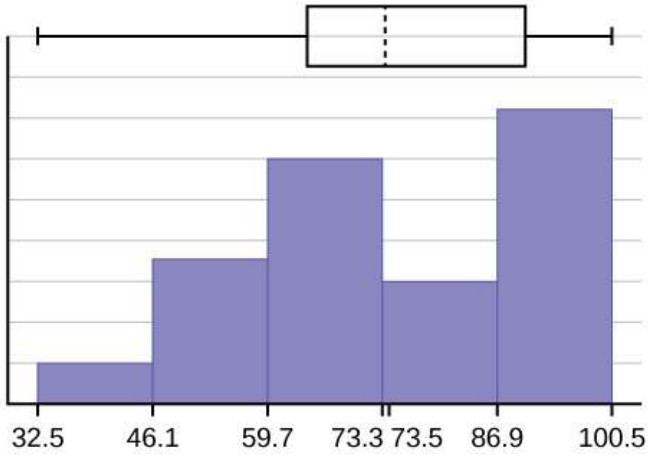


Figure 2.27

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

Table 2.33

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
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	$\bar{x}^2$	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, $\bar{x} = 7.58$, 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, $s_x = \sqrt{\frac{f(m - \bar{x})^2}{n - 1}}$, where s_x = sample standard deviation, $\bar{x}$ = 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.**

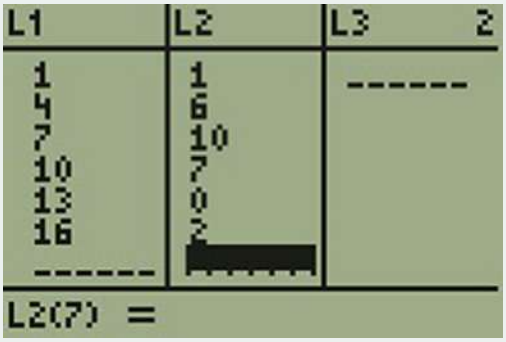


Figure 2.29

Select **STAT**, **CALC**, and **1: 1-Var Stats.**

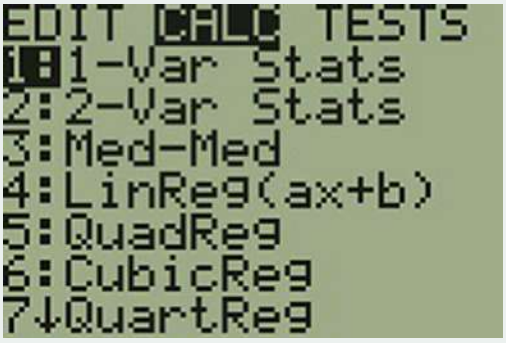


Figure 2.30

Select **2nd**, then **1**, then, **2nd**, then **2 Enter.**

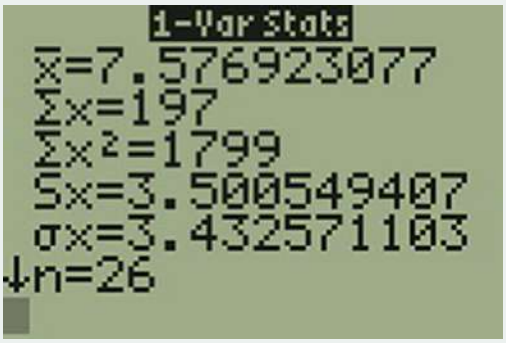


Figure 2.31

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

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 2.36

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

2.1 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.

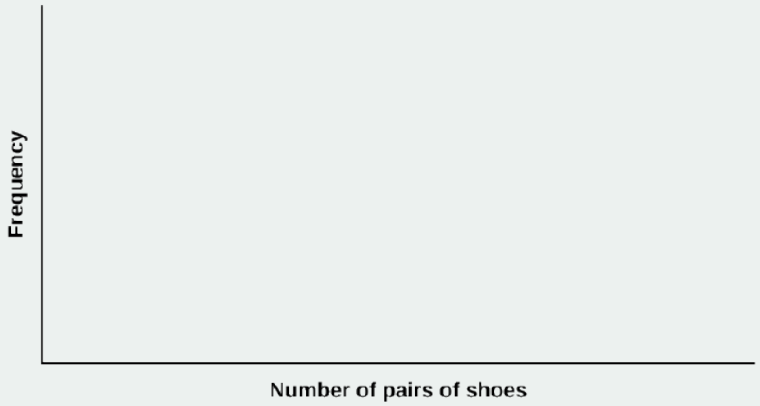


Figure 2.32

3. Calculate the following values:
   a. $\bar{x}$ = _____
   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 $\bar{x}$) is $\bar{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, 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 $x - \bar{x}$ where x is a value of the data and $\bar{x}$ 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

$s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2}$ where

s_x = sample standard deviation

$\bar{x}$ = sample mean

$z = \frac{(x - \bar{x})}{s}$ and

$z = \frac{(x - \mu)}{\theta}$.

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
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
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
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 85^{th} percentile of wait times. Is that good or bad? Write a sentence interpreting the 85^{th} 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 78^{th} 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 90^{th} percentile. Should the manufacturer and the consumer be pleased or upset by this result? Explain and write a sentence that interprets the 90^{th} 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 96^{th} 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 34^{th} percentile. The 34^{th} 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 = 50^{th} percentile = _____

36. Third quartile = _____

37. Interquartile range (*IQR*) = _____ – _____ = _____

38. 10^{th} percentile = _____

39. 70^{th} 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 = $\bar{x}$ = _____

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, 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.

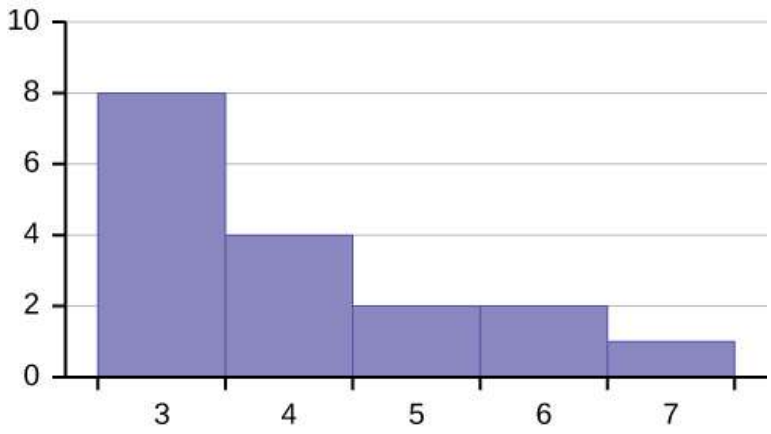


Figure 2.34

56. Describe the relationship between the mode and the median of this distribution.

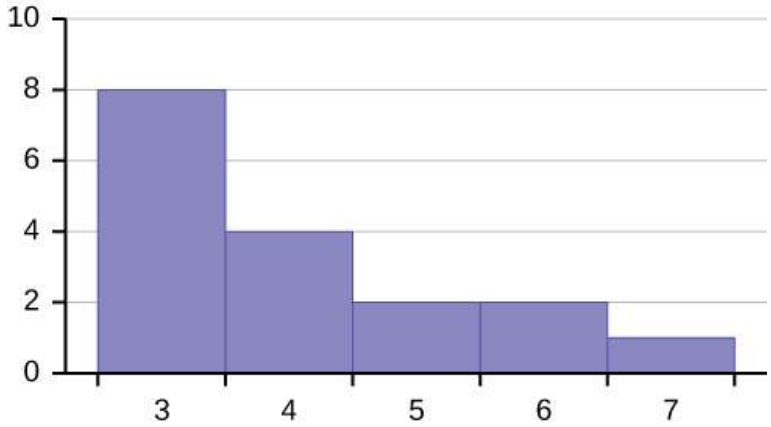


Figure 2.35

57. Describe the relationship between the mean and the median of this distribution.

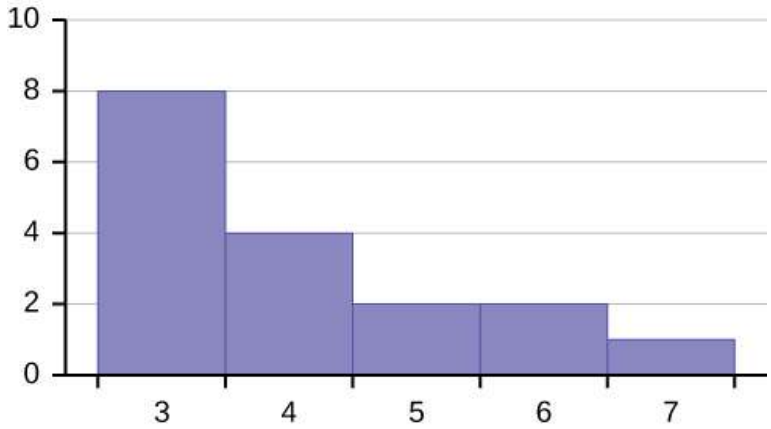


Figure 2.36

58. Describe the shape of this distribution.

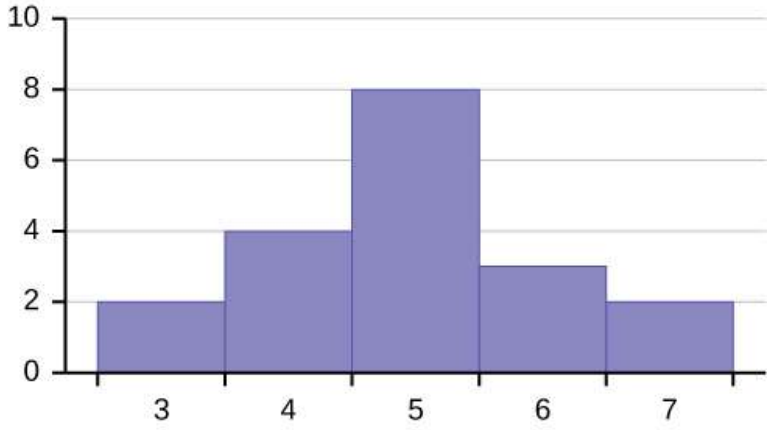


Figure 2.37

59. Describe the relationship between the mode and the median of this distribution.

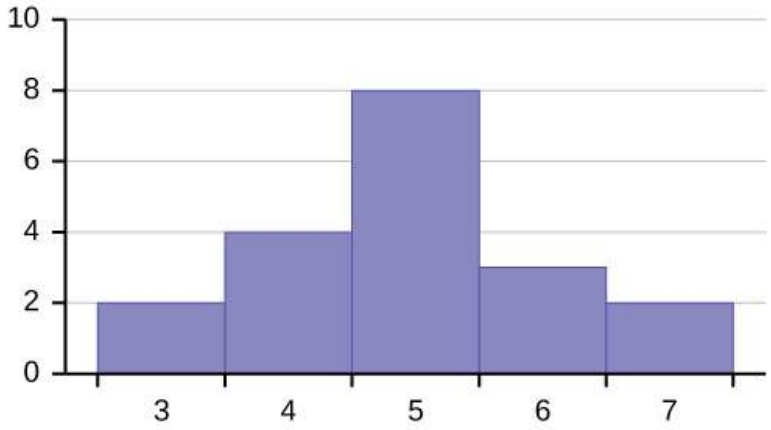


Figure 2.38

60. Are the mean and the median the exact same in this distribution? Why or why not?

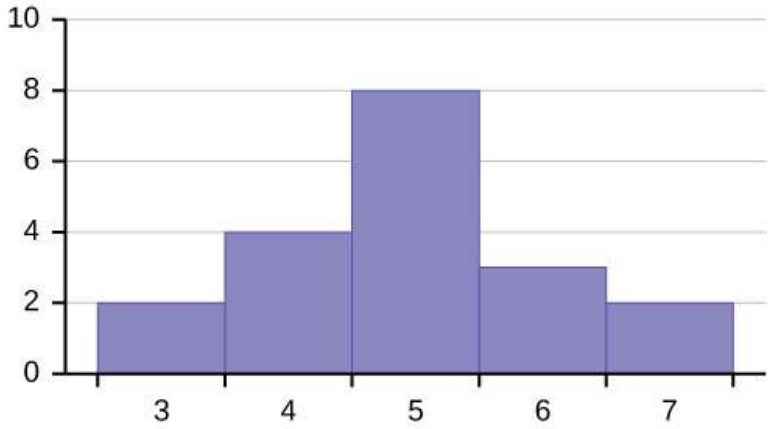


Figure 2.39

61. Describe the shape of this distribution.

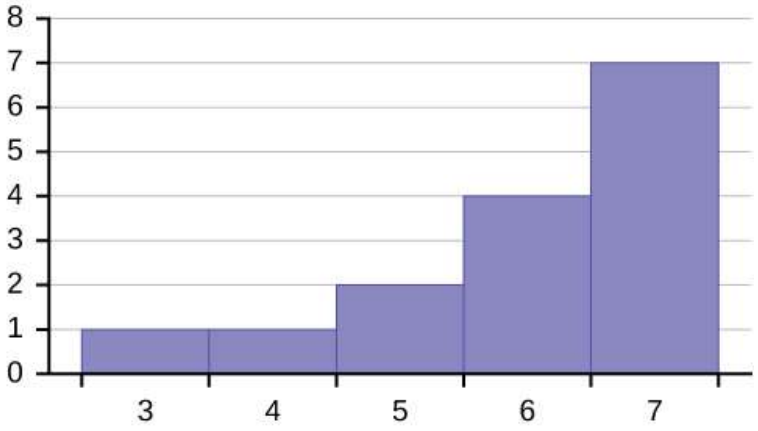


Figure 2.40

62. Describe the relationship between the mode and the median of this distribution.

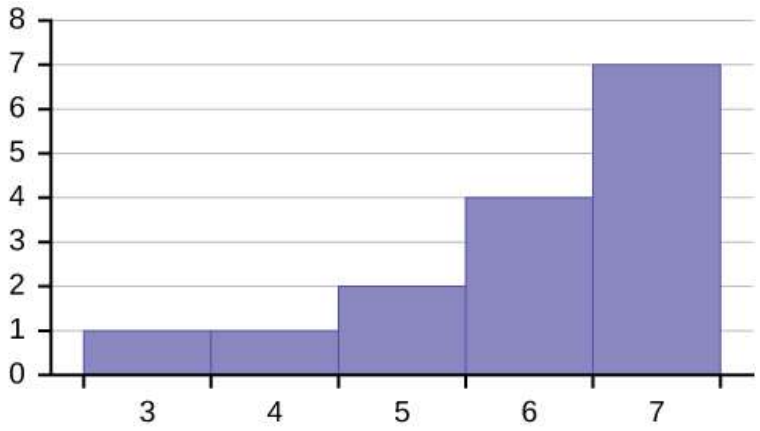


Figure 2.41

63. Describe the relationship between the mean and the median of this distribution.

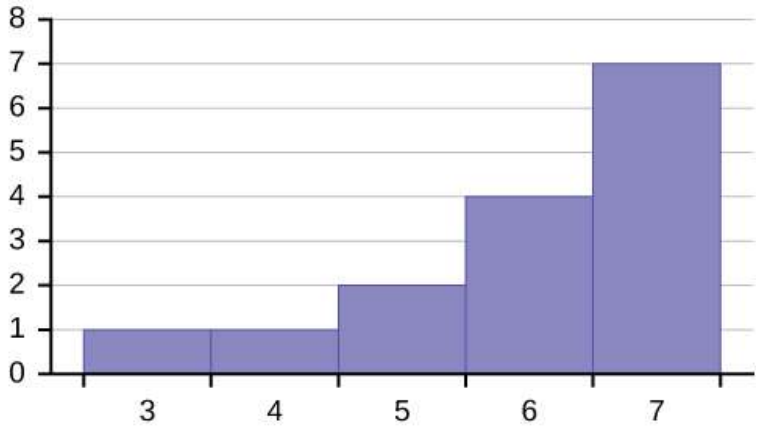


Figure 2.42

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
a. above the mean, and
b. 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
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	
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	
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.

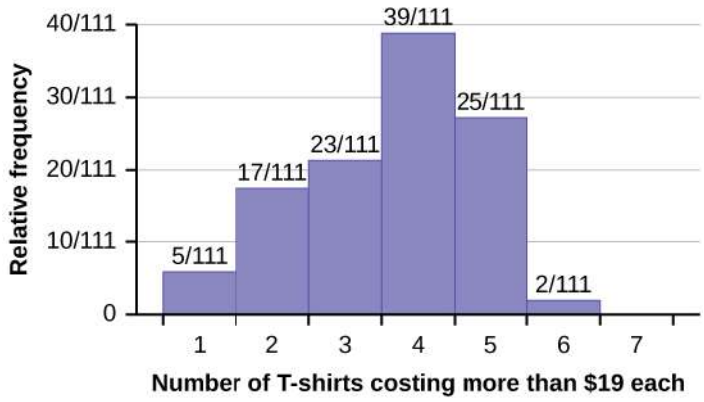
82. 25 randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

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
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.

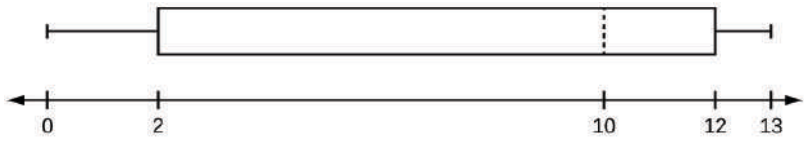


Figure 2.43

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:

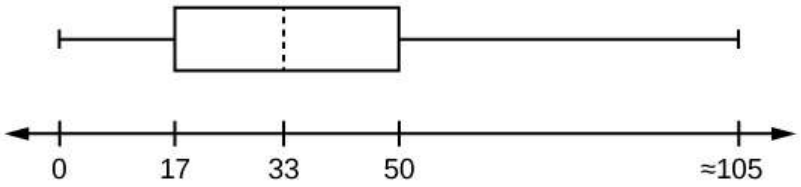


Figure 2.44

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:

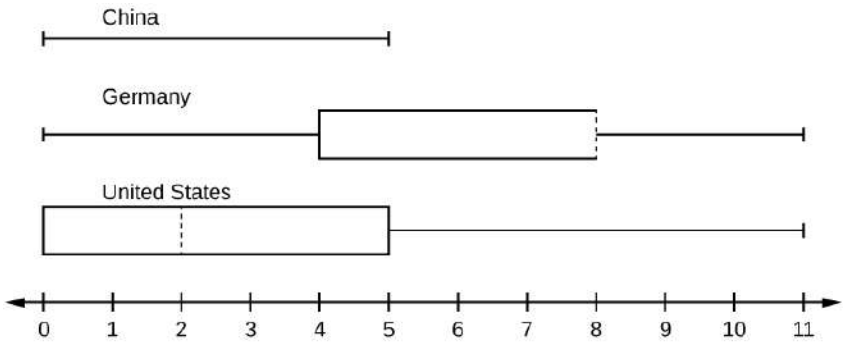


Figure 2.45

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.

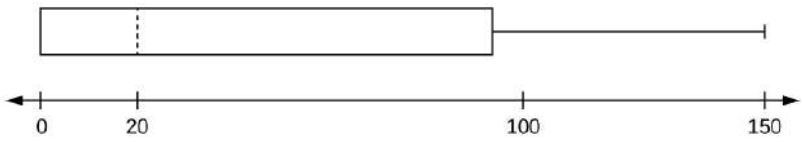


Figure 2.46

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.

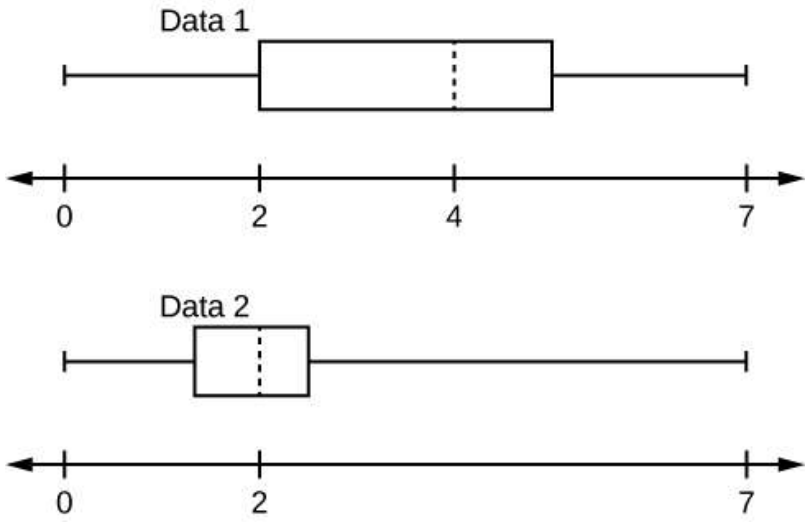


Figure 2.47

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:

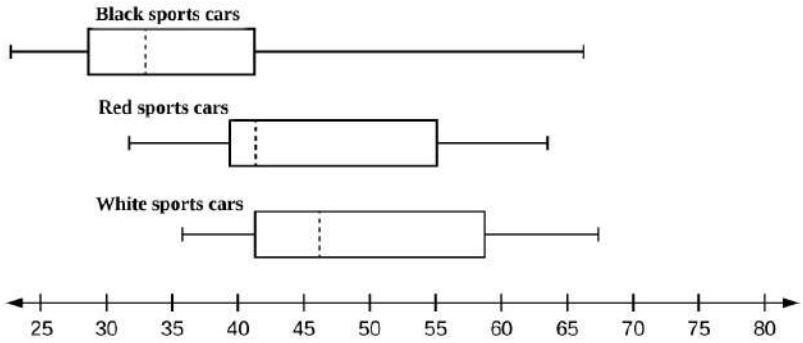


Figure 2.48

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
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?

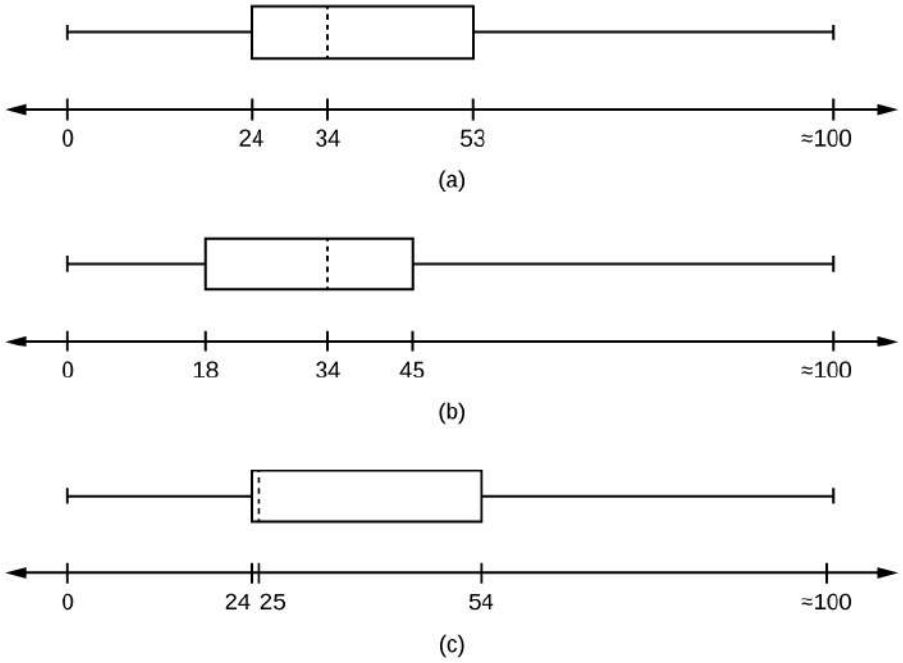


Figure 2.49

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
$\bar{x}$	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?

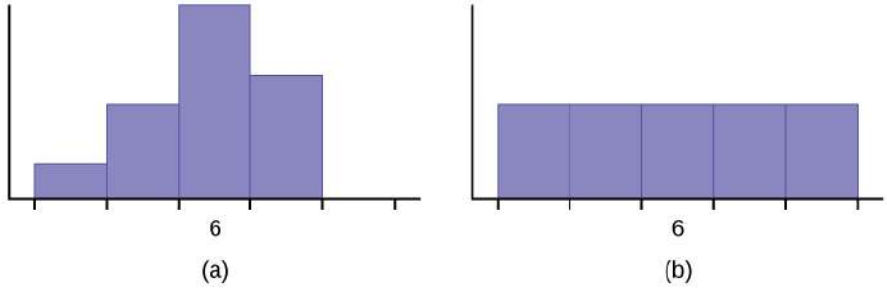


Figure 2.50

d. If the two box plots depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?

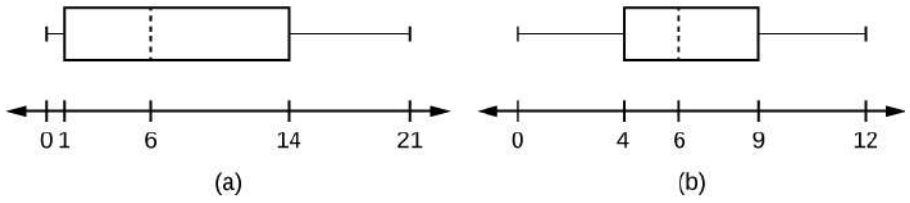


Figure 2.51

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
			Total = 20

Table 2.82

Number of Years	Frequency	Number of Years	Frequency
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 $\bar{x}$.
- 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, $\bar{x}$
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.

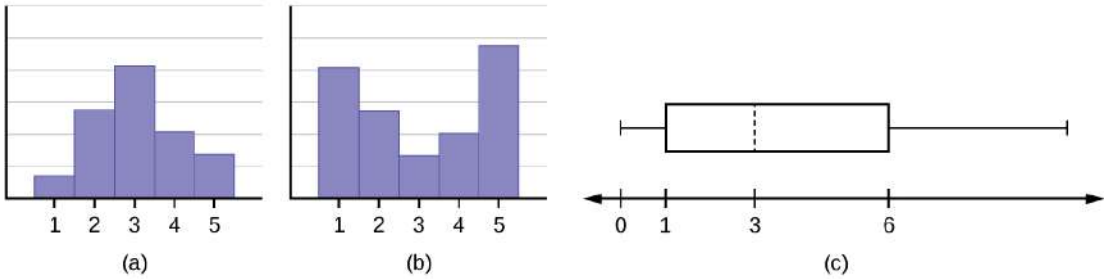


Figure 2.52

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 65^{th} percentile.
d. Find the 10^{th} 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 80^{th} 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

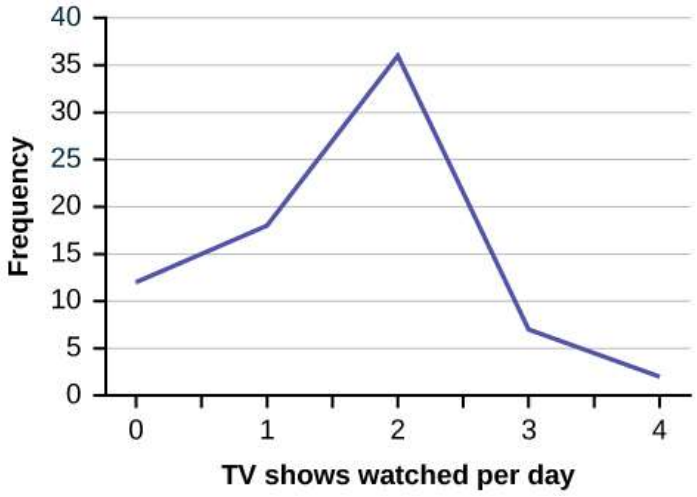


Figure 2.54

9

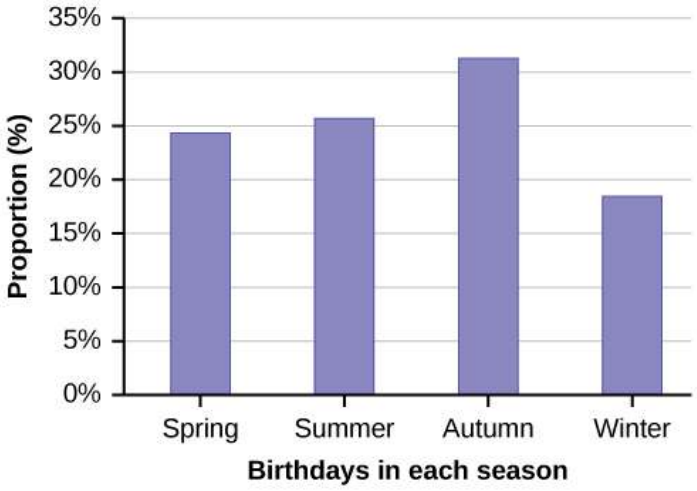


Figure 2.55

11

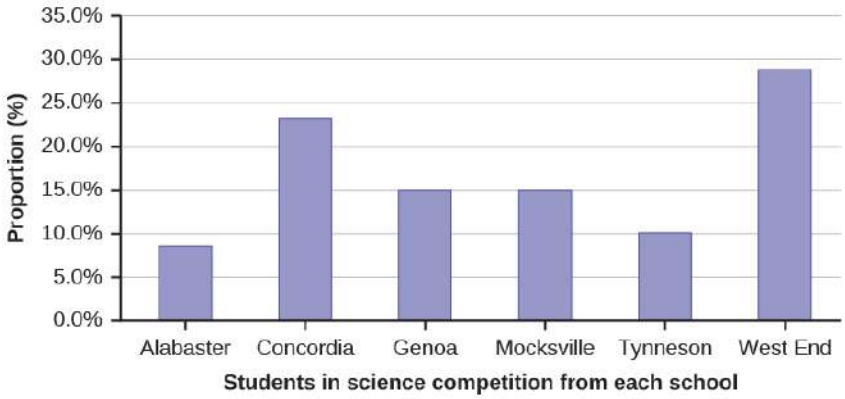


Figure 2.56

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.

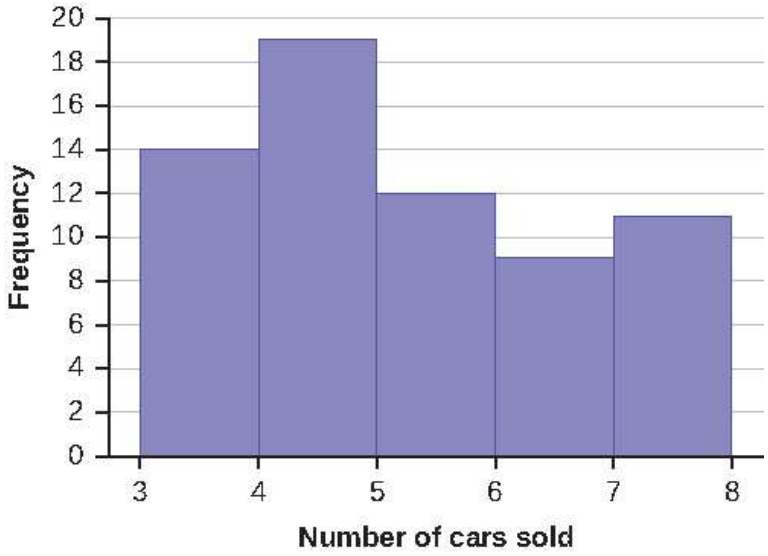


Figure 2.57

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.

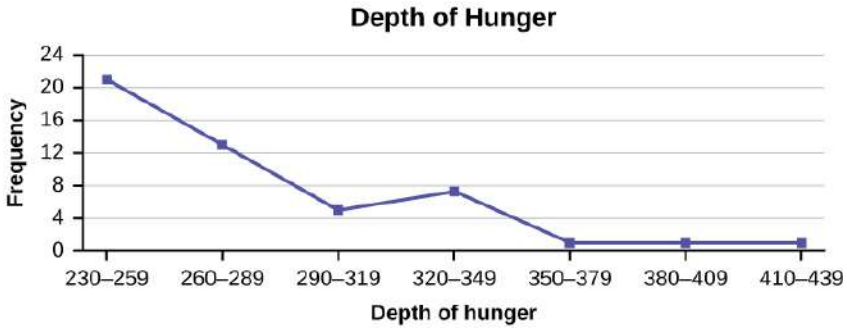


Figure 2.58

21

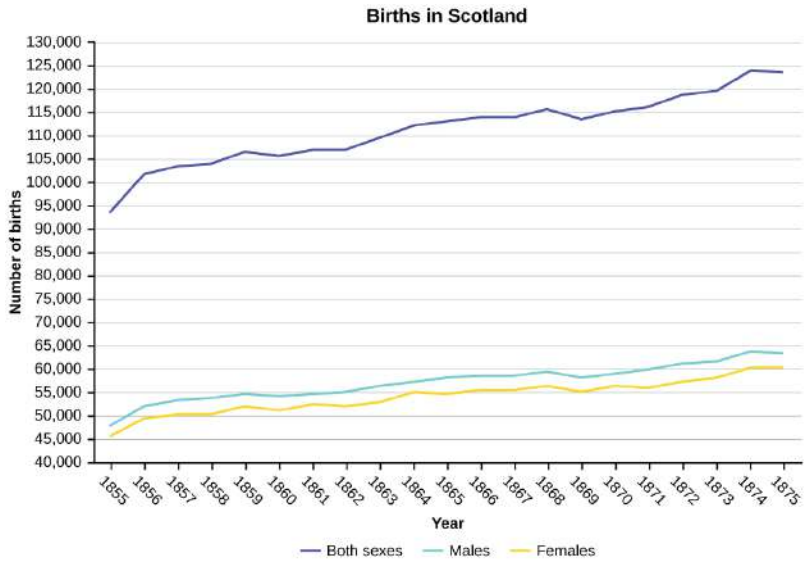


Figure 2.59

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. $s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{193{,}157.45}{30} - 79.5^2} = 10.88$

b. $s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{380{,}945.3}{101} - 60.94^2} = 7.62$

c. $s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{440{,}051.5}{86} - 70.66^2} = 11.14$

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}.

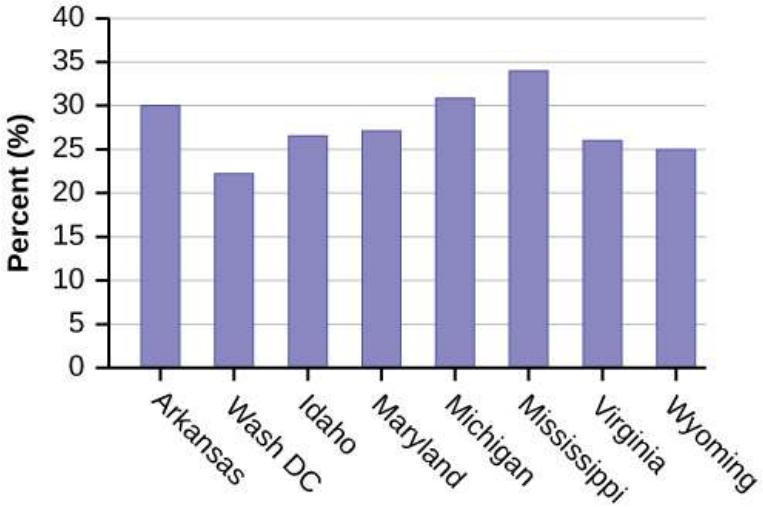


Figure 2.60

b.

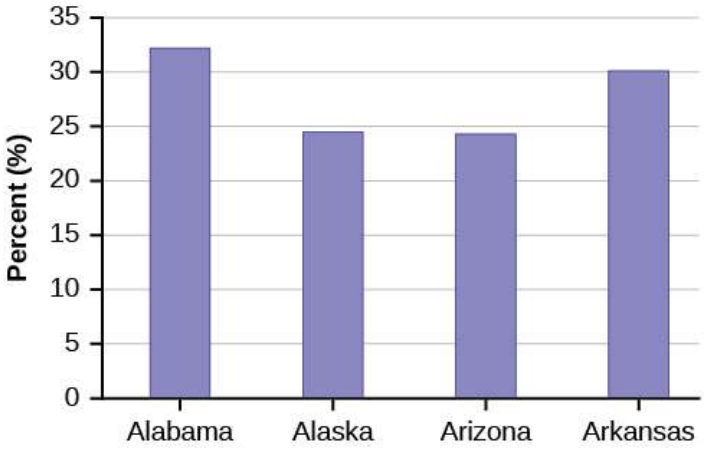
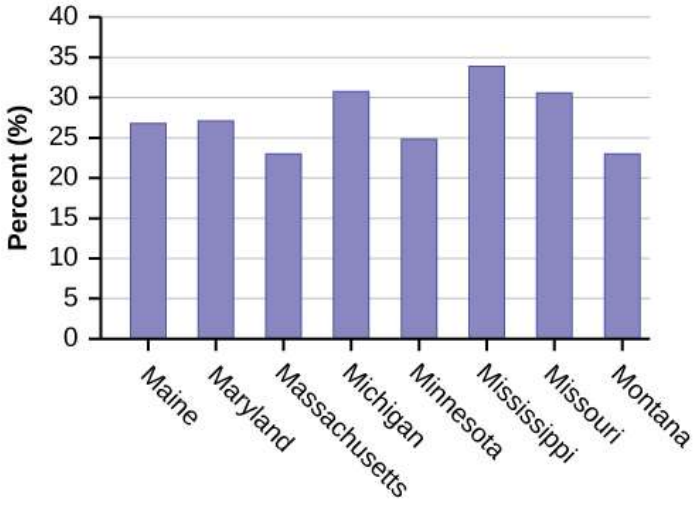


Figure 2.61



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 2.69** and **Table 2.69**.

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.

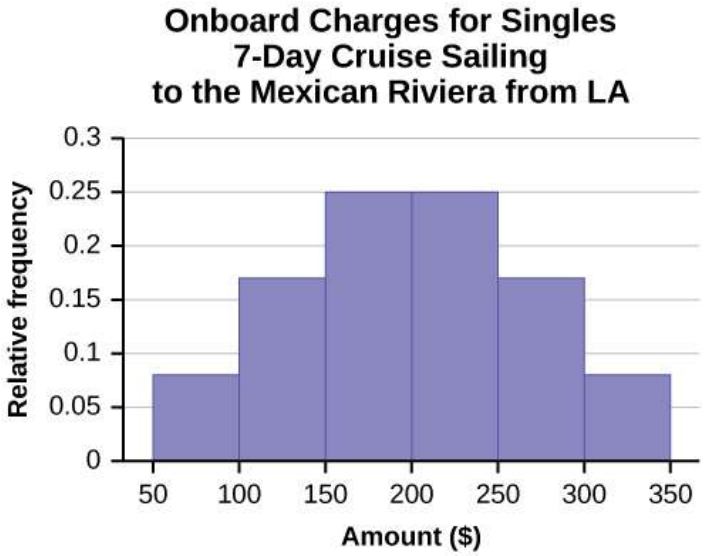


Figure 2.63

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.

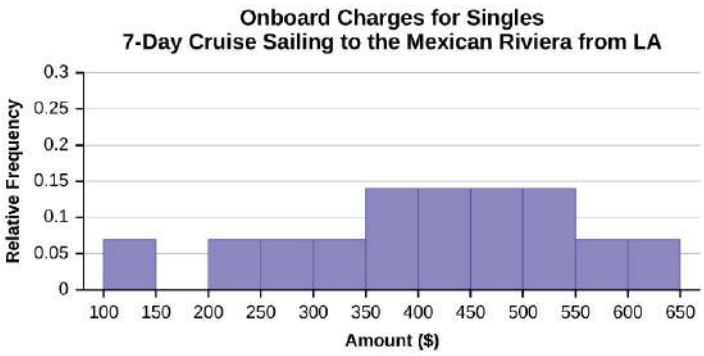


Figure 2.64

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:

i.
- 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* ~ 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, $\bar{x} = \frac{1{,}328.65}{50} = 26.75$

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

- $\bar{x} = 23.32$
- 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

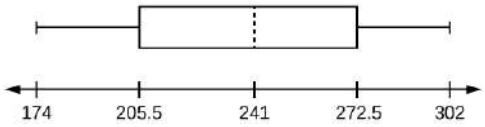
119

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)

Introduction

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

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)	HH

Table 3.1

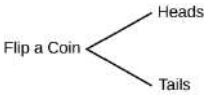
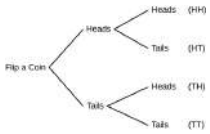
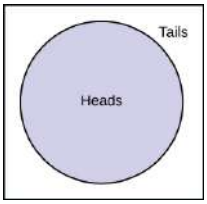
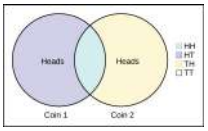
	Flipping a Fair Coin	Flipping Two Fair Coins
	tails (T)	HT TH TT
Tree Diagram*	 Figure 3.2	 Figure 3.3
Venn Diagram*	 Figure 3.4	 Figure 3.5
Set Notation	$S = \{H, T\}$	$S = \{HH, HT, TH, TT\}$

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.

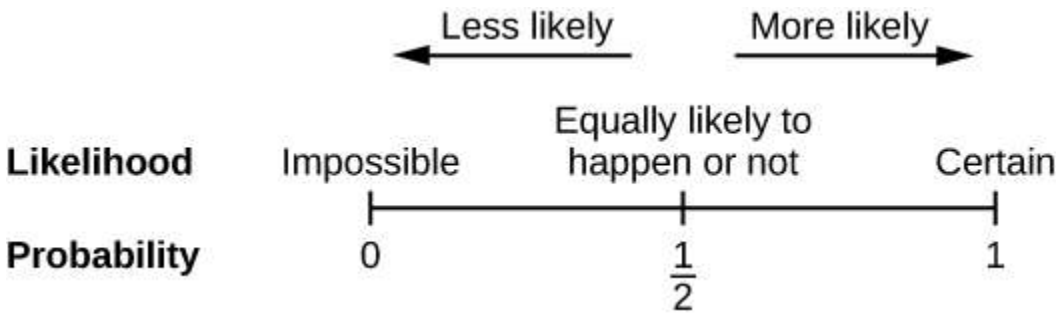


Figure 3.6

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

$$P(A) = \frac{\text{Number of outcomes in event A}}{\text{Total number of possible outcomes.}}$$

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

$$P(A) = \frac{\text{Number of times event A occurs.}}{\text{Total number of trials}}$$

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

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 3.1

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. $P(A) = \frac{\text{number of outcomes in } A}{\text{number of outcomes in } S} = \frac{9}{19}$, $P(B) = P(A) = \frac{\text{number of outcomes in } B}{\text{number of outcomes in } S} = \frac{6}{19}$
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

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 3.2

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

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 3.3

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. $P(R|M) = \frac{P(R \text{and} M)}{P(M)} = \frac{0.43}{0.52} = .8269$ (rounded to four decimal places)

l. $P(F|L) = \frac{P(F \text{and} L)}{P(L)} = \frac{0.04}{0.13} = .3077$ (rounded to four decimal places)

m. $P(L|F) = \frac{P(L \text{and} F)}{P(F)} = \frac{0.04}{0.48} = .0833$ (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.

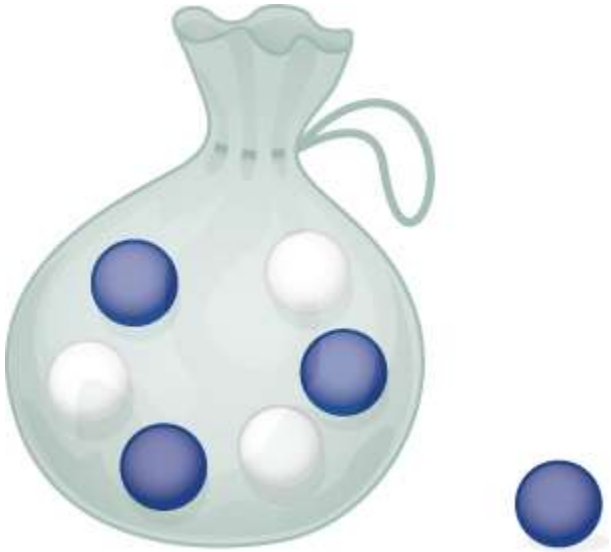


Figure 3.8

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.

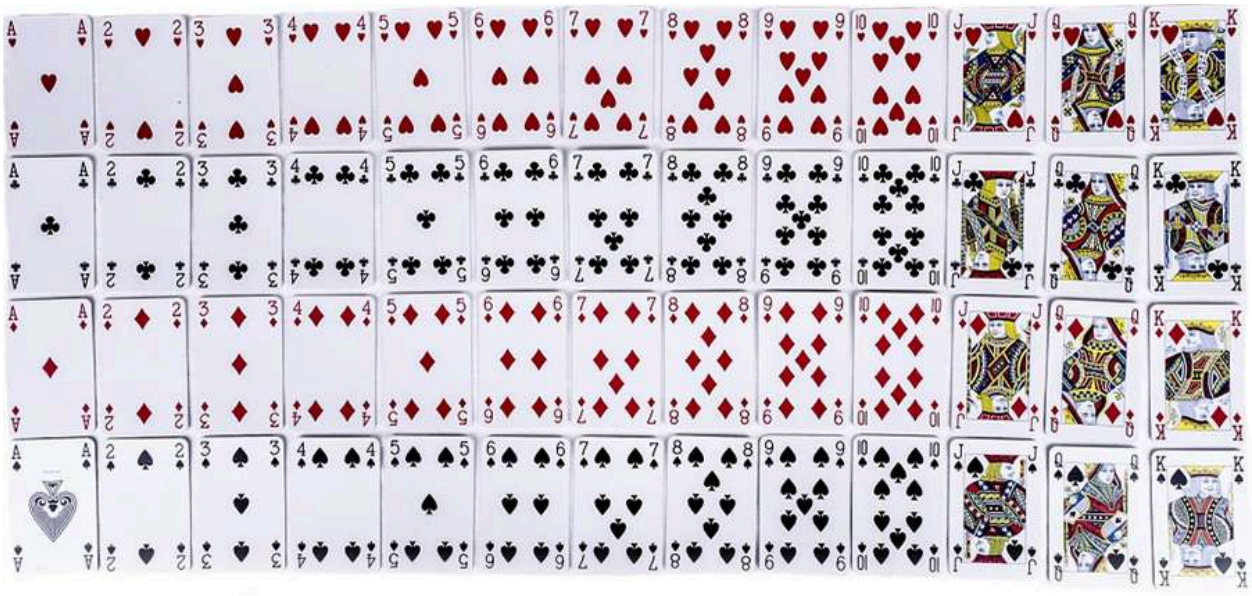


Figure 3.9

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

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 3.5

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 $P(A \text{ AND } B) = 0$.

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 $B = \{4, 5\}$. $P(A \text{ AND } B) = \frac{2}{10}$ 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 $P(A \text{ AND } C) = 0$. 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 $B = \{TT\}$, $P(B \text{ AND } C) = 0$. 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

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 3.7

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}.

Are C and E mutually exclusive events? Answer yes or no. Why or why not?

Solution 3.8

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:

- $P(A|B) = P(A)$
- $P(B|A)$
- $P(A \text{ AND } B) = P(A)P(B)$

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.

a. Show that $P(G|H) = P(G)$.

Solution 3.9

$$P(G|H) = \frac{P(G \text{ AND } H)}{P(H)} = \frac{.3}{.5} = .6 = P(G)$$

b. Show $P(G \text{ AND } H) = P(G)P(H)$.

Solution 3.9

$P(G)P(H) = (.6)(.5) = .3 = P(G \text{ AND } H)$

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

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 3.10

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:

- $P(F) = 0.60$; $P(L) = 0.50$
- $P(F \text{ AND } L) = 0.45$
- $P(L|F) = 0.75$

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 $P(F|L)$ 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 $P(L|F)$ equals $P(L)$. We are given that $P(L|F) = .75$, 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

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 $H1$ and $T6$.

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 $P(A \text{ AND } B)$? If $P(A \text{ AND } B) = 0$, then A and B are mutually exclusive.

h. Are A and B independent? Hint—Is $P(A \text{ AND } B) = P(A)P(B)$? If $P(A \text{ AND } B) = P(A)P(B)$, then A and B are independent. If not, then they are dependent.

Solution 3.13

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): $T1$, $T2$, $T3$, $T4$, $T5$, $T6$, $H1$, $H2$, $H3$, $H4$, $H5$, $H6$

e. $A = \{H2, H4, H6\}$; $P(A) = \frac{\text{number of outcomes in } A}{\text{number of possible outcomes}} = \frac{3}{12}$

f. $B = \{H3\}$; $P(B) = \frac{1}{12}$

g. Yes, because $P(A \text{ AND } B) = 0$

h. $P(A \text{ AND } B) = 0$. $P(A)P(B) = \left(\frac{3}{12}\right)\left(\frac{1}{12}\right)$. $P(A \text{ AND } B)$ 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 $P(A \text{ AND } B) = P(B)P(A|B)$.

This equation can be rewritten as $P(A \text{ AND } B) = P(B)P(A|B)$, the multiplication rule.

If A and B are **independent**, then $P(A|B) = P(A)$. In this special case, $P(A \text{ AND } B) = P(A|B)P(B)$ becomes $P(A \text{ AND } B) = P(A)P(B)$.

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 $P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$.

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.

$$P(H \text{ or } J) = P(H) + P(J) - P(H \text{ and } J)$$
$$= \frac{13}{52} + \frac{4}{52} - \frac{1}{52}$$
$$= \frac{16}{52}$$
$$= \frac{4}{13}$$
$$\approx .3077$$

If A and B are **mutually exclusive**, then $P(A \text{ AND } B) = 0$. Then $P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$ becomes $P(A \text{ OR } B) = P(A) + P(B)$.

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

$$P(H \text{ or } S) = P(H) + P(S)$$
$$= \frac{13}{52} + \frac{13}{52}$$
$$= \frac{26}{52}$$
$$= \frac{1}{2}$$
$$= .5$$

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 $P(B) = .35$.
- $P(A \text{ AND } B) = 0$ because Klaus can only afford to take one vacation.
- Therefore, the probability that he chooses either New Zealand or Alaska is $P(A \text{ OR } B) = P(A) + P(B) = .6 + .35 = .95$. 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.

a. What is the probability that he makes both goals?

Solution 3.15

a. The problem is asking you to find $P(A \text{ AND } B) = P(B \text{ AND } A)$. Since $P(B|A) = .90$: $P(B \text{ AND } A) = P(B|A) P(A) = (.90)(.65) = .585$.

Carlos makes the first and second goals with probability .585.

b. What is the probability that Carlos makes either the first goal or the second goal?

Solution 3.15

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.

c. Are A and B independent?

Solution 3.15

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 $P(B)P(A)$.

d. Are A and B mutually exclusive?

Solution 3.15

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.

a. What is the probability that the member is a novice swimmer?

Solution 3.16

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 $\frac{28}{150}$.

b. What is the probability that the member practices four times a week?

Solution 3.16

b. $\frac{40 + 30 + 10}{150} = \frac{80}{150}$

c. What is the probability that the member is an advanced swimmer and practices four times a week?

Solution 3.16

c. There are 40 advanced swimmers who practice four times per week, so the probability is $\frac{40}{150}$.

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 3.16

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.

e. Are being a novice swimmer and practicing four times a week independent events? Why or why not?

Solution 3.16

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.

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 3.17

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.

a. What is the probability that the woman develops the disease? What is the probability that woman tests negative?

Solution 3.18

a. $P(B) = .143$; $P(N) = .85$

b. Given that the woman develops the disease, what is the probability that she tests negative?

Solution 3.18

b. Among women who develop the disease, the test is negative 2 percent of the time, so $P(N|B) = .02$

c. What is the probability that the woman has the disease AND tests negative?

Solution 3.18

c. $P(B \text{ AND } N) = P(B)P(N|B) = (.143)(.02) = .0029$

d. What is the probability that the woman has the disease OR tests negative?

Solution 3.18

d. $P(B \text{ OR } N) = P(B) + P(N) - P(B \text{ AND } N) = .143 + .85 - .0029 = .9901$

e. Are having the disease and testing negative independent events?

Solution 3.18

e. No. $P(N) = .85$; $P(N|B) = .02$. So, $P(N|B)$ does not equal $P(N)$.

f. Are having the disease and testing negative mutually exclusive?

Solution 3.18

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

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 3.19

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:

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 3.20

a. This is the same as the marginal distribution (**Section 1.2**).

$$P(\text{Person uses a cell phone while driving}) = \frac{\text{number who use cell phones while driving}}{\text{number in study}} = \frac{305}{755} \approx .4040$$

b. The marginal distribution is

$$P(\text{Person had no violation in the last year}) = \frac{\text{number who had no violation}}{\text{number in study}} = \frac{685}{755} \approx .9073.$$

c. Find the number of participants who satisfy *both* conditions.

$$P(\text{Person had no violation in the last year AND uses a cell phone while driving}) = \frac{\text{number who had no violation AND uses cell phone while driving}}{\text{number in study}}$$
$$= \frac{280}{755}$$
$$\approx .3709$$

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.

$$P(\text{Person uses a cell phone while driving OR had no violation in the last year}) = \frac{25 + 405 + 280}{755}$$
$$= \frac{710}{755}$$
$$\approx .9404$$

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.

$$\text{(Person uses a cell phone while driving GIVEN the person had a violation in the last year)} = \frac{\text{number who used cell phone AND had a violation}}{\text{number in study who had a violation in the last year}}$$

$$= \frac{25}{70}$$

$$\approx .3571$$

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

a. Complete the table.

Solution 3.21

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

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 3.21

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.

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 3.21

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." $P(M|L) = \frac{25}{41}$

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 3.21

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.

a. Complete the probability contingency table. Calculate the entries for the totals. Verify that the lower-right corner entry is 1.

Solution 3.22

a.

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	$\mathbf{\frac{19}{60}}$
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	$\mathbf{\frac{41}{60}}$
Total	$\mathbf{\frac{5}{15}}$	$\mathbf{\frac{4}{12}}$	$\mathbf{\frac{2}{6}}$	1

Table 3.9 Door Choice

b. What is the probability that Alissa does not catch Muddy?

Solution 3.22

b. $\frac{41}{60}$

c. What is the probability that Muddy chooses Door One OR Door Two given that Muddy is caught by Alissa?

Solution 3.22

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

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 3.23

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.

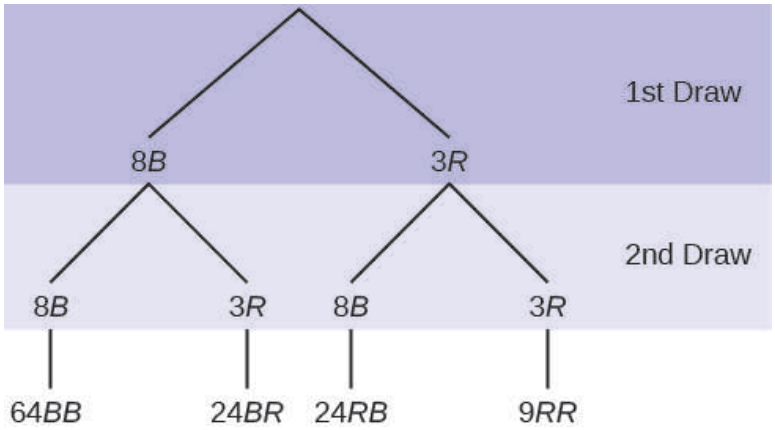


Figure 3.10 Total = 64 + 24 + 24 + 9 = 121.

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**.

a. List the 24 *BR* outcomes: *B*1*R*1, *B*1*R*2, *B*1*R*3, . . .

Solution 3.24

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

b. Calculate *P*(*RR*).

Solution 3.24

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*) = $\left(\frac{3}{11}\right)\left(\frac{3}{11}\right) = \frac{9}{121.}$

c. Calculate $P(RB \text{ OR } BR)$.

Solution 3.24

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 $P(RB \text{ OR } BR) = P(RB) + P(BR) = P(R)P(B) + P(B)P(R) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{11}\right) = \frac{48}{121.}$

d. Using the tree diagram, calculate $P(R \text{ on 1st draw AND } B \text{ on 2nd draw})$.

Solution 3.24

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 $3 \times 8 = 24$ ways to draw red then blue, so $P(RB) = \frac{24}{121}$.

Can you think of another way to find this probability? $P(R \text{ on 1st draw AND } B \text{ on 2nd draw}) = P(RB) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) = \frac{24}{121}$

e. Using the tree diagram, calculate $P(R \text{ on 2nd draw GIVEN } B \text{ on 1st draw})$.

Solution 3.24

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.

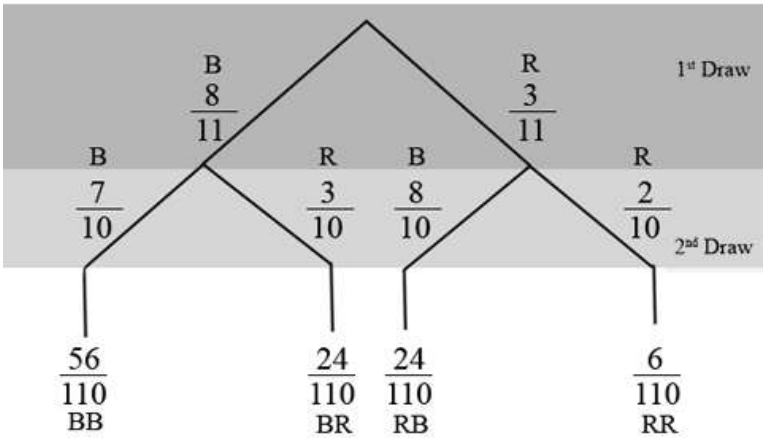


Figure 3.11

$P(R \text{ on 2nd draw GIVEN } B \text{ on 1st}) = P(R \text{ on 2nd} \mid B \text{ on 1st}) = \frac{3}{11}$

You can also use the formula

$$P(R \text{ on 2nd} \mid B \text{ on 1st}) = \frac{P(R \text{ on 2nd AND } B \text{ on 1st})}{P(B \text{ on 1st})} = \frac{\frac{24}{121}}{\frac{64+24}{121}} = \frac{24}{88} = \frac{3}{11.}$$

f. Using the tree diagram, calculate $P(BB)$.

Solution 3.24

f. $P(BB) = \frac{64}{121}$

g. Using the tree diagram, calculate $P(B$ on the 2nd draw GIVEN R on the first draw).

Solution 3.24

g. $P(B \text{ on 2nd draw} | R \text{ on 1st draw}) = \frac{8}{11}$

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 $P(FF)$.

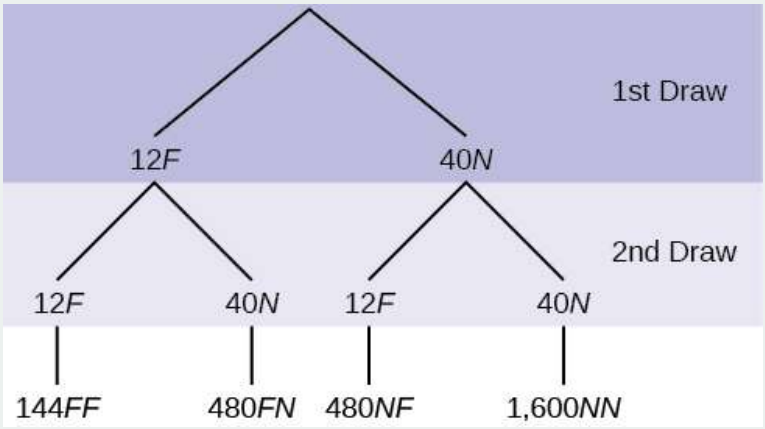


Figure 3.12

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, $P(RR) = \left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$.

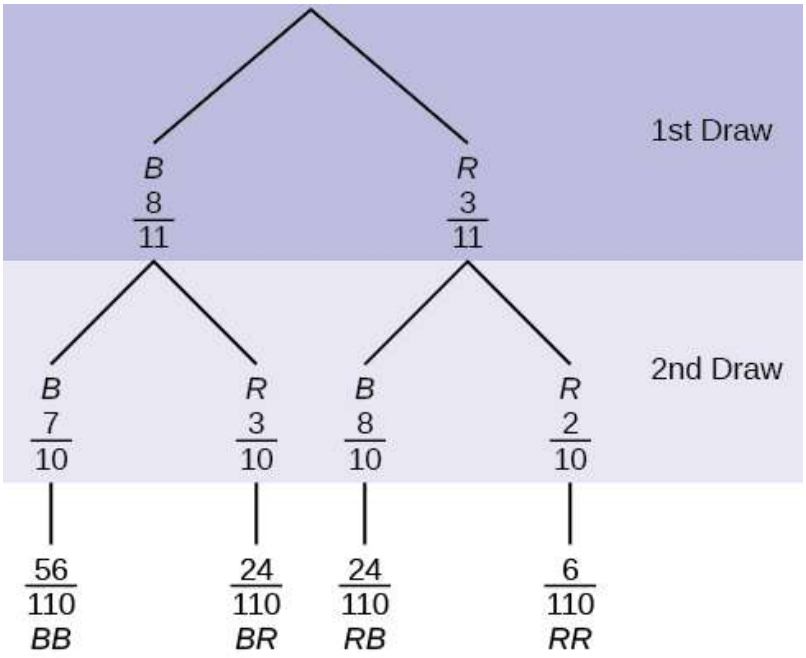


Figure 3.13 Total = $\frac{56+24+24+6}{110} = \frac{110}{110} = 1$.

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:

a. $P(RR)$ = _____

Solution 3.25

a. $P(RR) = \left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$

b. Fill in the blanks.

$P(RB \text{ OR } BR) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) + (\_\_\_\_\_)(\_\_\_\_\_) = \frac{48}{110}$

Solution 3.25

b. $P(RB \text{ OR } BR) = P(RB) + P(BR) = P(R \text{ on 1st})\, P(B \text{ on 2nd}) + P(B \text{ on 1st})\, P(R \text{ on 2nd}) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{10}\right)$

$= \frac{48}{110}$

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 $P(R \text{ on 2nd} | B \text{ on 1st}) = \frac{3}{10}$

Solution 3.25

c. $P(R \text{ on 2nd}|B \text{ on 1st}) = \frac{3}{10}$

d. Fill in the blanks.

$P(R \text{ on 1st AND } B \text{ on 2nd}) = P(RB) = (\_\_\_\_\_)(\_\_\_\_\_) = \frac{24}{100}$

Solution 3.25

d. $P(R \text{ on 1st AND } B \text{ on 2nd}) = P(RB) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) = \frac{24}{100}$

e. Find $P(BB)$.

Solution 3.25

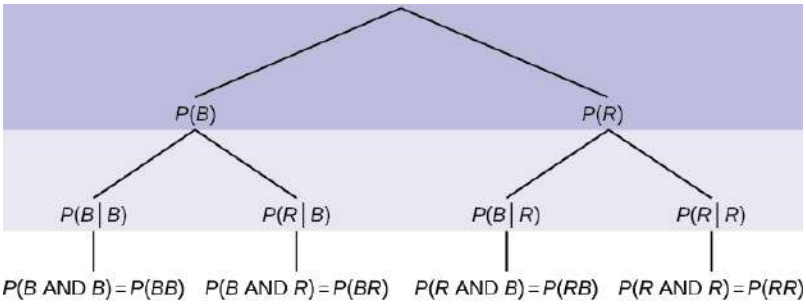
e. $P(BB) = \left(\frac{8}{11}\right)\left(\frac{7}{10}\right)$

f. Find $P(B \text{ on 2nd}|R \text{ on 1st})$.

Solution 3.25

f. Using the tree diagram, $P(B \text{ on 2nd}|R \text{ on 1st}) = P(R|B) = \frac{8}{10}$.

If we are using probabilities, we can label the tree in the following general way:



- $P(R|R)$ here means $P(R \text{ on 2nd}|R \text{ on 1st})$
- $P(B|R)$ here means $P(B \text{ on 2nd}|R \text{ on 1st})$
- $P(R|B)$ here means $P(R \text{ on 2nd}|B \text{ on 1st})$
- $P(B|B)$ here means $P(B \text{ on 2nd}|B \text{ on 1st})$

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.

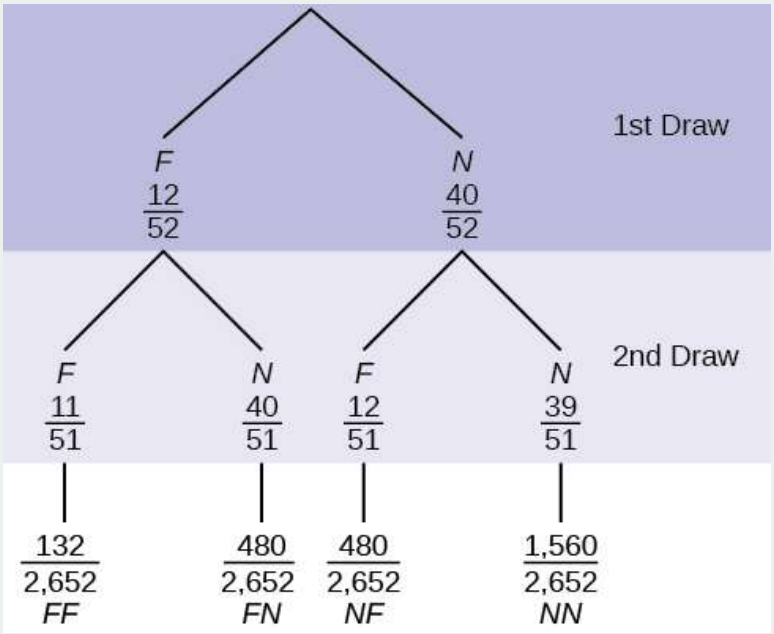


Figure 3.14

a. Find $P(FN \text{ OR } NF)$.

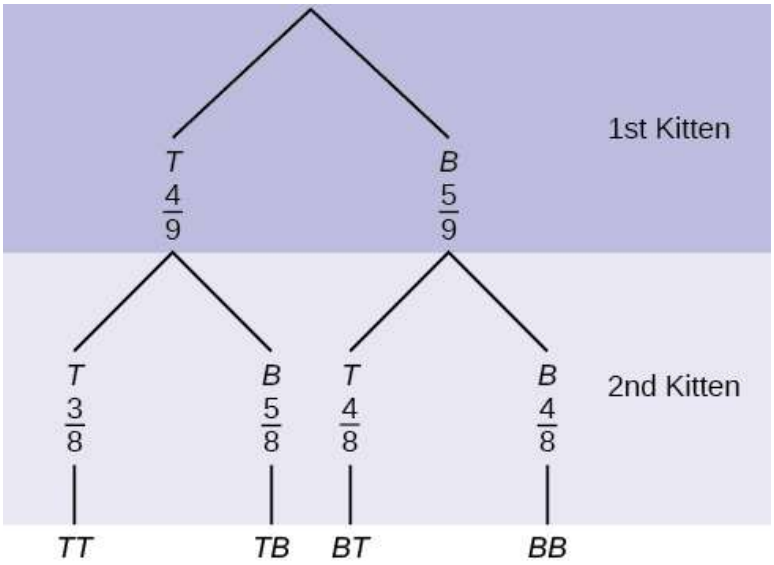
b. Find $P(N|F)$.

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.



a. Which shows the probability that both kittens are tabby?

a. $\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$ b. $\left(\frac{4}{9}\right)\left(\frac{4}{9}\right)$ c. $\left(\frac{4}{9}\right)\left(\frac{3}{8}\right)$ d. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$

b. What is the probability that one kitten of each coloring is selected?

a. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$ b. $\left(\frac{4}{9}\right)\left(\frac{5}{8}\right)$ c. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)+\left(\frac{5}{9}\right)\left(\frac{4}{9}\right)$ 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 3.26

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:

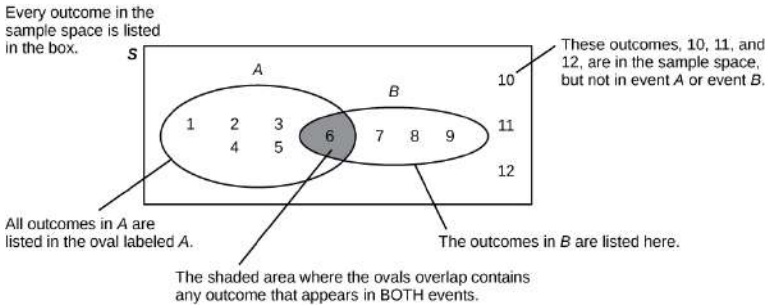


Figure 3.15

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:

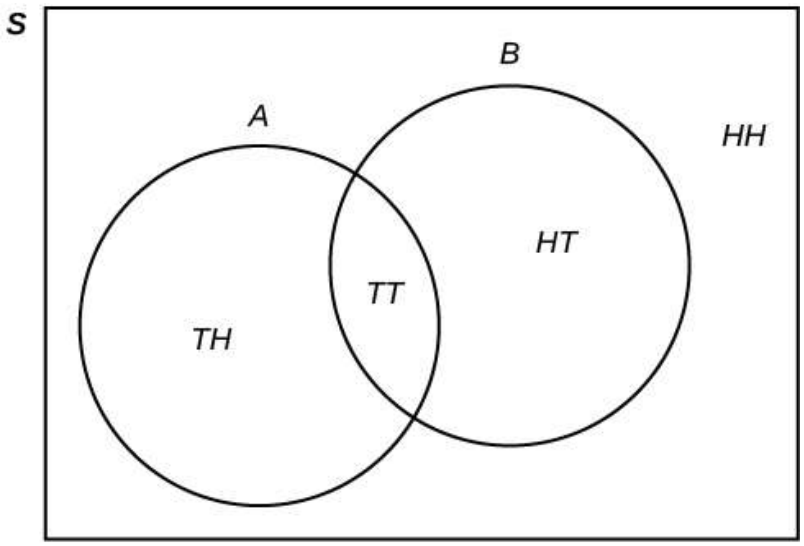


Figure 3.16

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

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.

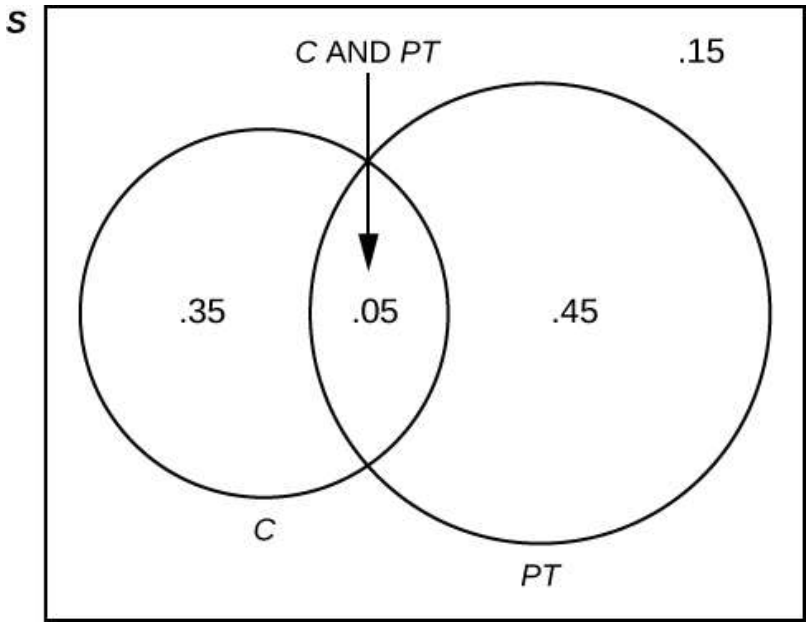


Figure 3.17

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 3.29

$P(C) = .40$

$P(PT) = .50$

$P(C \text{ AND } PT) = .05$

$$P(C|PT) = \frac{P(C \text{ AND } PT)}{P(PT)} = \frac{.05}{.50} = .1$$

$P(C \text{ OR } PT) = P(C) + P(PT) - P(C \text{ AND } PT) = .40 + .50 - .05 = .85$

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

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.

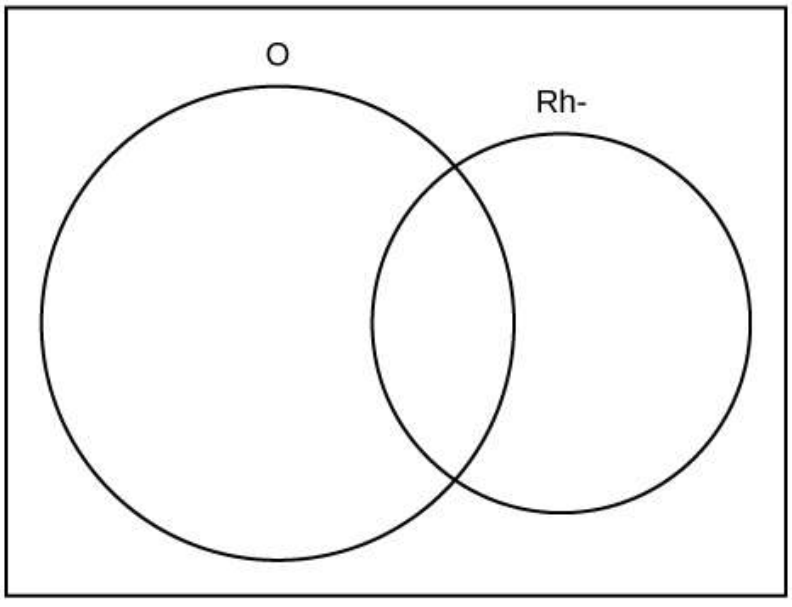


Figure 3.18

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 3.30

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

3.1 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 $G_2)$		
$P(G_2\|R_1)$		
P(no yellows)		
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 $G_2)$		
$P(G_2\|R_1)$		
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 $P(G_1 \text{ AND } R_2)$ and $P(R_1|G_2)$ 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.

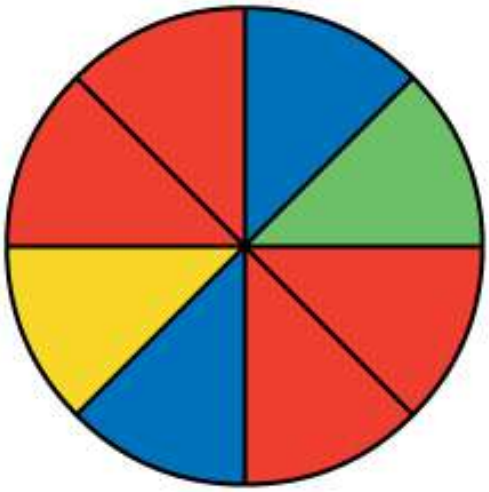


Figure 3.19

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.

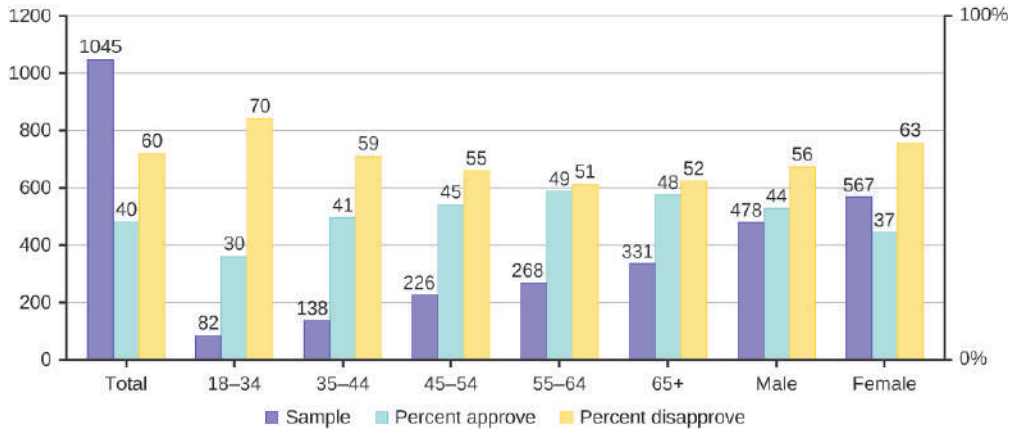


Figure 3.20 The graph in **Figure 3.20** displays the sample sizes and percentages of people in different age and gender groups who were polled concerning their approval of Mayor Ford's actions in office. The total number in the sample of all the age groups is 1,045.

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 $P(\text{Approve}|\text{Male})$.
i. Out of the age groups, what percent are more than 44 years old?
j. Find $P(\text{Approve}|\text{Age} < 35)$.

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.

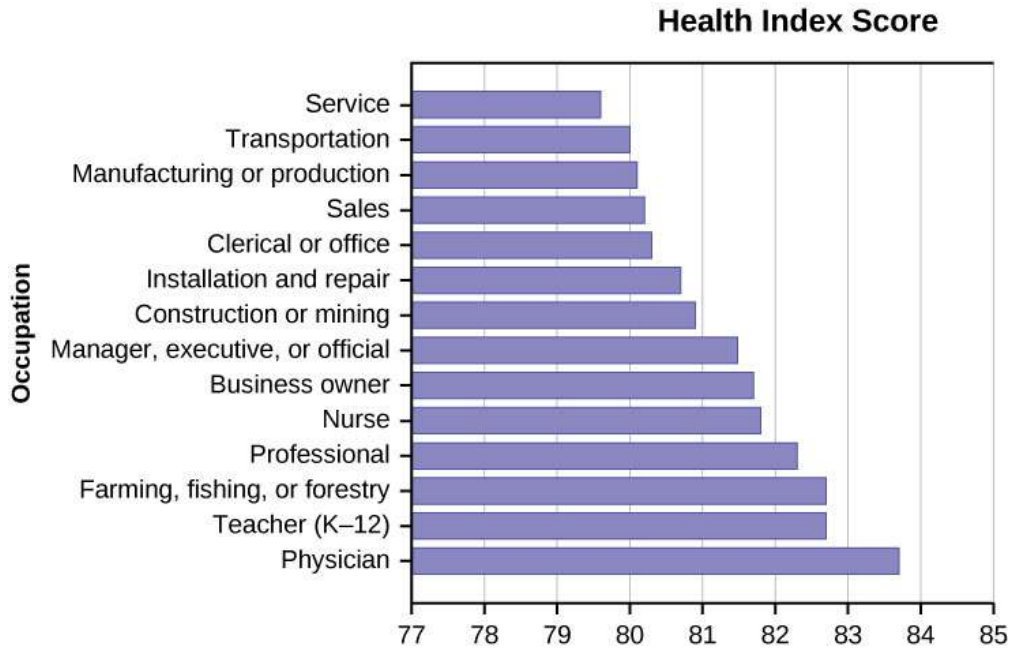


Figure 3.21

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*|*E*) = _____
d. *P*(*G* AND *E*) = _____
e. *P*(*G* 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*|*B*) represents. Find *P*(*A*|*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 $P(C|D) = .6$.

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 $P(D|C)$.

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: $P(H|G) = .4$.
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. $P(S\|E') =$	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(hit by Hank Aaron|hit is a double) = P(hit by Hank Aaron)
b. No, because P(hit by Hank Aaron|hit is a double) ≠ P(hit is a double)
c. No, because P(hit is by Hank Aaron|hit is a double) ≠ P(hit by Hank Aaron)
d. Yes, because P(hit is by Hank Aaron|hit is a double) = P(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 $P(E|D)$.
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	

Table 3.21

Up for Reelection:	Democratic Party	Republican Party	Other	Total
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.

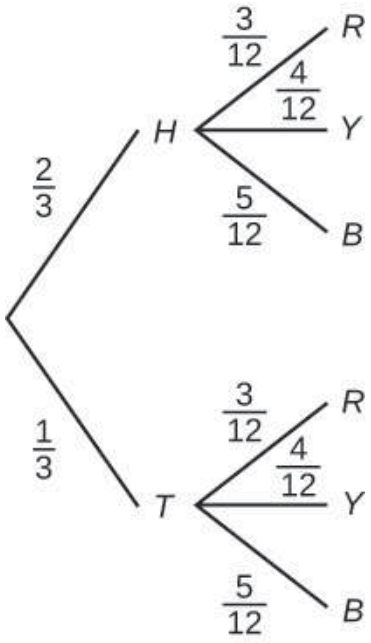


Figure 3.22

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(Shirt# 1–33|≤ 210 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. $P(P|C)$ = _____
c. $P(P|C')$ = _____
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 $P(G_2|G_1)$.
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 $P(G_2|G_1)$.
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. $P(M|L)$
e. $P(L|M)$
f. $P(S|F)$
g. $P(F|L)$
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 $P(H|I)$

29 $P(N|O)$

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 $C|L$ 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

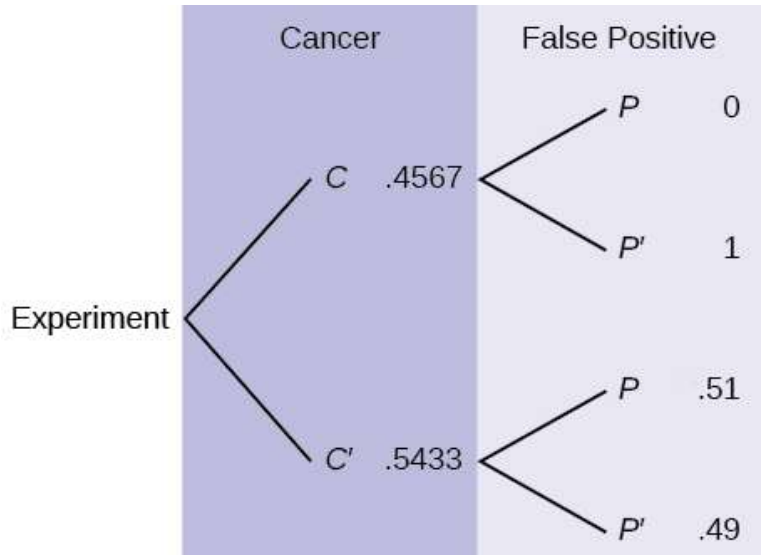


Figure 3.23

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 $P(\text{Occupation} < 81.3) = .5$

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. {$G1$, $G2$, $G3$, $G4$, $G5$, $Y1$, $Y2$, $Y3$}

b. $\frac{5}{8}$

c. $\frac{2}{3}$

d. $\frac{2}{8}$

e. $\frac{6}{8}$

f. No, because $P(G \text{ AND } E)$ 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. $P(A) = P(\text{blue})P(\text{head}) = \left(\frac{3}{10}\right)\left(\frac{1}{2}\right) = \frac{3}{20}$

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). $P(A \text{ AND } B) = 0$.

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). $P(A \text{ AND } C) = P(A) = \frac{3}{20.}$

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, $P(A \text{ AND } B) = 0$.

91

a. If Y and Z are independent, then $P(Y \text{ AND } Z) = P(Y)P(Z)$, so $P(Y \text{ OR } Z) = P(Y) + P(Z) – P(Y)P(Z)$.

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. $P(\text{type O OR Rh–}) = P(\text{type O}) + P(\text{Rh–}) – P(\text{type O AND Rh–})$

$0.52 = 0.43 + 0.15 – P(\text{type O AND Rh–})$; solve to find $P(\text{type O AND Rh–}) = .06$

6 percent of people have type O, Rh– blood

b. $P(\text{NOT(type O AND Rh–)}) = 1 – P(\text{type O AND Rh–}) = 1 – .06 = .94$

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. $P(\text{NEITHER chocolate NOR nuts}) = 1 – P(C \text{ OR } N) = 1 – .40 = .60$

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. $P(\text{NOT } (J \text{ AND } K)) = 1 - P(J \text{ AND } K) = 1 - 010 = .90$

c. $P(\text{NOT } (J \text{ OR } K)) = 1 - P(J \text{ OR } K) = 1 - .45 = .55$

120

a.

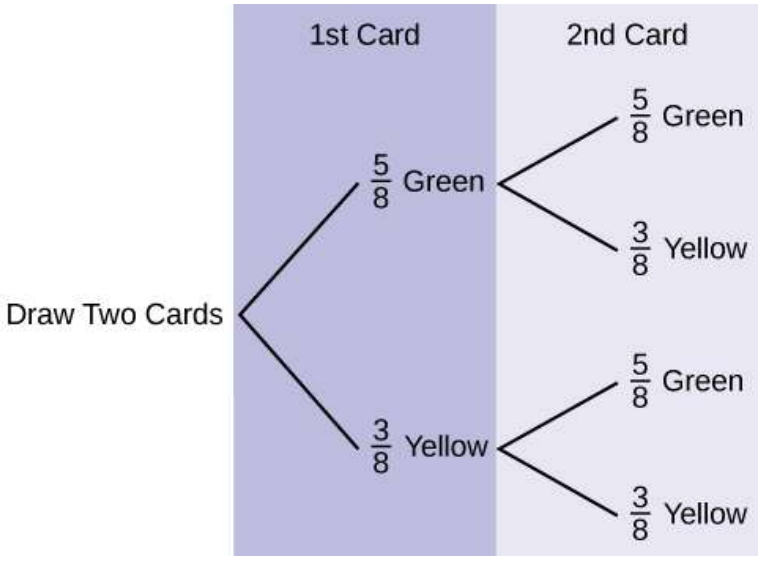


Figure 3.24

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. $P(G|G) = \frac{5}{8}$

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. $P(>64|F) = .1361$

d. $P(>64 \text{ and } F) = P(F)\,P(>64|F) = (.486)(.1361) = .0661$

e. $P(>64|F)$ 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: $P(\text{Alone}) = .7318 + .0390 = .7708$.

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.

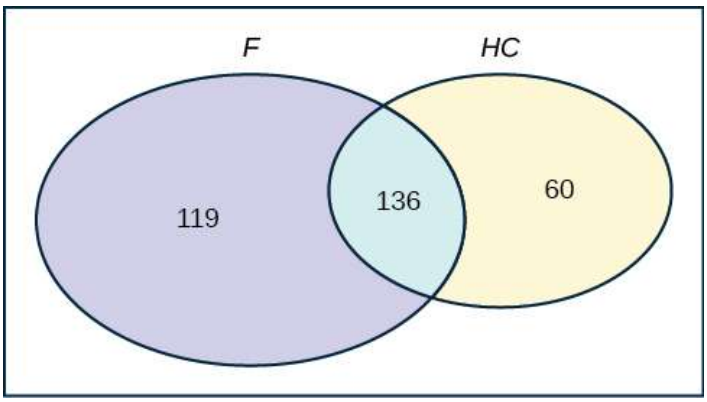


Figure 3.25

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)

Introduction

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.

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	$P(x = 2) = \frac{16}{50}$
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.

Describe the random variable in words. Let X = the number of days Nancy _____.

Solution 4.2

a. Let X = the number of days Nancy attends class per week.

b. In this example, what are possible values of X?

Solution 4.2

b. 0, 1, 2, and 3

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 4.2

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

$$\sigma^2 = \sum (x - \mu)^2 P(x).$$

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.

$$E(X) = \mu = \sum xP(x) = 0 + \frac{11}{50} + \frac{46}{50} + \frac{27}{50} + \frac{16}{50} + \frac{5}{50} = \frac{105}{50} = 2.1$$

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 $(x-\mu)^2$ and the fifth column $(x - \mu)^2 \cdot P(x)$ to get the following table:

x	$P(x)$	$xP(x)$	$(x-\mu)^2$	$(x-\mu)^2 \cdot P(x)$
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.

$$\sigma^2 = .1764 + .2662 + .0046 + .1458 + .2888 + .1682 = 1.05$$

To get the standard deviation σ, we simply take the square root of variance σ^2.

$$\sigma = \sqrt{\sigma^2} = \sqrt{1.05} \approx 1.0247$$

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?

a. Define a random variable X.

Solution 4.6

a. X = amount of profit

b. Complete the following expected value table.

	x	____	____
WIN	10	$\frac{1}{3}$	____
LOSE	____	____	$\frac{-12}{3}$

Table 4.10

Solution 4.6

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

c. What is the expected value, μ? Do you come out ahead?

Solution 4.6

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. $P(\text{red}) = \frac{2}{5}$, $P(\text{blue}) = \frac{2}{5}$, and $P(\text{green}) = \frac{1}{5}$. 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

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 4.7

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

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 4.10

$P(x > 10)$

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.

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 4.11
a. failure

b. If we are interested in the number of students who do their homework on time, then how do we define X?

Solution 4.11
b. X = the number of statistics students who do their homework on time

c. What values does x take on?

Solution 4.11
c. 0, 1, 2, . . ., 50

d. What is a *failure*, in words?

Solution 4.11

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 $n = 50$.

e. If $p + q = 1$, then what is q?

Solution 4.11
e. $q = .30$

f. The words *at least* translate as what kind of inequality for the probability question $P(x$ ____ $40)$?

Solution 4.11
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

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 4.13

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. $P(x = 2) = \text{binompdf}\left(100, \frac{8}{560}, 2\right) = .2466$

   ii. $P(x \leq 6) = \text{binomcdf}\left(100, \frac{8}{560}, 6\right) = .9994$

iii. $P(x > 3) = 1 - P(x \leq 3) = 1 - \text{binomcdf}\left(100, \frac{8}{560}, 3\right) = 1 - .9443 = .0557$

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.

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 4.14

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**:
$P(x \leq 8) = \text{binomcdf}(200, .0128, 8) = .9988$

d. $P(x = 5) = \text{binompdf}(200, .0128, 5) = .0707$

$P(x = 6)$ = 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.

$$\mu = \frac{1}{p},\ \sigma^2 = \left(\frac{1}{p}\right)\left(\frac{1}{p} - 1\right),\ \sigma = \sqrt{\left(\frac{1}{p}\right)\left(\frac{1}{p} - 1\right)}$$

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).

a. Let X = the number of _____ you must ask _____ one says yes.

Solution 4.18
a. Let X = the number of *students* you must ask *until* one says yes.

b. What values does X take on?

Solution 4.18
b. 1, 2, 3, . . ., (total number of students)

c. What are p and q?

Solution 4.18
c. $p = .55$; $q = .45$

d. The probability question is P(_____).

Solution 4.18
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* = 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* ≤ 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

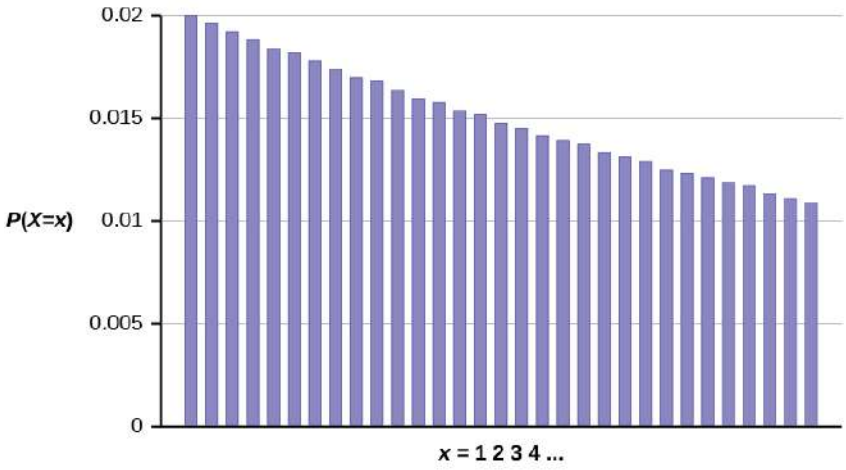


Figure 4.2

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

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 4.20

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

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 4.21
$P(x = 35)$

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

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, . . . , 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 4.22
$P(x \leq 2)$

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).

a. Are you choosing with or without replacement?

Solution 4.23
a. without

b. What is the group of interest?

Solution 4.23
b. the men

c. How many are in the group of interest?

Solution 4.23
c. 15 men

d. How many are in the other group?

Solution 4.23
d. 18 women

e. Let X = _____ on the committee. What values does X take on?

Solution 4.23
e. Let X = *the number of men* on the committee. $x = 0, 1, 2, \ldots, 7$.

f. The probability question is P(_____).

Solution 4.23
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

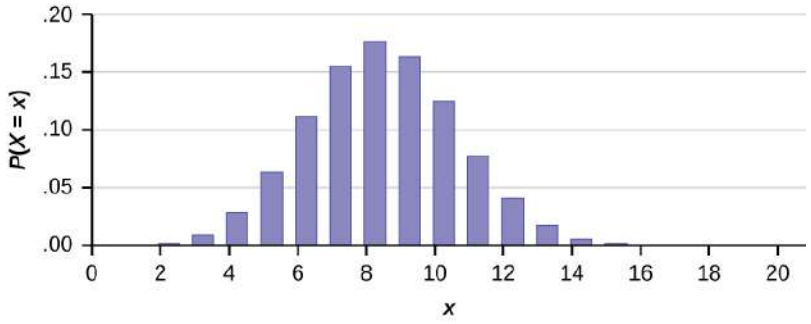


Figure 4.3

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

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 4.26

$P(x < 5)$

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.

a. What is the interval of interest?

Solution 4.27

a. one broadcast

b. What is the average number of times the news reporter says "uh" during one broadcast?

Solution 4.27

b. 2

c. Let X = _____. What values does X take on?

Solution 4.27
c. Let X = the number of times the news reporter says "uh" during one broadcast.
x = 0, 1, 2, 3, . . .

d. The probability question is P(_____).

Solution 4.27
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 2nd 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

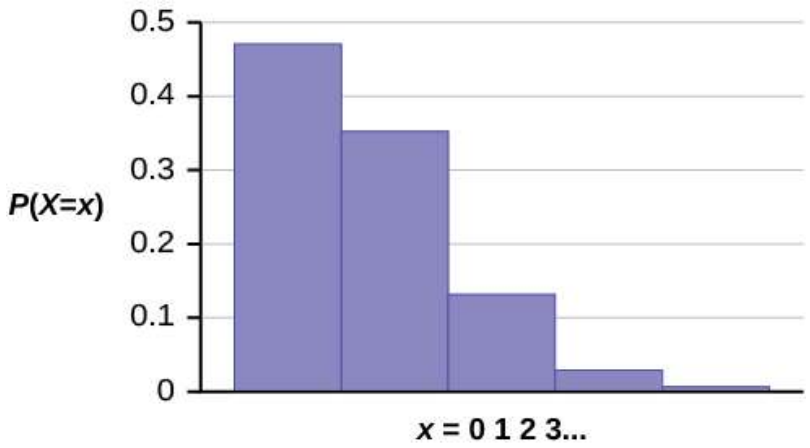


Figure 4.4

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.

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 4.29

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 Σ

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 $x = 0, 1, 2 \ldots$. 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.

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 4.30

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 Σ

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

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 4.31

Let X = the number of days with low seismic activity.

Using the binomial distribution

- $P(x = 10)$ = binompdf(200, .0102, 10) ≈ .000039

Using the Poisson distribution

- Calculate $\mu = np = 200(.0102) \approx 2.04$
- $P(x = 10) = \text{poissonpdf}(2.04, 10) \approx .000045$

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

4.1 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	_____	_____
3	_____	_____
4	_____	_____
5	_____	_____

x	Frequency	Relative Frequency
6	_____	_____
7	_____	_____
8	_____	_____
9	_____	_____
10	_____	_____

Table 4.15

2. Calculate the following:
   a. $\bar{x}$ = _____
   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	_____	_____
4	_____	_____
5	_____	_____
6	_____	_____

Table 4.16

X	Frequency	Relative Frequency
7	_____	_____
8	_____	_____
9	_____	_____
10	_____	_____

Table 4.16

2. Calculate the following:
   a. $\bar{x}$ = _____
   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	
5	
6	
7	
8	

x	*P*(*x*)
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

4.2 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. $\bar{x}$ = _____
   b. s_x = _____
   c. $\bar{y}$ = _____
   d. s_y = _____
3. Explain what $\bar{x}$ represents.

4. Explain what $\bar{y}$ 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.

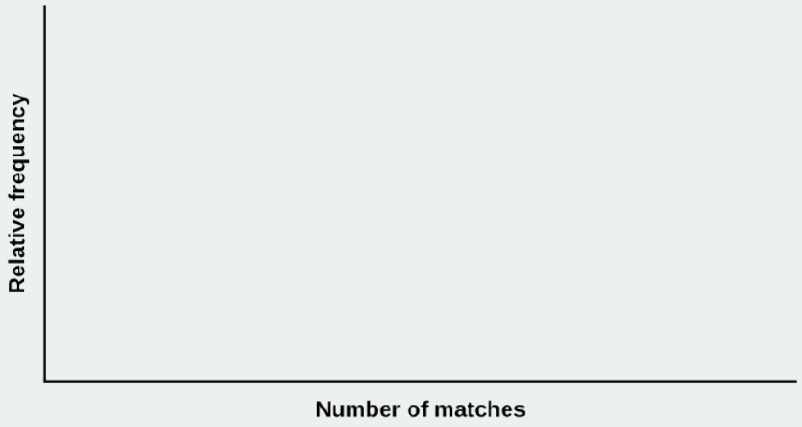


Figure 4.8

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. $\mu_x =$ _____
   b. $\sigma_x =$ _____
   c. $\mu_x =$ _____
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.

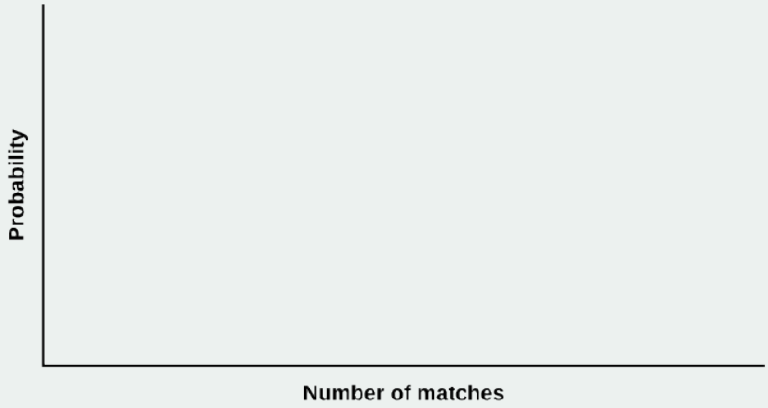


Figure 4.9

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 foreve; 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 $\bar{x}$) is $\bar{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$ 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 $X, Y, Z, \ldots$; 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 $\sigma = \sqrt{\frac{rbn(r+b-n)}{(r+b)^2\,(r+b-1)}}$.

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
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
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	
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. $X \sim$ _____(_____,_____)

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
2	.24
3	
4	.70
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. $X \sim$ _____(_____,_____)
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. $X \sim$ _____(_____,_____)
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. $X \sim$ _____(_____,_____)
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. $X \sim$ _____(_____,_____)
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. $X \sim$ _____(_____,_____)
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. $X \sim$ _____(_____,_____)
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. $X \sim$ _____(_____,_____)
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. $X \sim$ _____(_____,_____)
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. $X \sim$ _____(_____,_____)
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. $X \sim$ _____(_____,_____)
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. $X \sim$ _____
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. $X \sim$ _____(_____,_____)
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. $X \sim$ _____(_____,_____)
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. $X \sim$ _____(_____,_____)
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. $X \sim$ _____(_____,_____)
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 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. $X \sim$ _____(_____,_____)
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

- 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	$\frac{250}{10{,}000} = .025$

Table 4.45

x	$P(x)$
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. $X \sim B(15, .281)$

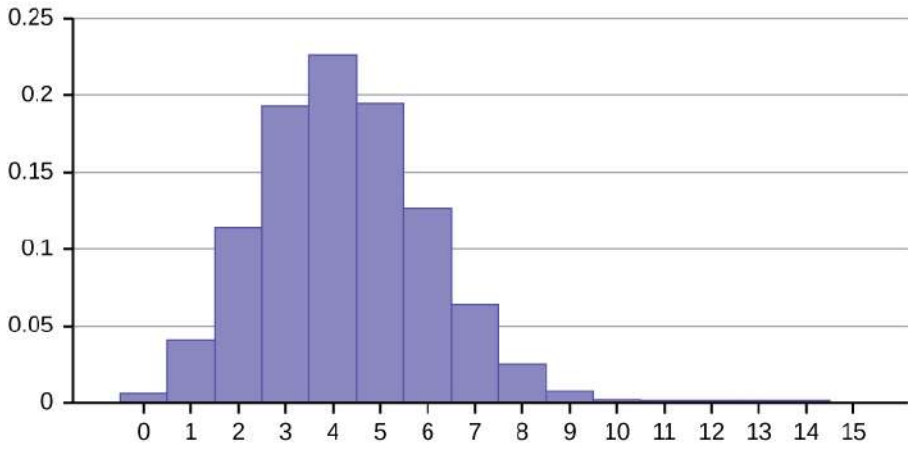


Figure 4.10

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. $X \sim P(2)$

d. 2

e. .1353

f. .3233

127

a. X = the number of shell pieces in one cake

b. 0, 1, 2, 3,...

c. $X \sim P(1.5)$

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)

Introduction

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

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:

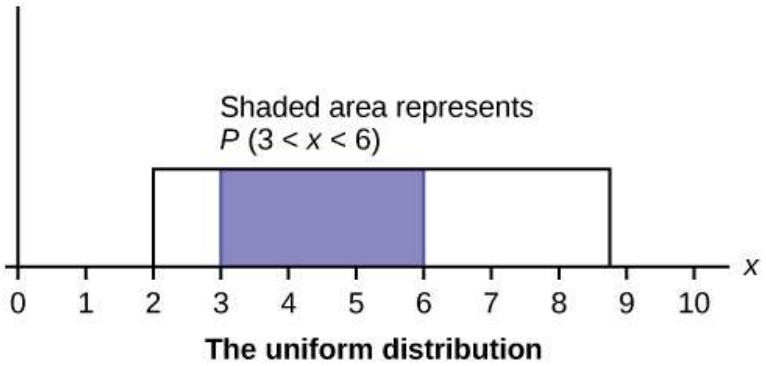


Figure 5.2 The graph shows a uniform distribution with the area between $x = 3$ and $x = 6$ shaded to represent the probability that the value of the random variable X is in the interval between three and six.

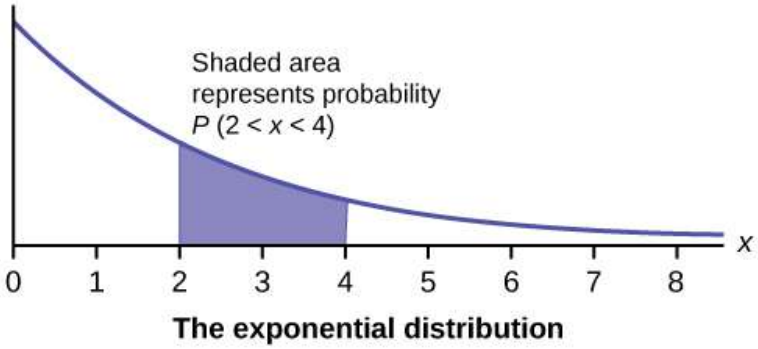


Figure 5.3 The graph shows an exponential distribution with the area between $x = 2$ and $x = 4$ shaded to represent the probability that the value of the random variable X is in the interval between two and four.

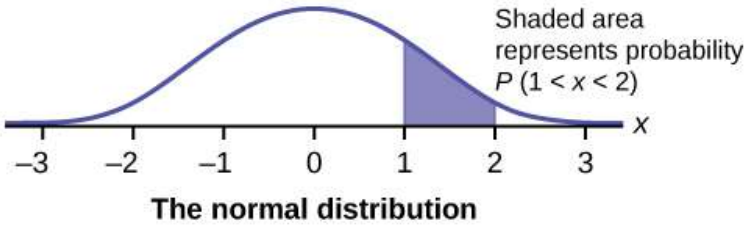


Figure 5.4 The graph shows the standard normal distribution with the area between $x = 1$ and $x = 2$ shaded to represent the probability that the value of the random variable X is in the interval between one and two.

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 $0 \le x \le 20$. x = a real number. The graph of $f(x) = \frac{1}{20}$ is a horizontal line.

However, since $0 \le x \le 20$, $f(x)$ is restricted to the portion between $x = 0$ and $x = 20$, inclusive.

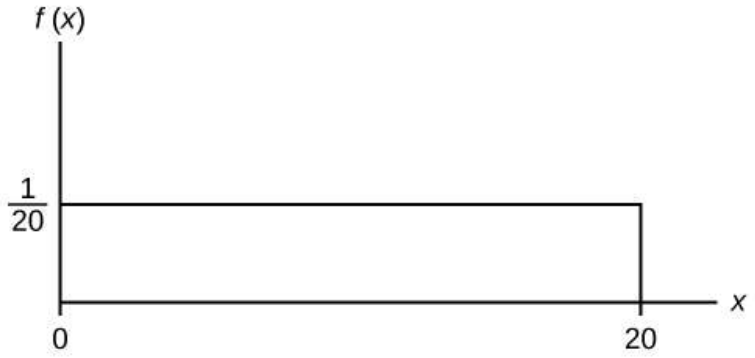


Figure 5.5

$f(x) = \frac{1}{20}$ **for** $0 \le x \le 20$.

The graph of $f(x) = \frac{1}{20}$ is a horizontal line segment when $0 \le x \le 20$.

The area between $f(x) = \frac{1}{20}$ where $0 \le x \le 20$ 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 $0 < x < 2$.

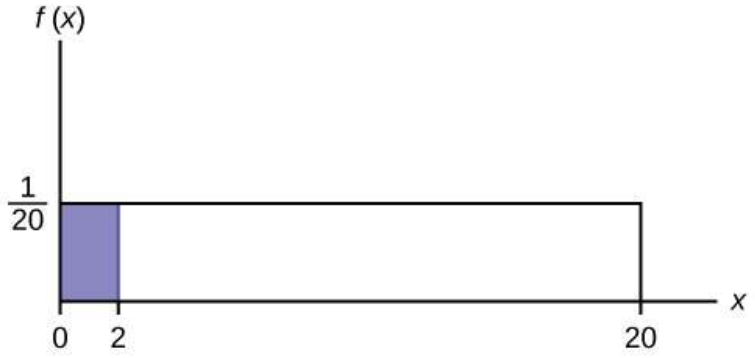


Figure 5.6

$\text{AREA} = (2 - 0)\left(\frac{1}{20}\right) = 0.1$

$(2 - 0) = 2 =$ 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 $4 < x < 15$.

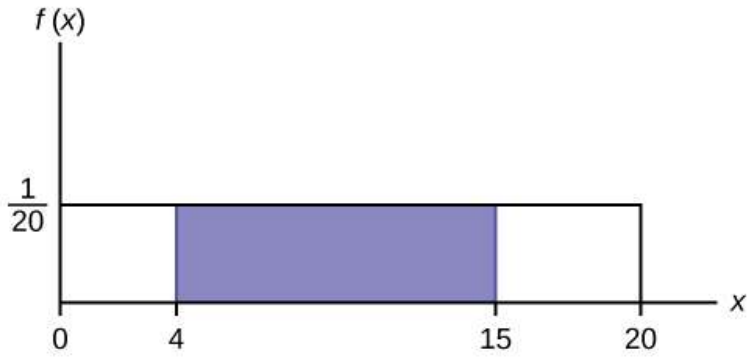


Figure 5.7

$$\text{AREA} = (15 - 4)\left(\frac{1}{20}\right) = 0.55$$

$(15 - 4) = 11 =$ 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, $P(x = 15) = (\text{base})(\text{height}) = (0)\left(\frac{1}{20}\right) = 0$

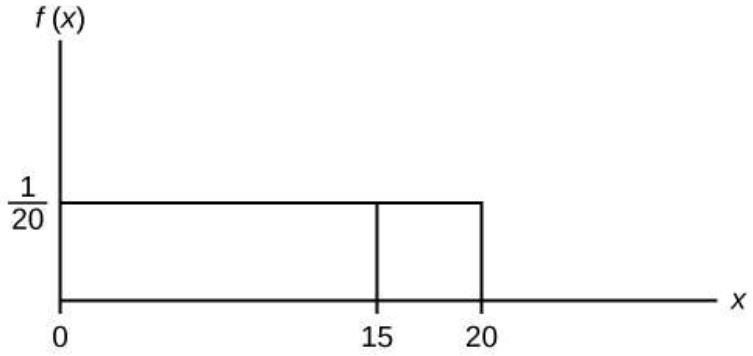


Figure 5.8

$P(X <= x)$, 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: $P(X > x) = 1 - P(X < x)$.

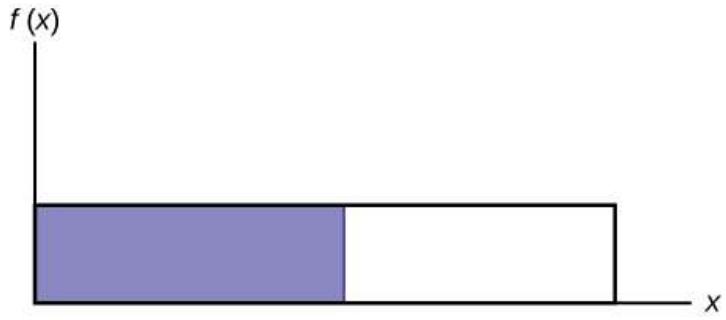


Figure 5.9

Label the graph with $f(x)$ and x. Scale the x and y axes with the maximum x and y values. $f(x) = \frac{1}{20}$, $0 \le x \le 20$.

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.

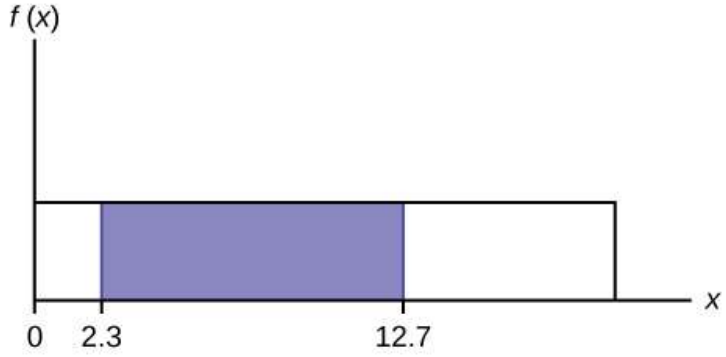


Figure 5.10

$$P(2.3 < x < 12.7) = (\text{base})(\text{height}) = (12.7 - 2.3)\left(\frac{1}{20}\right) = 0.52$$

Try It

5.1 Consider the function $f(x) = \frac{1}{8}$ for $0 \le x \le 8$. 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

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 5.3

$P(2 < x < 18) = (\text{base})(\text{height}) = (18 - 2)\left(\frac{1}{23}\right) = \frac{16}{23}$

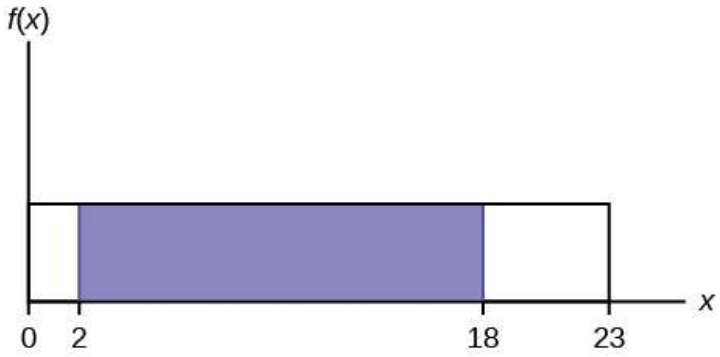


Figure 5.11

b. Find the 90th percentile for an eight-week-old baby's smiling time.

Solution 5.3

b. Ninety percent of the smiling times fall below the 90th percentile, k, so $P(x < k) = 0.90$.

$$P(x < k) = 0.90$$
$$(\text{base})(\text{height}) = 0.90$$
$$(k - 0)\left(\frac{1}{23}\right) = 0.90$$

$$k = (23)(0.90) = 20.7$$

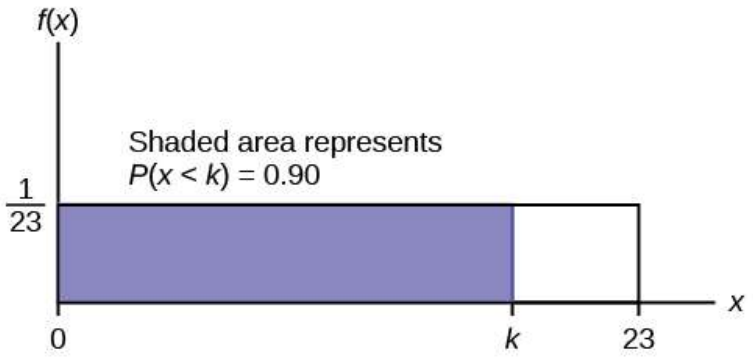


Figure 5.12

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 5.3

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 $P(x > 12|x > 8)$ 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 $8 < x < 23$.

$P(x > 12|x > 8) = (23 - 12)\left(\frac{1}{15}\right) = \frac{11}{15}$

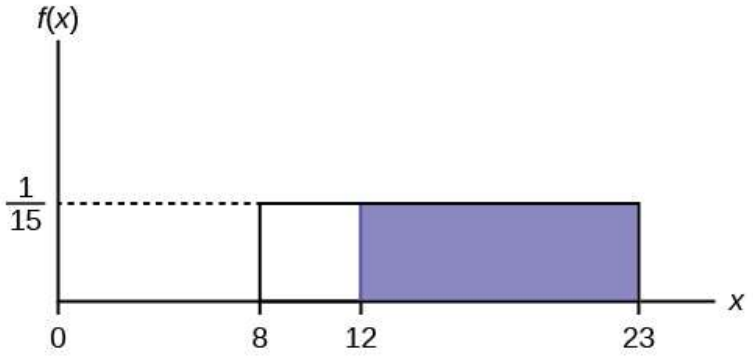


Figure 5.13

For the second way, use the conditional formula from **Probability Topics** with the original distribution.

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)}$$

For this problem, A is $(x > 12)$ and B is $(x > 8)$.

So, $P(x > 12|x > 8) = \frac{(x > 12 \text{ AND } x > 8)}{P(x > 8)} = \frac{P(x > 12)}{P(x > 8)} = \frac{\frac{11}{23}}{\frac{15}{23}} = \frac{11}{15}$

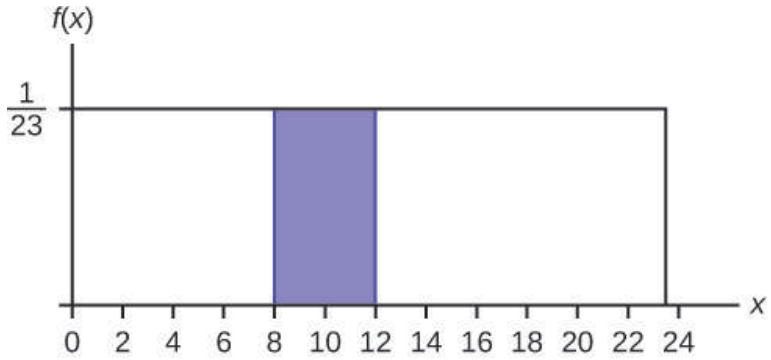


Figure 5.14

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.

a. What is the probability that a person waits fewer than 12.5 minutes?

Solution 5.4

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.

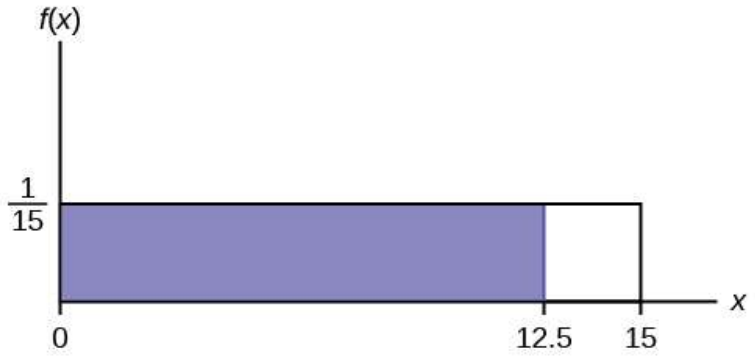


Figure 5.15

b. On the average, how long must a person wait? Find the mean, μ, and the standard deviation, σ.

Solution 5.4

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.

c. Ninety percent of the time, the minutes a person must wait falls below what value?

This question asks for the 90th percentile.

Solution 5.4

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.

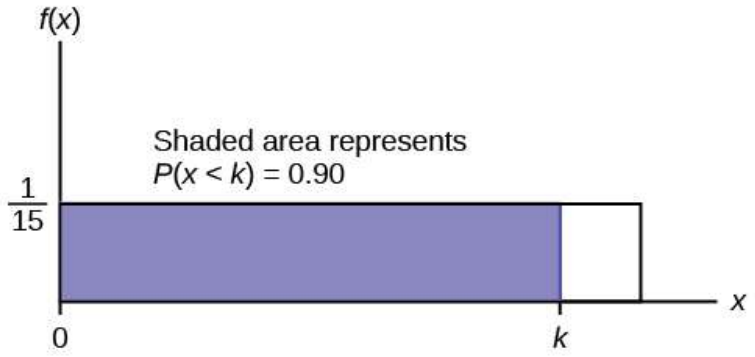


Figure 5.16

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 $X \sim U(0.5, 4)$.

a. The probability that a randomly selected nine-year-old child eats a doughnut in at least two minutes is _____.

Solution 5.5

a. 0.5714

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 $P(x > 2|x > 1.5)$. Draw a graph.

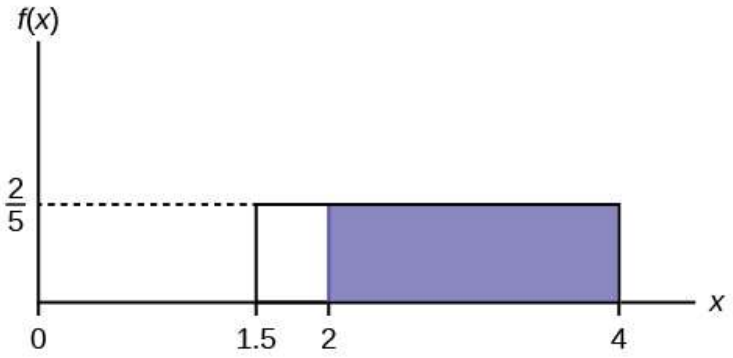


Figure 5.17

$$P(x > 2 | x > 1.5) = (\text{base})(\text{new height}) = (4 - 2)\left(\frac{2}{5}\right) = \frac{4}{5}$$

Solution 5.5

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

$$P(x > 2 | x > 1.5) = \frac{P(x > 2 \quad AND \quad x > 1.5)}{P(x > 1.5)} = \frac{P(x > 2)}{P(x > 1.5)} = \frac{\frac{2}{3.5}}{\frac{2.5}{3.5}} = 0.8 = \frac{4}{5}$$

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 $x \sim U(1.5, 4)$.

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 30th 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 5.6

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

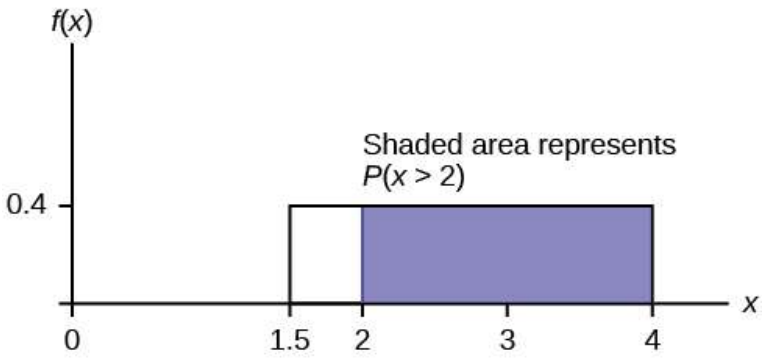


Figure 5.18 Uniform distribution between 1.5 and four with shaded area between two and four representing the probability that the repair time x is greater than two

Solution 5.6

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.

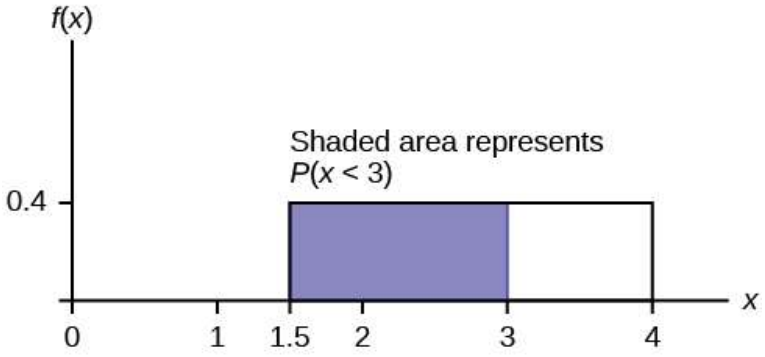


Figure 5.19 Uniform distribution between 1.5 and four with shaded area between 1.5 and three representing the probability that the repair time x is less than three

Solution 5.6

c.

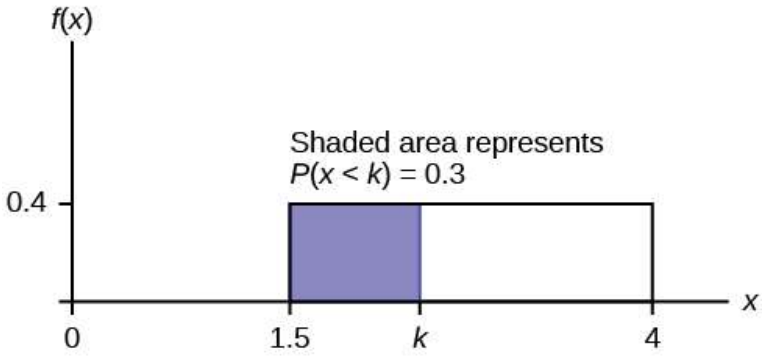


Figure 5.20 Uniform distribution between 1.5 and 4 with an area of 0.30 shaded to the left, representing the shortest 30 percent of repair times.

$P(x < k) = 0.30$
$P(x < k)$ = (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 5.6

d.

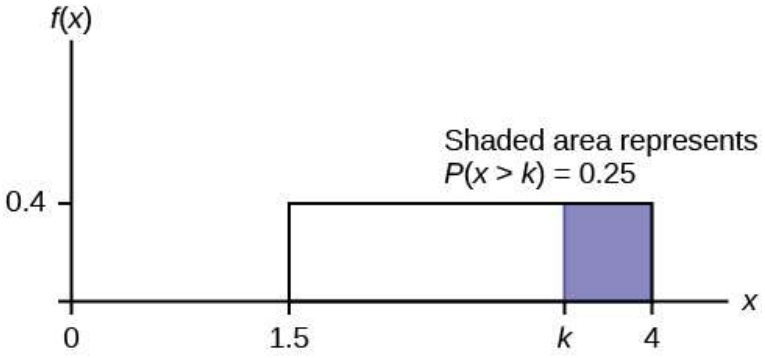


Figure 5.21 Uniform distribution between 1.5 and 4 with an area of 0.25 shaded to the right representing the longest 25 percent of repair times.

$P(x > k) = 0.25$
$P(x > k)$ = (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 5.6

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 50th 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 $e = 2.71828182846...$ 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:

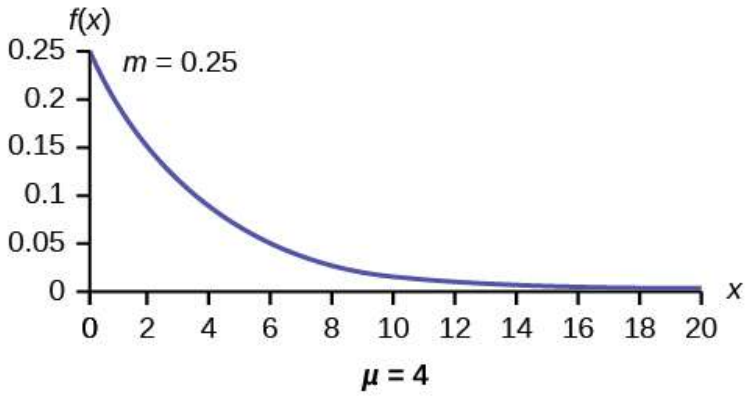


Figure 5.22

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

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 5.8

a. Find $P(4 < x < 5)$. The **cumulative distribution function (CDF)** gives the area to the left.

$$P(x < x) = 1 - e^{-mx}$$

$$P(x < 5) = 1 - e^{(-0.25)(5)} = 0.7135 \text{ and } P(x < 4) = 1 - e^{(-0.25)(4)} = 0.6321$$

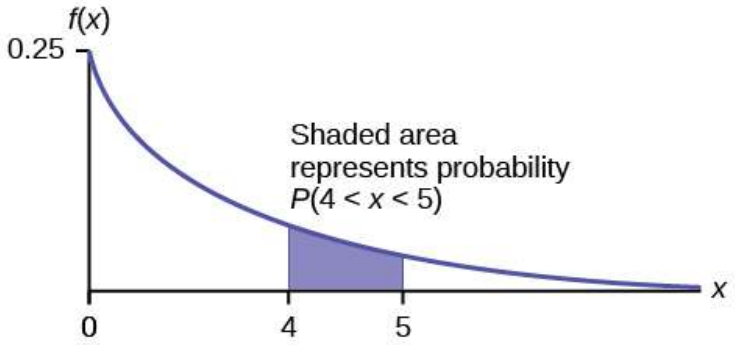


Figure 5.23

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).

b. Half of all customers are finished within how long? (Find the 50th percentile).

Solution 5.8

b. Find the 50th percentile.

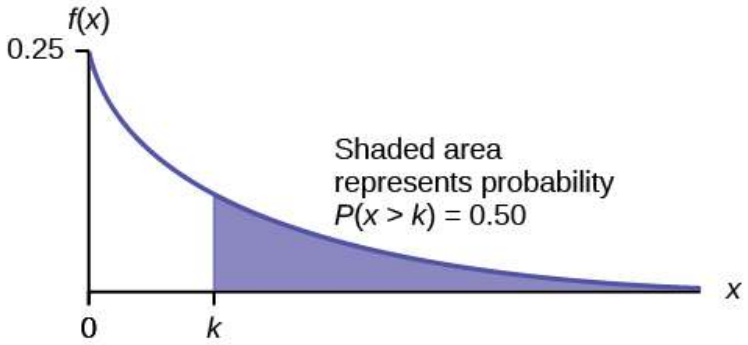


Figure 5.24

$P(x < k) = 0.50$, $k = 2.8$ 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.

c. Which is larger, the mean or the median?

Solution 5.8

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.

a. What is the probability that a computer part lasts more than seven years?

Solution 5.9

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.

$$P(x > 7) \quad = \quad 1 \quad - P(x < 7).$$

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).

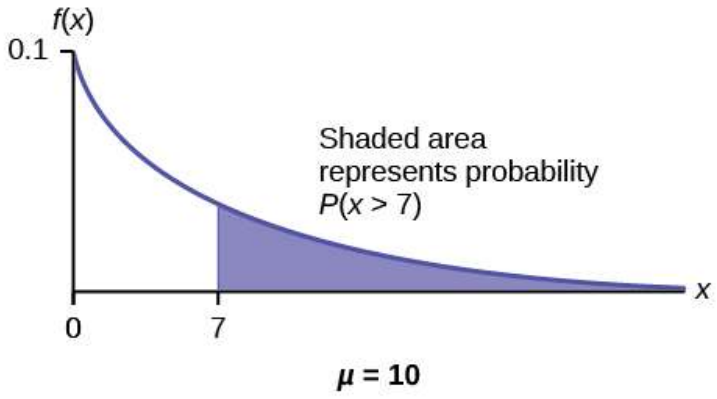


Figure 5.25

b. On the average, how long would five computer parts last if they are used one after another?

Solution 5.9

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.

c. Eighty percent of computer parts last at most how long?

Solution 5.9

c. Find the 80th percentile. Draw the graph. Let k = the 80th percentile.

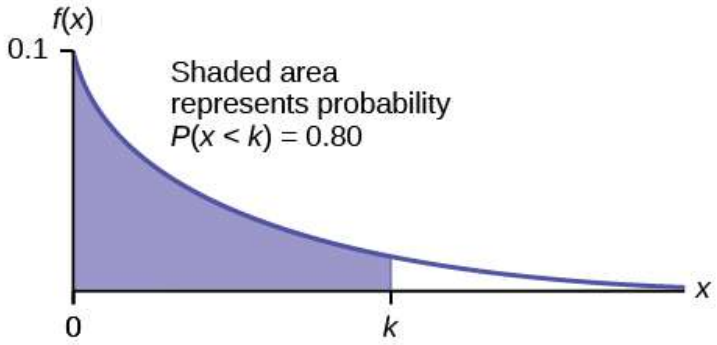


Figure 5.26

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 $\frac{\ln(1 - 0.80)}{-0.1}$.

d. What is the probability that a computer part lasts between nine and 11 years?

Solution 5.9

d. Find $P(9 < x < 11)$. Draw the graph.

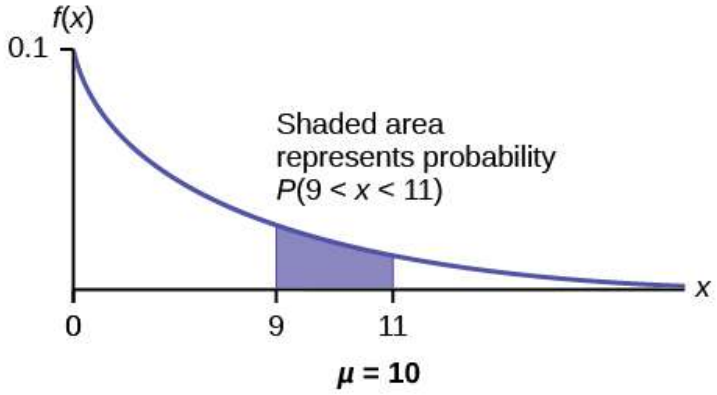


Figure 5.27

$P(9 < x < 11) = P(x < 11) - P(x < 9) = (1 - e^{(-0.1)(11)}) - (1 - e^{(-0.1)(9)}) = 0.6671 - 0.5934 = 0.0737$. 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.

What is m, μ, and σ? The probability that you must wait more than five minutes is _____ .

Solution 5.10

- $m = \frac{1}{12}$
- $\mu = 12$
- $\sigma = 12$

$P(x > 5) = 0.6592$

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.

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 5.11

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 $P(X < 1) = 1 - e^{(-0.5)(1)} \approx 0.3935$.

1 - e^(–0.5) ≈ 0.3935

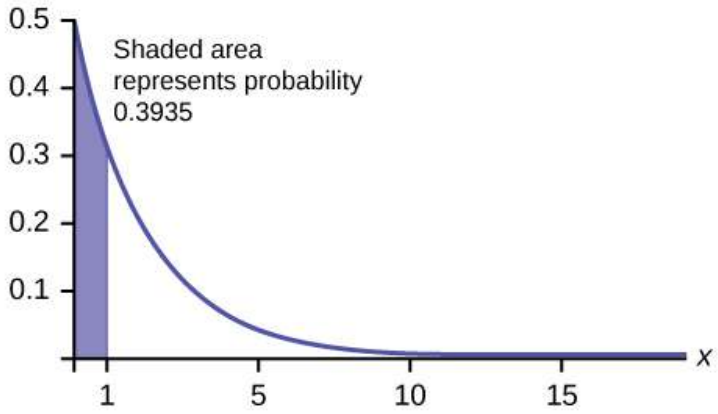


Figure 5.28

d. $P(X > 5) = 1 - P(X < 5) = 1 - (1 - e^{(-0.5)(5)}) = e^{-2.5} \approx 0.0821$.

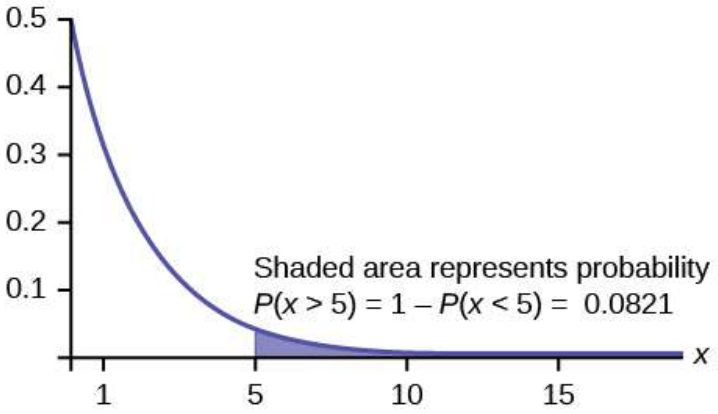


Figure 5.29

$$1 - (1 - e\text{^}(-0.50)(5)) \text{ or } e\text{^}(-0.50)(5)$$

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

$$k = \frac{ln(1 - Area\_To\_The\_Left\_Of\_k)}{(-m)}$$

$$k = \frac{ln(1 - 0.70)}{(-0.5)} \approx 2.41 \text{ minutes}$$

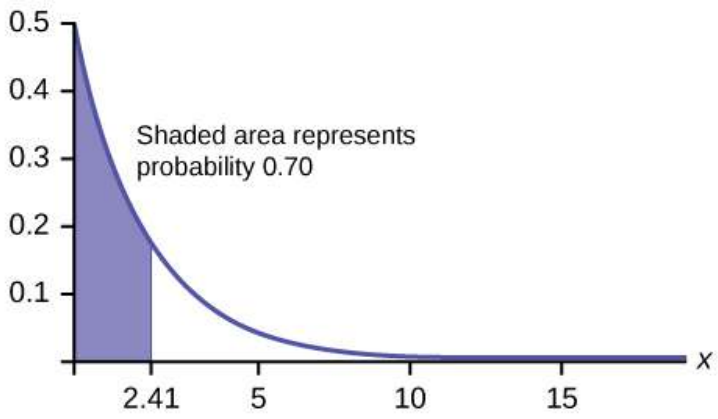


Figure 5.30

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 $P(X > 17|X > 10) = P(X > 7) = 0.4966$.

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 $P(X > 7|X > 4)$. The **memoryless property** says that $P(X > 7|X > 4) = P(X > 3)$, so we just need to find the probability that a customer spends more than three minutes with a postal clerk.

This is $P(X > 3) = 1 - P(X < 3) = 1 - (1 - e^{-0.25 \cdot 3}) = e^{-0.75} \approx 0.4724$.

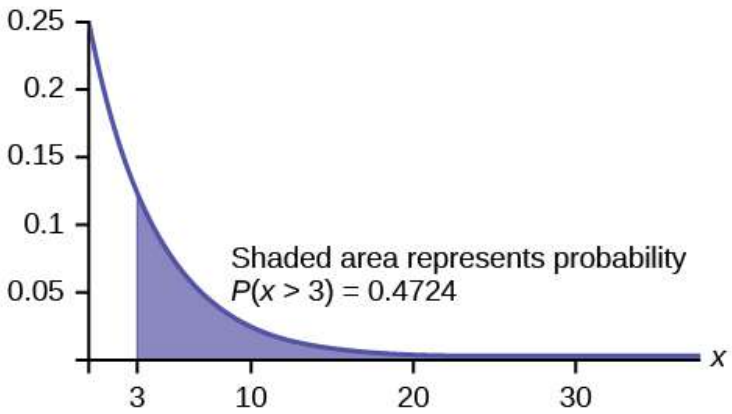


Figure 5.31

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 2^{nd}, VARS(DISTR), C: poissonpdf(λ, k). To compute $P(X \leq k)$, enter 2^{nd}, 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.

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 5.13

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 $P\left(T < \frac{1}{6}\right) = 1 – e^{(-4)\left(\frac{1}{6}\right)} \approx 0.4866$.

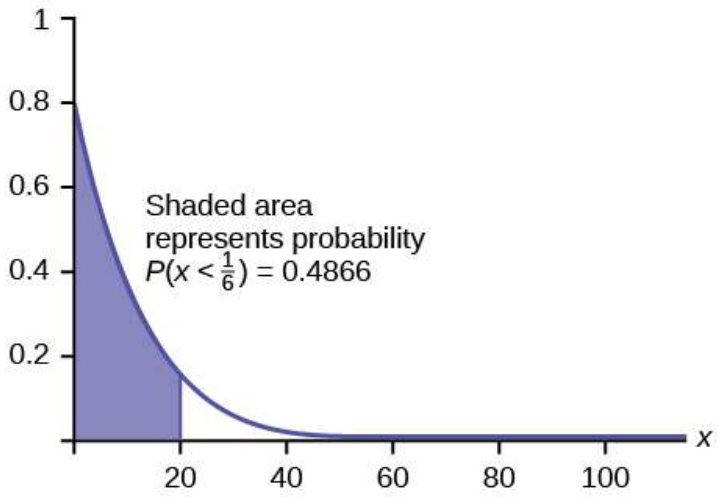


Figure 5.32

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))

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 $P(X \leq 4) = 0.6288$.

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, $P(Y > 40) = 1 - P(Y \leq 40) = 1 - 0.9294 = 0.0706$.

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

5.1 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. $\bar{x}$ = _____
   b. s = _____
   c. first quartile = _____
   d. third quartile = _____
   e. median = _____

Organize the Data

1. Construct a histogram of the empirical data. Make eight bars.

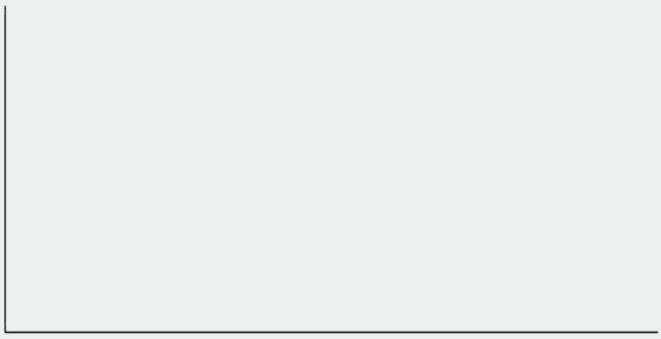


Figure 5.33

2. Construct a histogram of the empirical data. Make five bars.

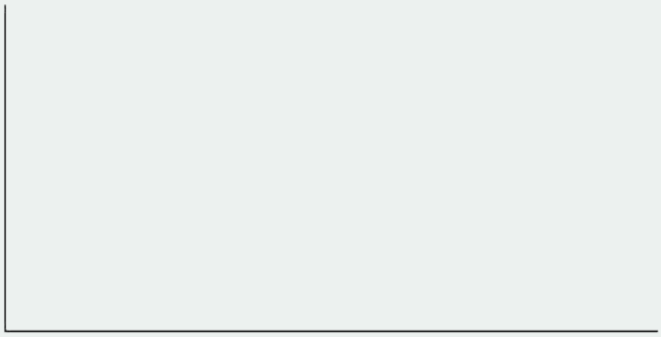


Figure 5.34

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 $P(a < x < b)$.

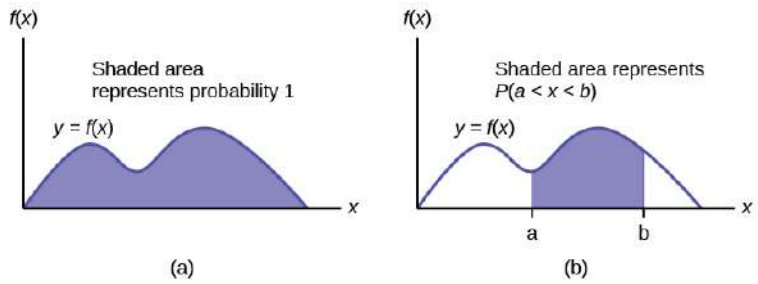


Figure 5.35

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.

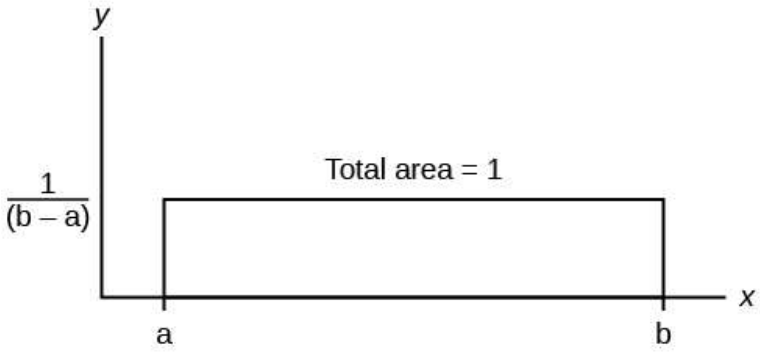


Figure 5.36

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 $\sigma = \sqrt{\frac{(b-a)^2}{12}}$
- $P(c < X < d) = (d - c)\left(\frac{1}{b-a}\right)$

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
- $k! = k*(k-1)*(k-2)*(k-3)*\ldots 3*2*1$

PRACTICE

5.1 Continuous Probability Functions

1. Which type of distribution does the graph illustrate?

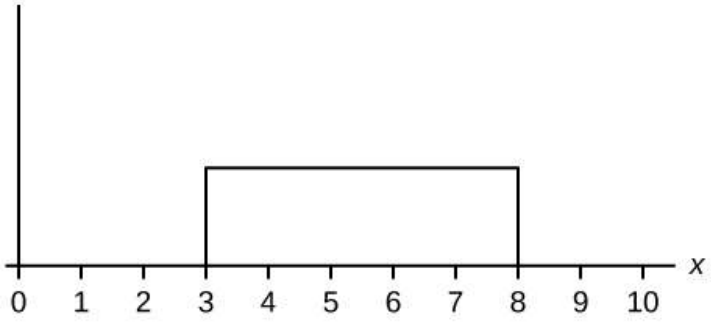


Figure 5.37

2. Which type of distribution does the graph illustrate?

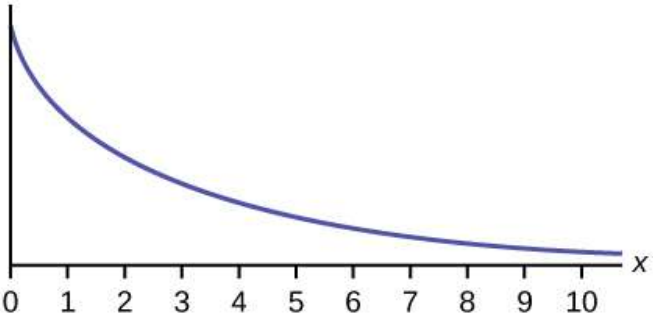


Figure 5.38

3. Which type of distribution does the graph illustrate?

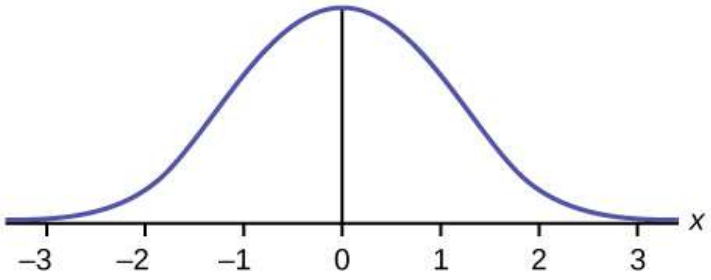


Figure 5.39

4. What does the shaded area represent? $P(\_\_\_ < x < \_\_\_)$

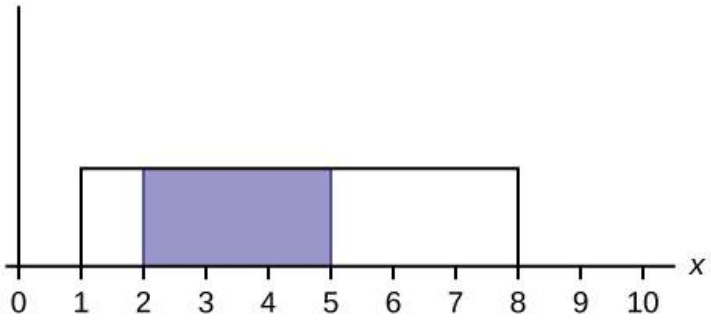


Figure 5.40

5. What does the shaded area represent? $P(\_\_\_ < x < \_\_\_)$

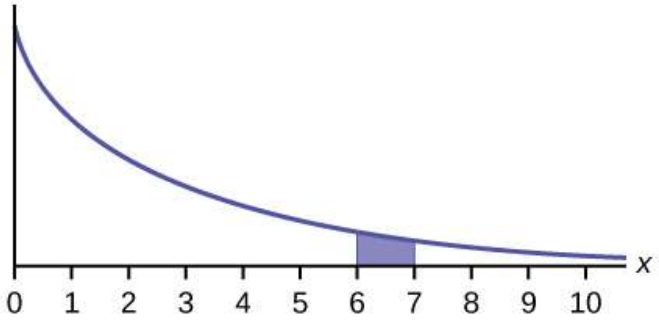


Figure 5.41

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.

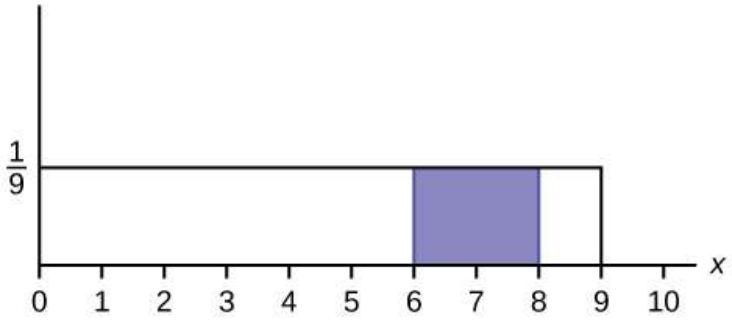


Figure 5.42

13. Find the probability that x falls in the shaded area.

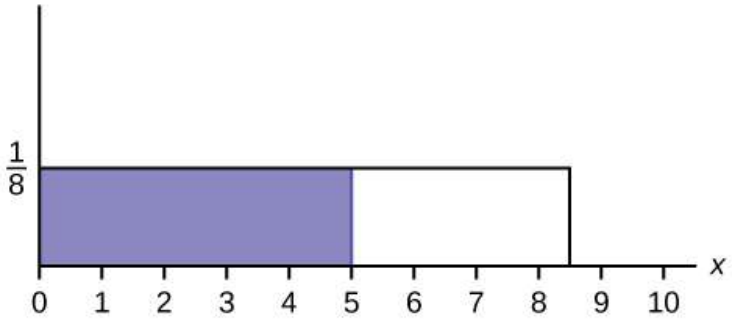


Figure 5.43

14. Find the probability that x falls in the shaded area.

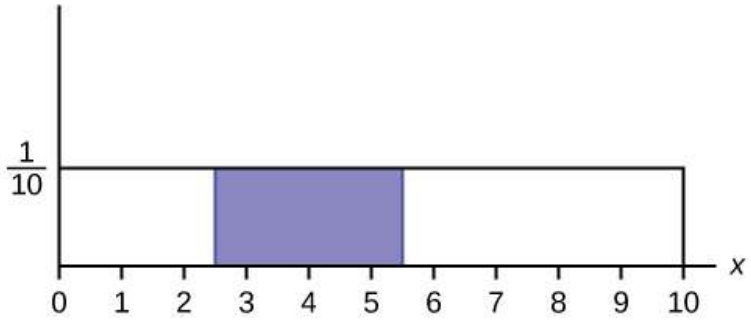


Figure 5.44

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 $P(x < 3.5 | x < 4)$?

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 $\bar{x}$: _____

ii. Highest value for $\bar{x}$: _____

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. $P(x < 4|x < 7.5) =$ _____

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 70th 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 30th 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. $P(12 < x | x < 28) =$ _____
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. $P(x > 5 | x > 3) =$ _____
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. $X \sim$ _____
c. Graph the probability distribution.
d. $f(x) =$ _____
e. $\mu =$ _____
f. $\sigma =$ _____
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. $P(7 < x < 13 | x > 9) =$ _____. 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. $X \sim$ _____
c. Graph the probability distribution.
d. $f(x) =$ _____
e. $\mu =$ _____
f. $\sigma =$ _____
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. $X \sim$ _____
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 70th 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 30th 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. μ = _____
f. σ = _____
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 90th percentile. This means that 90 percent of the time, the time is less than _____ minutes.
k. Find the 75th 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. $\mu =$ _____
e. $\sigma =$ _____
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. $\mu =$ _____
e. $\sigma =$ _____
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. $\mu =$ _____
d. $\sigma =$ _____
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. $\mu =$ _____
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

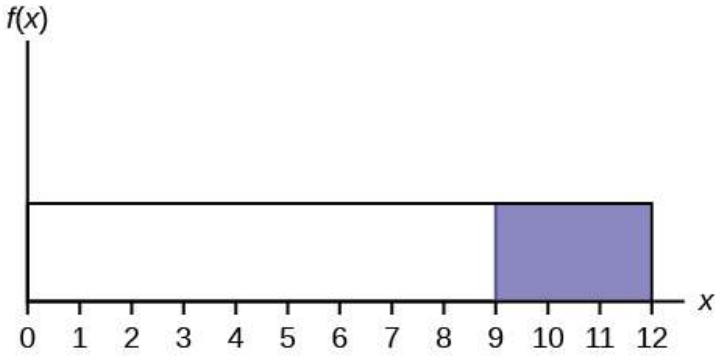


Figure 5.52

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

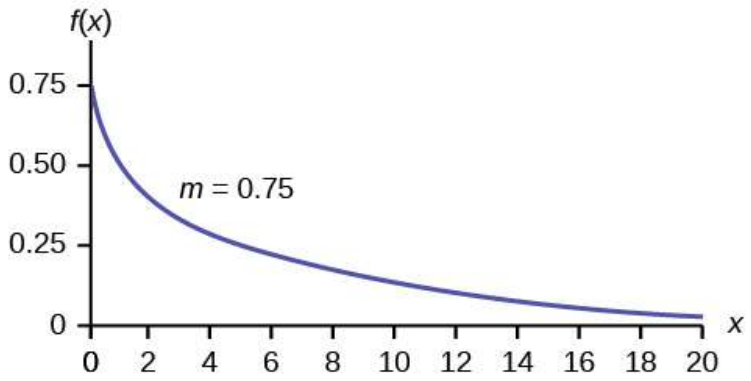


Figure 5.53

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 $\frac{1}{25 - 16} = \frac{1}{9}$.

$P(X > 19) = (25 - 19)\left(\frac{1}{9}\right) = \frac{6}{9} = \frac{2}{3}$.

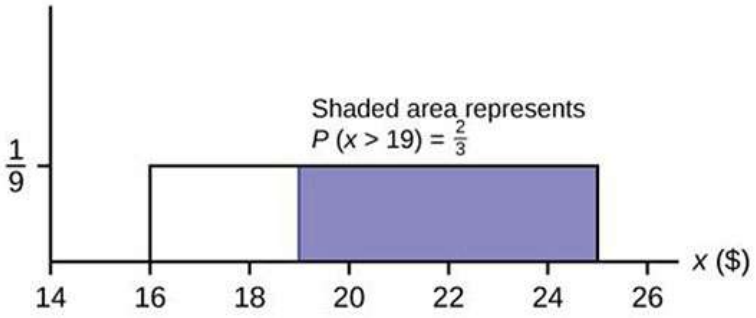


Figure 5.54

b. $P(19 < X < 22) = (22 - 19)\left(\frac{1}{9}\right) = \frac{3}{9} = \frac{1}{3}$.

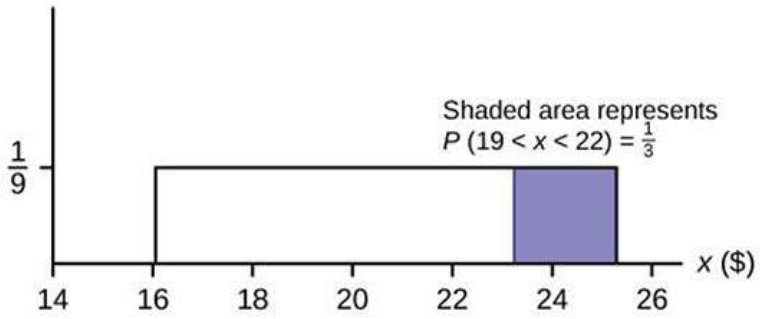


Figure 5.55

c. The area must be 0.25, and 0.25 = (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, $P(x > 21|x > 18) = (25 - 21)\left(\frac{1}{7}\right) = 4/7$.

- Use the formula: $P(x > 21|x > 18) = \frac{P(x > 21 \text{ AND } x > 18)}{P(x > 18)}$

  $= \frac{P(x > 21)}{P(x > 18)} = \frac{(25-21)}{(25-18)} = \frac{4}{7}$.

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, $P(6 < t < 10) - P(t < 6)$
$= = \left(1 - e^{-\frac{1}{8}*10}\right) - \left(1 - e^{-\frac{1}{8}*6}\right) \approx 0.7135 - 0.5276 = 0.1859$

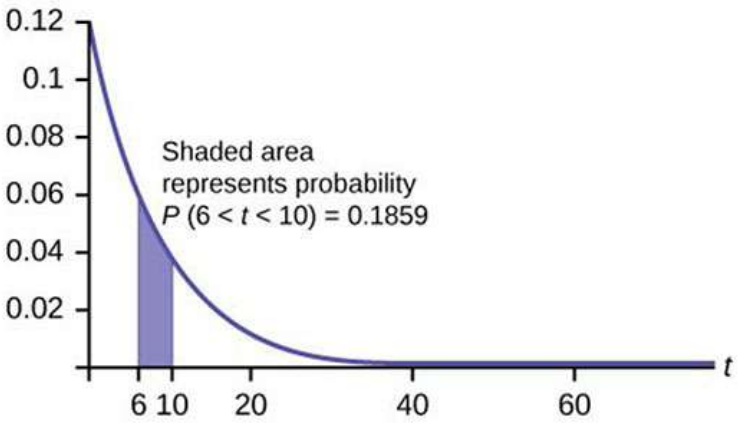


Figure 5.56

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 .

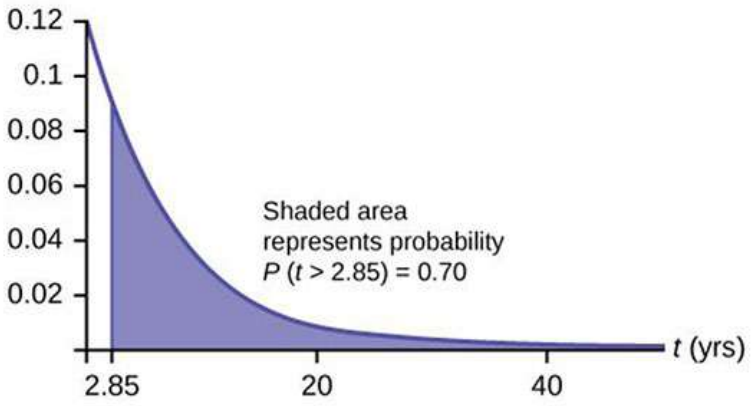


Figure 5.57

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 $\frac{\ln(\text{area\_to\_the\_right})}{(-\text{m})} = \frac{\ln(1 - 0.2)}{-\frac{1}{8}} = 0.1616$.

e. We must find $P(T < 8|T > 7)$.

Notice that by the rule of complement events, $P(T < 8|T > 7) = 1 - P(T > 8|T > 7)$.

By the memoryless property ($P(X > r + t|X > r) = P(X > t)$).

So $P(T > 8|T > 7) = P(T > 1) = 1 - \left(1 - e^{-\frac{1}{8}}\right) = e^{-\frac{1}{8}} \approx 0.8825$

Therefore, $P(T < 8|T > 7) = 1 - 0.8825 = 0.1175$.

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 $P(T > 2|T > 1)$, 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. $P(T > 15|T > 10) = P(T > 5) = 1 - \left(1 - e^{-\frac{5}{7}}\right) = e^{-\frac{5}{7}} \approx 0.4895$.

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)

Introduction

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

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

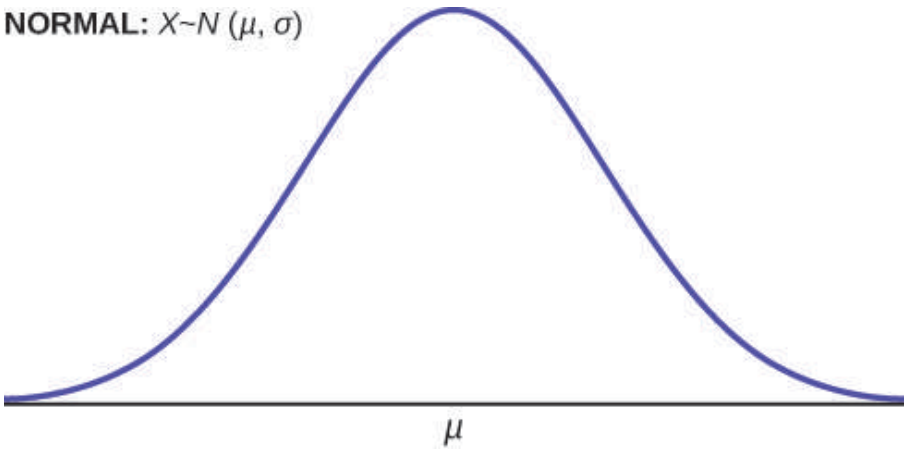


Figure 6.2

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 1 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

$$z = \frac{11 - 5}{2} = 3$$

or

$$5 + (z)(2) = 11,$$

we can solve for *z*.

$$2z = 6$$
$$z = 3$$

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.

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 6.2

a. This z-score tells you that $x = 10$ is **2.5** standard deviations to the *right* of the mean *five*.

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 6.2

b. $z =$ **–4**. This z-score tells you that $x = –3$ is *four* standard deviations to the *left* of the mean.

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 6.2

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 -2σ are +2 and –2, respectively.
- The z-scores for $+3\sigma$ and -3σ 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*.

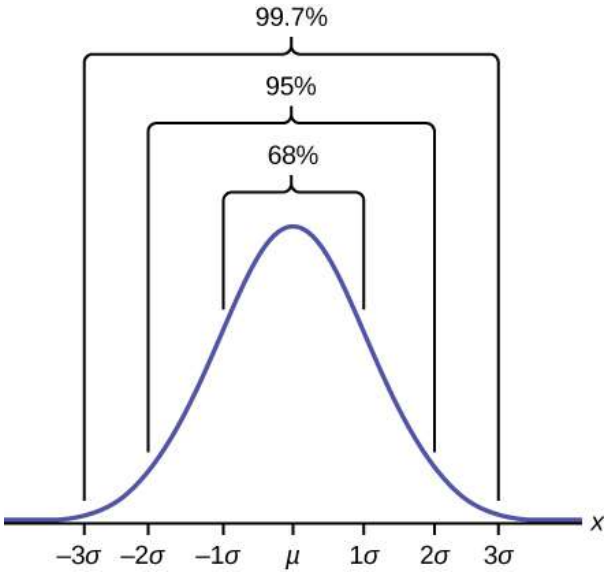


Figure 6.3

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 $X \sim N(170, 6.28)$.

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 6.3

a. –0.32, 0.32, left, 170

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 6.3

b. 177.98 cm, 1.27, right

Try It Σ

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

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 6.4

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 Σ

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

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 6.6

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 1153.34 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 $P(X < x)$.

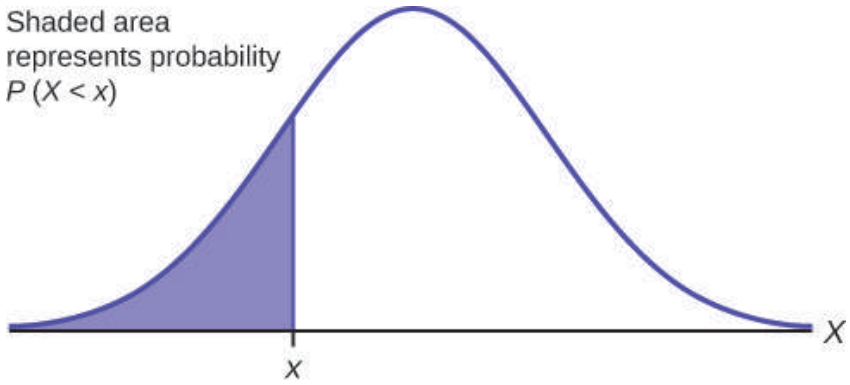


Figure 6.4

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** 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.

a. Find the probability that a randomly selected student scored more than 65 on the exam.

Solution 6.8

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:

$$z = \frac{x - \mu}{\sigma} = \frac{65 - 63}{5} = \frac{2}{5} = .40$$

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 $P(x > 65)$.

$$P(x > 65) = 0.3446$$

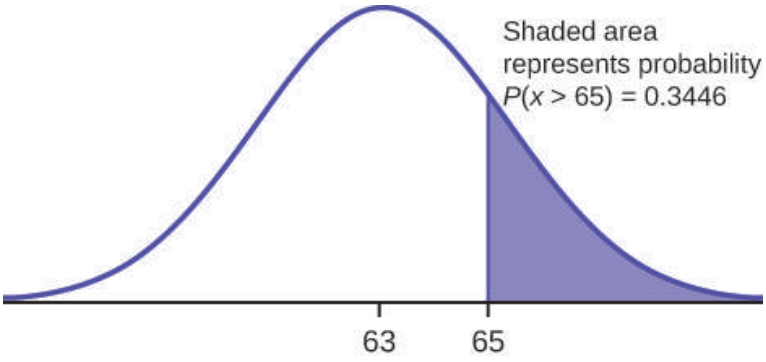


Figure 6.5

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.

Using the TI-83, 83+, 84, 84+ Calculator

Calculate the *z*-score

*Press 2nd Distr
*Press 3:invNorm(
*Enter the area to the left of *z* followed by)
*Press ENTER.
For this Example, the steps are
2nd Distr
3:invNorm(.6554) ENTER
The answer is 0.3999, which rounds to 0.4.

b. Find the probability that a randomly selected student scored less than 85.

Solution 6.8

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.

c. Find the 90th 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 6.8

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.

$$k = 69.4$$

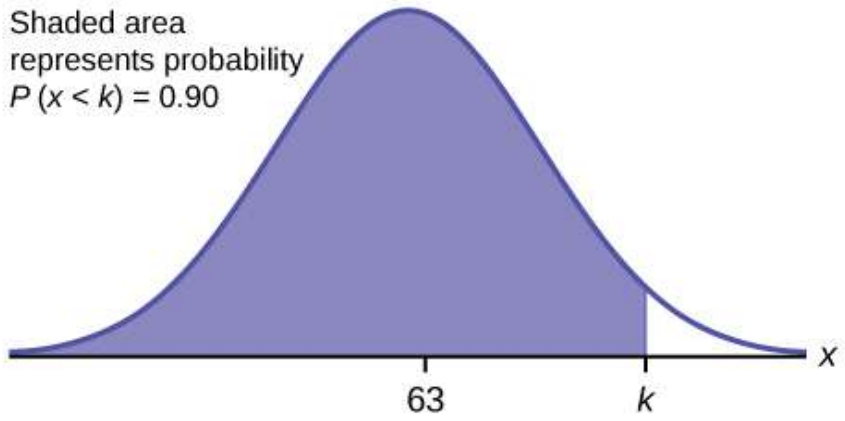


Figure 6.6

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

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 6.8

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.

a. Find the probability that a household personal computer is used for entertainment between 1.8 and 2.75 hours per day.

Solution 6.9

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.

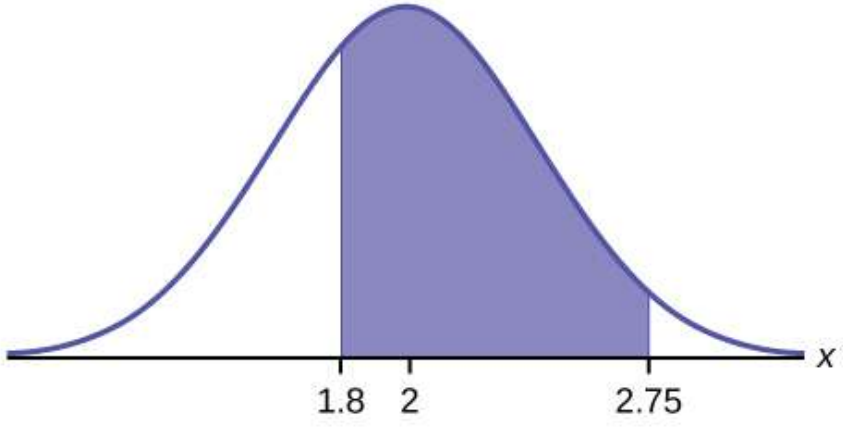


Figure 6.7

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.

b. Find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment.

Solution 6.9

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 $P(x < k) = 0.25$.

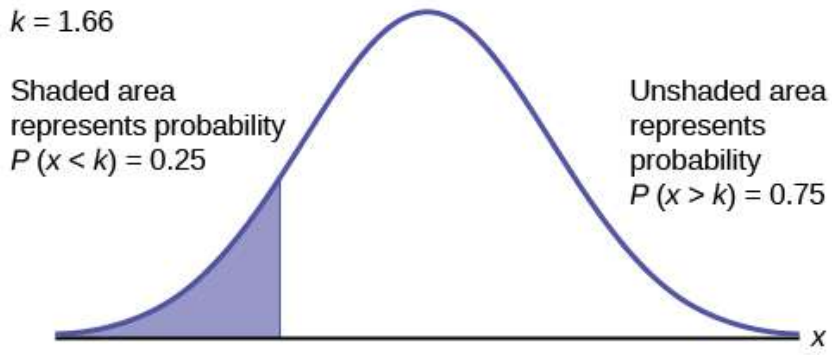


Figure 6.8

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+ 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.

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 6.10

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.

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 6.10

b. normalcdf(-10^{99},50.8,36.9,13.9) = 0.8413

c. Find the 80th percentile of this distribution, and interpret it in a complete sentence.

Solution 6.10

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.

a. Calculate the interquartile range (*IQR*).

Solution 6.11

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

b. Forty percent of the ages that range from 13 to 55+ are at least what age?

Solution 6.11

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.

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 6.12

a. normalcdf(6,10^99,5.85,0.24) = 0.2660

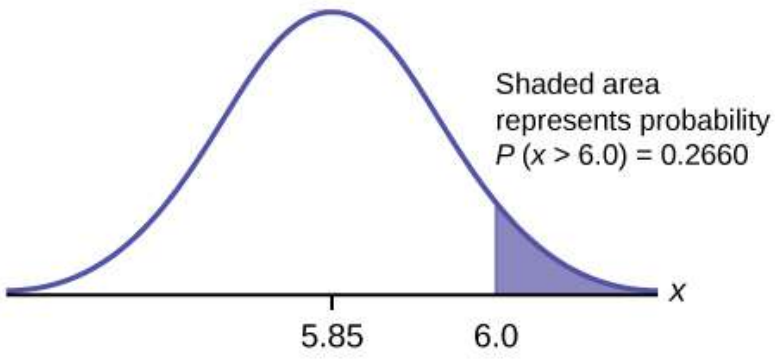


Figure 6.9

b. The middle 20 percent of mandarin oranges from this farm have diameters between _____ and _____.

Solution 6.12

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.

c. Find the 90th percentile for the diameters of mandarin oranges, and interpret it in a complete sentence.

Solution 6.12
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 16^{th} percentile, and interpret it in a complete sentence.

6.3 | Normal Distribution—Lap Times

Stats Lab

6.1 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.

Use the data from **Appendix C**. 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.

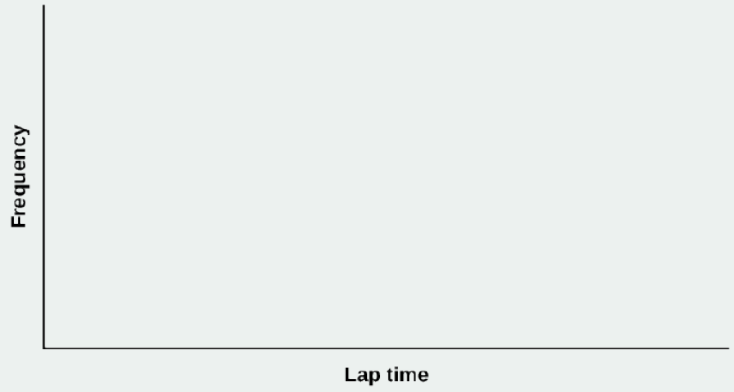


Figure 6.10

Calculate the following:

a. $\bar{x}$ = _____

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?

- $X \sim$ _____(_____,_____)
- 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 15^{th} percentile is _____.
- The 85^{th} 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 85^{th} 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 15^{th} percentile is _____.
- The 85^{th} percentile is _____.
- The median is _____.
- The probability that a randomly chosen lap time is more than 130 seconds is _____.
- Explain the meaning of the 85^{th} 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

6.2 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.

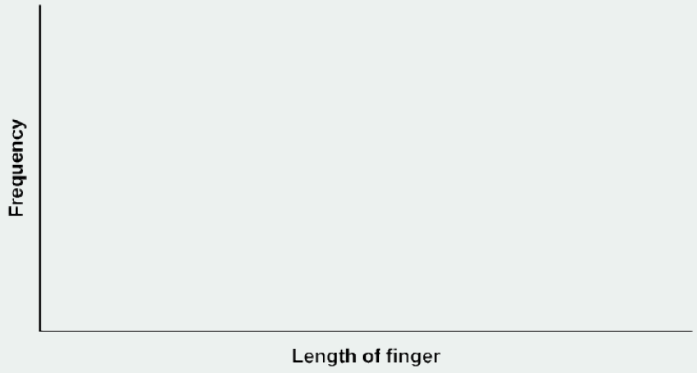


Figure 6.11

3. Calculate the following:
   a. $\bar{x}$ = _____
   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?

- $X \sim$ _____ (_____, _____)
- 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 15th percentile is _____.
- The 85th 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 85th 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 15th percentile is _____.
- The 85th 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 85th 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

6.0 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. $X \sim N(1, 2)$

$\sigma =$ _____

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. $X \sim N(3, 5)$

$\sigma =$ _____

7. $X \sim N(-2, 1)$

$\mu =$ _____

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?

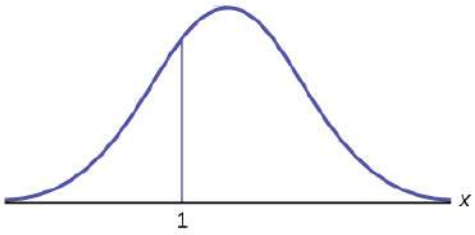


Figure 6.12

44. What is the area to the right of one?

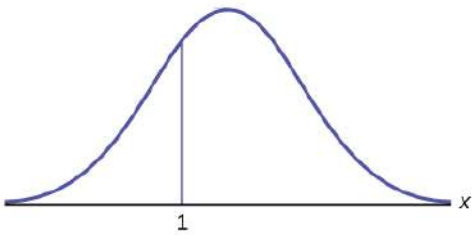


Figure 6.13

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?

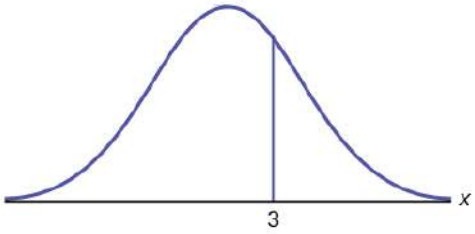


Figure 6.14

47. What is the area to the right of three?

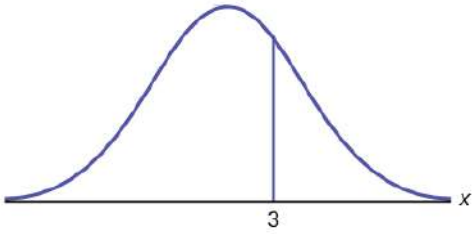


Figure 6.15

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.

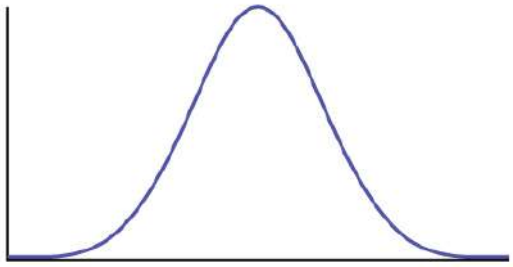


Figure 6.16

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.

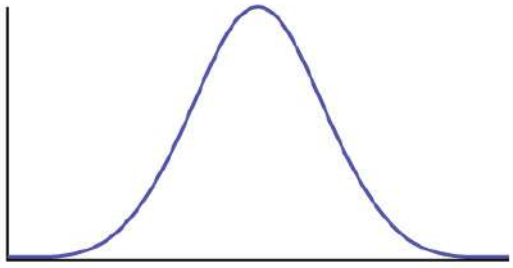


Figure 6.17

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.

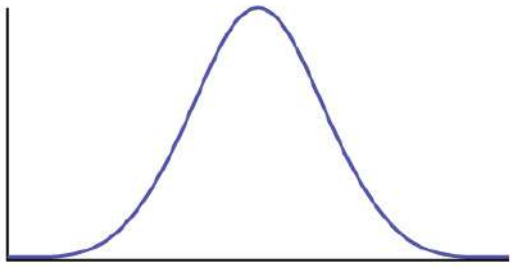


Figure 6.18

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. $X \sim$ _____ (_____ ,_____)
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. $X \sim$ _____ (_____, _____)
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. $X \sim$ _____ (_____, _____)
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 $X \sim$ _____ (_____, _____)
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. $X \sim$ _____ (_____, _____)
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. $X \sim$ _____ (_____, _____)
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. $X \sim$ _____ (_____, _____)
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
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
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 μ = 20 and σ = 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 95^{th} 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)

Introduction

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

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 $\bar{X}$, which consists of sample means, tends to be **normally distributed** and

$$\bar{X} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$$

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 $\bar{X}$, 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 $\bar{X}$ has a different z-score associated with it from that of the random variable X. The mean $\bar{x}$ is the value of $\bar{X}$ in one sample.

$$z = \frac{\bar{x} - \mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)},$$

μ_X is the average of both X and $\bar{X}$.

$\sigma_{\bar{x}} = \frac{\sigma x}{\sqrt{n}}$ = standard deviation of $\bar{X}$ 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.

a. Find the probability that the **sample mean** is between 85 and 92.

Solution 7.1

a. Let X = one value from the original unknown population. The probability question asks you to find a probability for the **sample mean**.

Let $\bar{X}$ = the mean of a sample of size 25. Because $\mu_x = 90$, $\sigma_x = 15$, and $n = 25$,

$$\bar{X} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$$

Find $P(85 < \bar{x} < 92)$. Draw a graph.

$P(85 < \bar{x} < 92) = 0.6997$

The probability that the sample mean is between 85 and 92 is 0.6997.

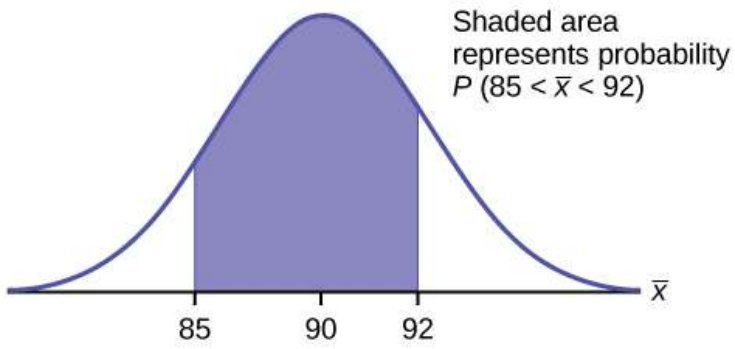


Figure 7.2

Find $P(85 < \bar{x} < 92)$. Draw a graph.

$$P(85 < \bar{x} < 92) = 0.6997$$

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`$(85,92,90,\frac{15}{\sqrt{25}}) = 0.6997$

b. Find the value that is two standard deviations above the expected value, 90, of the sample mean.

Solution 7.1

b. To find the value that is two standard deviations above the expected value 90, use the following formula

$$\text{value} = \mu_x + (\#\text{ of STDEVs})\left(\frac{\sigma_x}{\sqrt{n}}\right)$$

$$\text{value} = 90 + 2\left(\frac{15}{\sqrt{25}}\right) = 96.$$

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

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 7.2

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 $\bar{X}$ = the **mean** time, in hours, it takes to play one soccer match.

If $\mu_X =$ _____, $\sigma_X =$ _____, 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 $P(1.8 < \bar{x} < 2.3)$. Draw a graph.

$$P(1.8 < \bar{x} < 2.3) = 0.9977$$

`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.

2^{nd} DIStR
3:invNorm

$$k = \text{invNorm}\left(\text{area to the left of } k, \quad \text{mean}, \quad \frac{\text{standard deviation}}{\sqrt{\text{sample size}}}\right)$$

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

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 95^{th} percentile for the sample mean age (to one decimal place).

Solution 7.3

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 $P(X > 30) =$ `normalcdf`$(30,E99,34,1.5) = 0.9962$.

d. Let k = the 95^{th} 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

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 90^{th} 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 7.4

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 $\mu = 16.00$ ounces (as labeled) and $\sigma = 0.143$ 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:

2^{nd} 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.

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 7.5

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 $\mu_X = 90$, $\sigma_X = 15$, 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 $P(\Sigma x > 7500)$

$P(\Sigma x > 7500) = 0.0127$

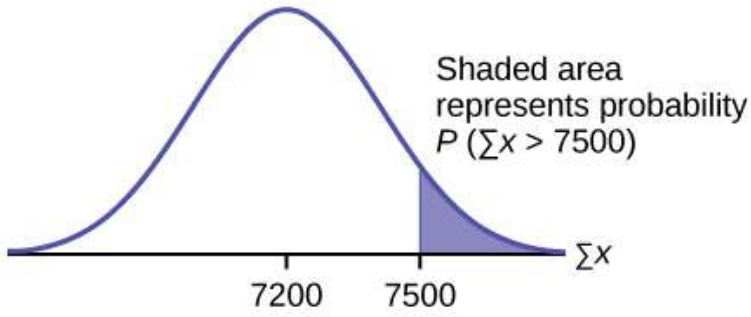


Figure 7.3

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)(μ_X, ($\sqrt{n}$) (σ_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 Σ

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

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 80^{th} percentile for the sum of the 50 ages.

Solution 7.6

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 80^{th} percentile.
k = invNorm(0.80,(50)(34), $(\sqrt{50})(15)$) = 1789.3

Try It Σ

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

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 7.7

a. $\mu_{\Sigma x} = n\mu_x = 70(8.2) = 574$ minutes and $\sigma_{\Sigma x} = (\sqrt{n})(\sigma_x) = (\sqrt{70})(1) = 8.37$ 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 Σ

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 $\bar{x}$ 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 $\bar{X}$ is $\frac{\sigma}{\sqrt{n}}$.) This means that the sample mean $\bar{x}$ 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),

$$\mu_X = \frac{a+b}{2} = \frac{1 + 5}{2} = 3$$

$$\sigma_X = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(5-1)^2}{12}} = 1.15$$

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 $\bar{X}$ = the mean stress score for the 75 students. Then,

$$\bar{X} \sim N\left(3, \frac{1.15}{\sqrt{75}}\right) \text{ where } n = 75.$$

a. Find $P(\bar{x} < 2)$. Draw the graph.

Solution 7.8

a. $P(\bar{x} < 2) = 0$

The probability that the mean stress score is less than 2 is about zero.

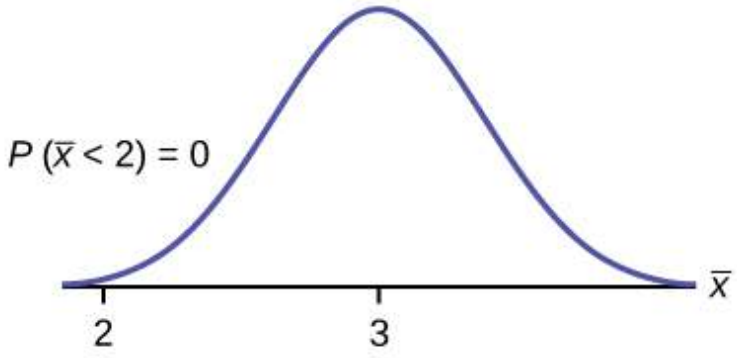


Figure 7.4

$$\texttt{normalcdf}\left(1,2,3,\frac{1.15}{\sqrt{75}}\right) = 0$$

REMINDER

The smallest stress score is one.

b. Find the 90th percentile for the mean of 75 stress scores. Draw a graph.

Solution 7.8

b. Let k = the 90th precentile.

Find k, where $P(\bar{x} < k) = 0.90$.

$$k = 3.2$$

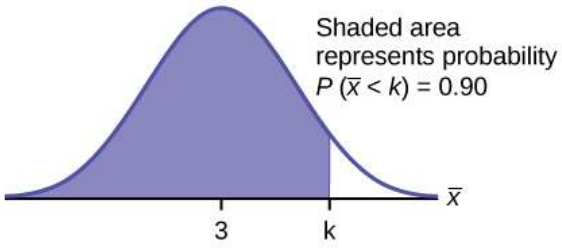


Figure 7.5

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.

$$\texttt{invNorm}\left(0.90,3,\frac{1.15}{\sqrt{75}}\right) = 3.2$$

For problems (c) and (d), let ΣX = the sum of the 75 stress scores. Then, $\Sigma X \sim N[(75)(3), (\sqrt{75})\ (1.15)]$.

c. Find $P(\Sigma x < 200)$. Draw the graph.

Solution 7.8

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 $(\sqrt{75})(1.15) = 9.96$.

$$P(\Sigma x < 200) = 0$$

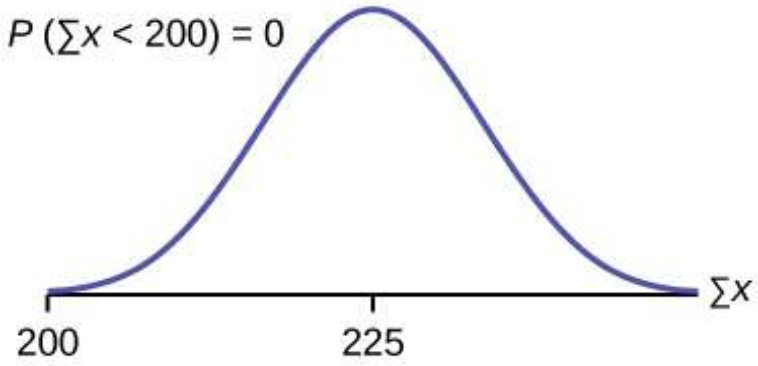


Figure 7.6

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.

d. Find the 90th percentile for the total of 75 stress scores. Draw a graph.

Solution 7.8

d. Let k = the 90th percentile.

Find k where $P(\Sigma x < k) = 0.90$.

$$k = 237.8$$

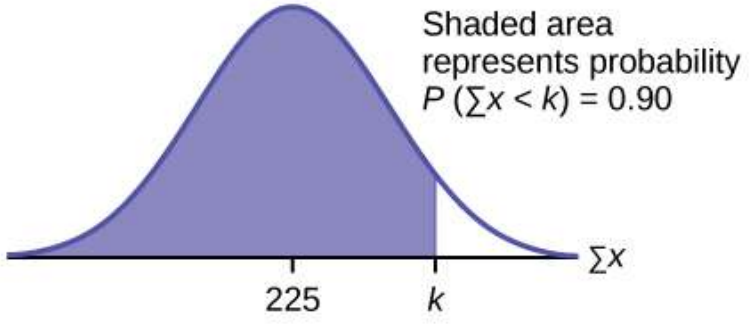


Figure 7.7

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 Σ

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 $\bar{X}$ = the mean excess time used by a sample of $n = 80$ customers who exceed their contracted time allowance.

$\bar{X} \sim N\left(22, \frac{22}{\sqrt{80}}\right)$ by the central limit theorem for sample means.

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 7.9

a. Find: $P(\bar{x} > 20)$

$P(\bar{x} > 20) = 0.79199$ using $\texttt{normalcdf}\left(20,1E99,22,\frac{22}{\sqrt{80}}\right)$

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.

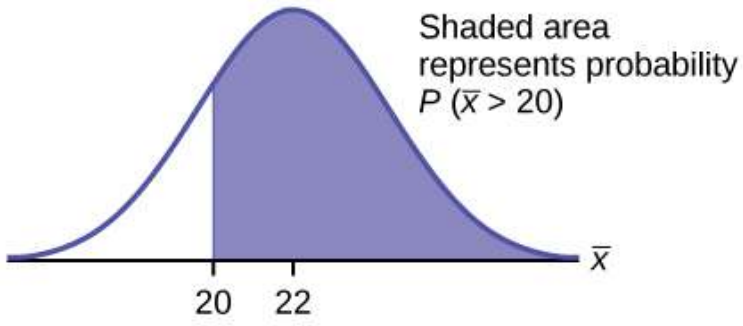


Figure 7.8

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 $P(\bar{x} > 20) = 0.7919$

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.

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 7.9

Let k = the 95th percentile. Find k where $P(\bar{x} < k) = 0.95$.

$k = 26.0$ using $\texttt{invNorm}\left(0.95, 22, \frac{22}{\sqrt{80}}\right) = 26.0$

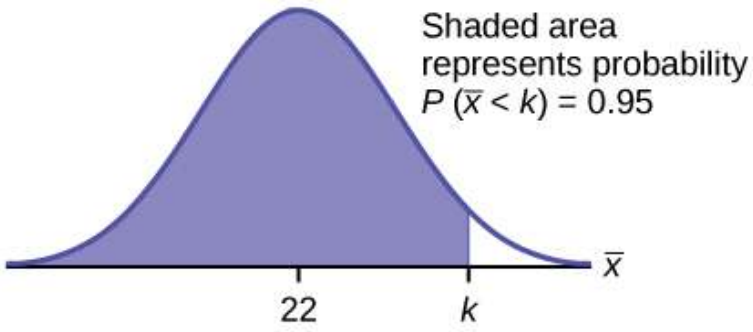


Figure 7.9

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 $P(20 < \bar{x} < 30)$.

b. Find $P(\Sigma x$ 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.

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 7.10

a. We have $\mu_x = \mu = 2$ and $\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{0.5}{10} = 0.05$. Therefore,

1. 50th percentile = $\mu_x = \mu = 2$,
2. 25th percentile = *invNorm*(0.25,2,0.05) = 1.97, and

3. 75th 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. 50th percentile = $\mu_{\Sigma x} = n(\mu_x) = 100(2) = 200$,
2. 25th percentile = invNorm(0.25,200,5) = 196.63, and
3. 75th percentile = invNorm(0.75,200,5) = 203.37.

c. $P(1.75 < \bar{x} < 1.85)$ = `normalcdf`(1.75,1.85,2,0.05) = 0.0013

d. Using the *z*-score equation, $z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$, and solving for *x*, we get $x = 2(0.05) + 2 = 2.1$.

e. The *IQR* is 75th percentile – 25th 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

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 95th percentile for the sample mean age of 65 people. Interpret the results.

f. Find the 90th percentile for the sum of the ages of 65 people. Interpret the results.

Solution 7.11

a. $P(\bar{x} < 35)$ = `normalcdf`(-*E*99,35,30.9,1.8) = 0.9886

b. $P(\bar{x} > 50)$ = `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

7.1 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.

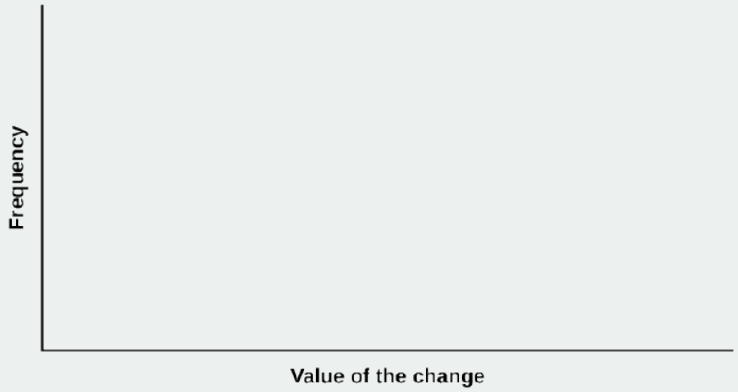


Figure 7.10

4. Calculate the following (n = 1, surveying one person at a time):
   a. $\bar{x}$ = _____
   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.

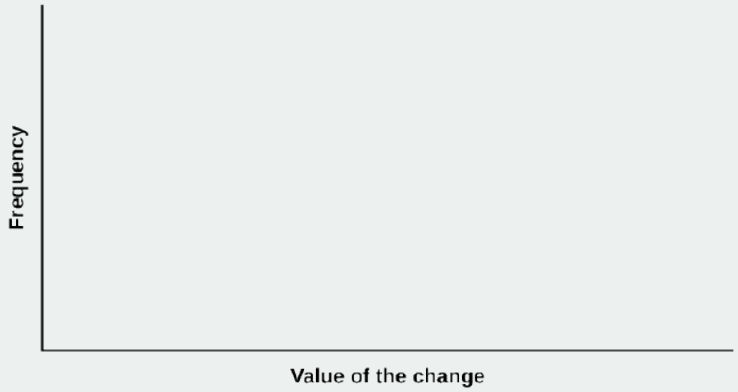


Figure 7.11

4. Calculate the following (n = 2, surveying two people at a time):
   a. $\bar{x}$ = _____
   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

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.

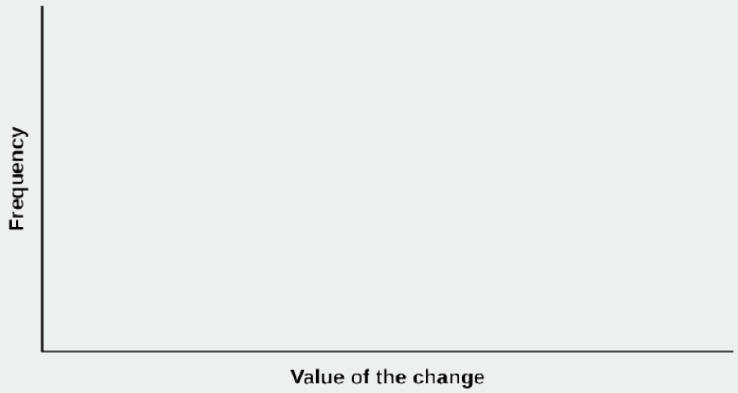


Figure 7.12

4. Calculate the following (n = 5, surveying five people at a time):
   a. $\bar{x}$ = _____
   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? $\bar{X}$ ~ _____(_____, _____)
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

7.2 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:
Means:	$\bar{x}$ = ____	$\bar{x}$ = ____	$\bar{x}$ = ____	$\bar{x}$ = ____	

Table 7.5

2. Calculate the following:
   a. $\bar{x}$ = _____
   b. $s_{\bar{x}}$ = _____
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:	$\bar{x}$ = ____	$\bar{x}$ = ____	$\bar{x}$ = ____	$\bar{x}$ = ____	

Table 7.6

4. Calculate the following:
   a. $\bar{x}$ = _____
   b. $s_{\bar{x}}$ = _____
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.

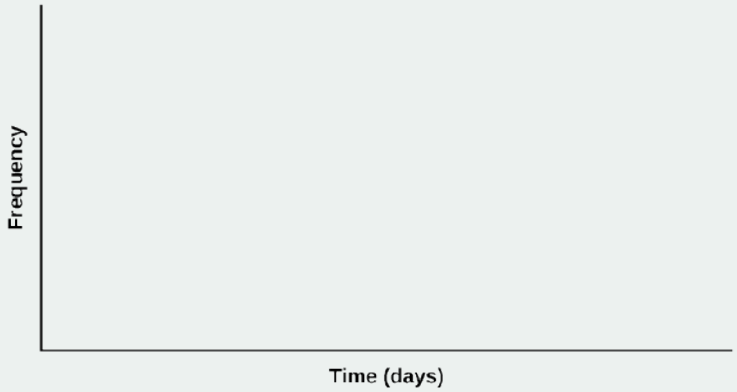


Figure 7.13

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.

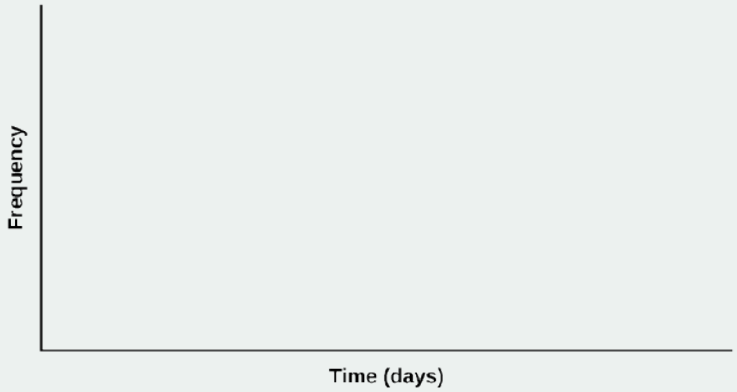


Figure 7.14

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.

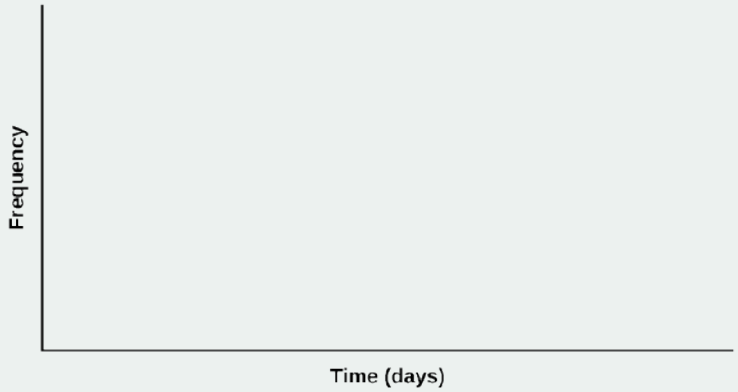


Figure 7.15

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. $n = 5$: $\bar{x}$ ~ _____(_____, _____)
   b. $n = 10$: $\bar{x}$ ~ _____(_____, _____)
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, $\bar{X}$, and the sample sum, ΣX

If the size (n) of the sample is sufficiently large, then $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$ 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 $\bar{x}$) is $\bar{x} = \frac{\text{sum of all values in the sample}}{\text{number of values in the sample}}$, 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, $\bar{x}$, gets to μ.

FORMULA REVIEW

7.1 The Central Limit Theorem for Sample Means (Averages)

Central limit theorem for sample means: $\bar{X} \sim N\left(\mu_x, \frac{\sigma x}{\sqrt{n}}\right)$

Mean $\bar{X}$: μ_x

Central limit theorem for sample means z-score and standard error of the mean: $z = \frac{\bar{x} - \mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)}$

Standard error of the mean (standard deviation ($\bar{X}$)): $\frac{\sigma_x}{\sqrt{n}}$

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 $= \frac{\Sigma x - (n)(\mu_X)}{(\sqrt{n})(\sigma_X)}$

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 $\bar{X}$ 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. $\bar{X} \sim$ _____(_____, _____)

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.

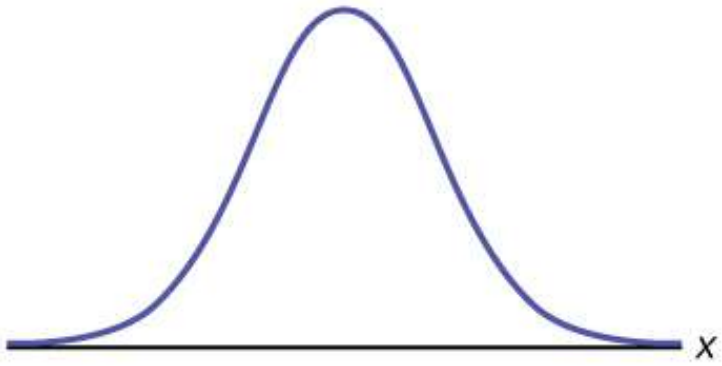


Figure 7.16

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.

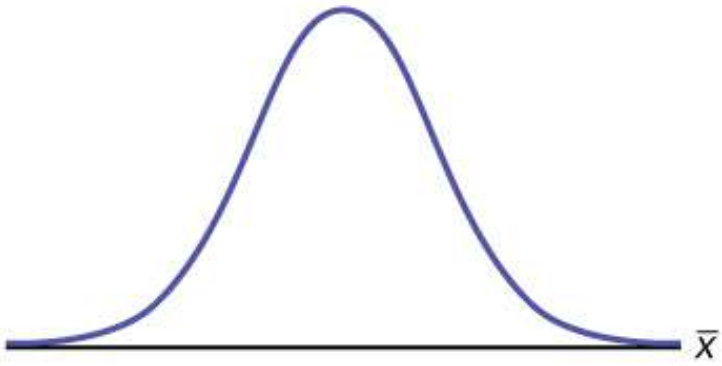


Figure 7.17

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.

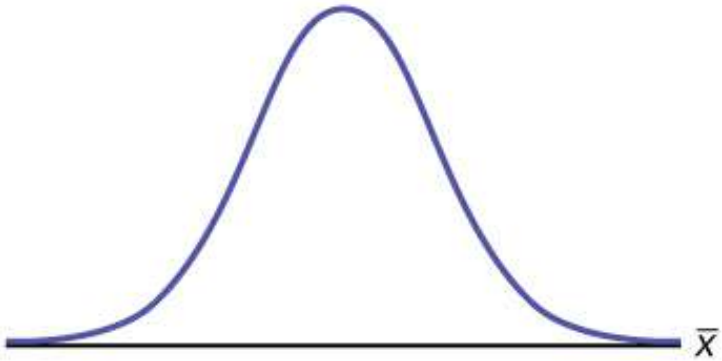


Figure 7.18

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 (16) and (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 90^{th} 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 90^{th} 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. $X \sim$ _____(_____, _____)
c. In words, $\bar{X}$ = _____.
d. $\bar{X} \sim$ _____ (_____, _____)
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 $\bar{X}$ = average distance in feet for 49 fly balls, then $\bar{X} \sim$ _____(_____, _____).
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 $\bar{X}$. 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, $\bar{X}$ = _____.
c. $\bar{X}$ ~ _____(_____, _____)
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 $\bar{X}$ be the average of the 49 races.

a. $\bar{X}$ ~ _____(_____, _____)
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, $\bar{X}$ = _____.
d. $\bar{X}$ ~ _____(_____, _____)
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, $\bar{X}$ = _____.
c. $\bar{X}$ ~ _____(_____, _____)
d. The IQR for $\bar{X}$ 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 $\bar{X}$ is approximately equal to the mean of X.
b. When the sample size is large, $\bar{X}$ is approximately normally distributed.
c. When the sample size is large, the standard deviation of $\bar{X}$ 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 $\bar{X}$ = average percentage of fat calories.

a. $\bar{X}$ ~ _____(_____, _____)
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, $\bar{X}$ = _____.
c. $\bar{X}$ ~ _____(_____, _____)
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. $\bar{X} \sim N(4.59, 0.10)$
b. $\bar{X} \sim N\left(4.59, \frac{0.10}{\sqrt{16}}\right)$
c. $\bar{X} \sim N\left(4.59, \frac{16}{0.10}\right)$
d. $\bar{X} \sim N\left(4.59, \frac{\sqrt{16}}{0.10}\right)$

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 35th 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 90th percentile for an individual manager's salary.
g. Find the 90th 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, $\bar{X}$ = _____.
d. $\bar{X}$ ~ _____(_____, _____)
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 $\bar{X}$ 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
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

Table 7.7

Company	Closing Stock Prices
35	29.25

Table 7.7

a. In words, X = _____.
b. i. $\bar{x}$ = _____
   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. $\bar{x}$ = _____
   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, $\bar{X}$ ~ _____(_____, _____).

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 12^{th} 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 $\bar{X}$ 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 $\bar{X}$, and find the probability.

a. Sketch the distributions of X and $\bar{X}$ on the same graph.
b. $\bar{X} \sim$ _____(_____, _____)
c. $P(\bar{x} < 60) =$ _____
d. Find the 30^{th} percentile for the mean.
e. $P(56 < \bar{x} < 62) =$ _____
f. $P(18 < \bar{x} < 58) =$ _____
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. μ_x = _____
d. σ_x = _____
e. In words, $\bar{X}$ = _____.
f. $\bar{X} \sim$ _____(_____, _____)
g. In words, ΣX = _____.
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, $\bar{X}$ = _____.
c. $\bar{X} \sim$ _____(_____, _____)
d. In words, ΣX = _____.
e. $\Sigma X \sim$ _____(_____, _____)
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
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. $X \sim E(0.88, 0.88)$

c. $\bar{X}$ = average amount of change carried by a sample of 25 students.

d. $\bar{X} \sim N(0.88, 0.176)$

e. 0.0819

f. 0.1882

g. The distributions are different. Part (a) is exponential and part (b) 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. $\bar{X} \sim N\left(2{,}000, \frac{8{,}000}{\sqrt{1{,}000}}\right)$

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.

$P(2{,}000 < \bar{X} < 2{,}100) = 0.1537$

$P(2{,}100 < \bar{X} < 2{,}200) = 0.1317$

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

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. $\bar{X} \sim N\left(60, \frac{9}{\sqrt{25}}\right)$

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 $\mu = 17$, $\sigma = 0.8$, $\bar{x} = 16.7$, and $n = 30$. To calculate the probability, we use normalcdf(lower, upper, μ, $\frac{\sigma}{\sqrt{n}}$) = $\text{normalcdf}\left(E-99,16.7,17,\frac{0.8}{\sqrt{30}}\right) = 0.0200$.
- 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 $n = 100$, $\bar{x} = 0.862$, and $s = 0.05$.

b. $\Sigma \bar{x} = 85.65$, $\Sigma s = 5.18$

c. normalcdf(396.9,E99,(465)(0.8565),(0.05)($\sqrt{465}$)) ≈ 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 $\text{normalcdf}\left(E-99,1.1,1,\frac{1}{\sqrt{70}}\right) = 0.7986$. 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 $\text{normalcdf}\left(5.111,5.291,5.201,\frac{0.065}{\sqrt{280}}\right) \approx 1$, 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.