Empiricism at Work on Mars

The planet Mars contains numerous populations that scientists have wanted to learn about since the planet was first discovered. Scientists have been trying to estimate all sorts of things about the planet, including the planet's climate, geology, and history. NASA has sent 14 successful missions to sample various aspects of this planet: four Mariner missions between 1965-1971, two Viking Landers in 1976, Mars Pathfinder and Mars Surveyor in 1997, the Mars Odyssey in 2001, the Mars Exploration Rovers in 2004, the Mars Reconnaissance Orbiter in 2006, Phoenix in 2008, Curiosity in 2012, and MAVEN in 2014. Three "rover" missions (Curiosity in 2012, Insight in 2018, and Perseverance in 2021) have landed on Mars. If you are interested in seeing some of the images or looking at some of the sample data collected from these missions, go to mars.jpl.nasa.gov.

term "empirical" (from the Greek word *empeiria,* meaning experience) to the English language. This work and the inductive method were highly influential on the development of science and the evolution of the scientific method.

Induction embraces the philosophical principle "what is true of the many is true of the whole" and can be thought of as bottom-up thinking. All applications of statistical inference use inductive reasoning.

Let's consider an example. Down syndrome (DS) is a relatively common disorder in humans. Geneticists tell us that individuals with DS have three copies of chromosome 21 instead of two. How did they reach this conclusion? A French geneticist in 1959 analyzed chromosomes of individuals with what was believed to be Down syndrome and observed an extra copy of chromosome 21 in each of them.[4] This discovery was later confirmed by further studies using more advanced techniques. Is this absolute proof that an extra copy of chromosome 21 is the only cause of DS? At some time in the future, could a person exhibit DS and not have an extra copy of chromosome 21?

Unlike deductive reasoning, induction is not guaranteed to produce absolutely true conclusions. Thus, inductive inference must be associated with *probable* conclusions and some degree of uncertainty. As the data increases, the more probable the inductive conclusion. Probability is the language of uncertainty and will be used to define the degree of belief we have in our conclusions. Being able to express uncertainty in a precise manner is one of the reasons we study probability in a statistics course before statistical inference.

Inductive reasoning in one of the standards used to determine whether someone's beliefs about the world are "justified." As the statistician R. A. Fisher said, "inductive inference is the only process known to us by which essentially new knowledge comes into the world… (and was a) … contribution to the intellectual development of mankind".[5] As the philosopher Robert Audi put it, induction is empiricism's "role … in grounding rationality."[6]

1.1 Exercises

Basic Concepts

1. Complete the sentence: An empirical claim would be one that ….
2. What is the difference between knowledge and belief?
3. Is the idea that three copies of chromosome 21 causes Down Syndrome closer to a belief or to knowledge?
4. Do a Google search to determine the difference between anecdotal evidence and empirical evidence.
5. Restate the sentences to have the same meaning but without using the word "empirical":
   a. "Only empirical research can decide whether bankruptcy law helps or hurts entrepreneurs."
   b. "The book reviews the new theory carefully and in language accessible to the general reader, and then subject it to a detailed empirical examination."
   c. "It was established as an empirical matter, that when average incomes rise, the average incomes of the poorest fifth of society rise proportionately."
   d. "And empirical estimates are being replaced by mathematical exactness."

Exercises

6. If a doctor uses an empirical therapy on your illness, what exactly would that mean?

7. Sherlock Holmes pleaded to his friend Watson "Data! Data! Data! I can't make bricks without clay". What were the "bricks" and "clay" that Holmes was referring to?

8. Retailers often have customer loyalty or rewards programs in which a person supplies some basic contact and demographic information in exchange for discounts. It probably does not surprise you to learn that the company is collecting data from your shopping preferences and habits. Give examples of the information that might be gathered. How is the "story" that the data tells potentially useful to the retail company?

9. Suppose there were two bags of marbles. In one bag there are 500 marbles and the other bag contains 100 marbles. You reach in and take 50 marbles out of each bag. If all 50 marbles selected from each bag are red, could you reasonably state a hypothesis that every marble in each bag is red? With respect to the conclusion, do you feel there is a difference in the likelihood of the conclusion based on the number of marbles in each bag?

10. Research shows the London taxicab drivers who have to memorize a map of London to get their taxicab license have larger brain regions devoted to spatial, or mapping, memories. Was this empirical research? Was the research deductive or inductive?

11. For the past 10 years geese have come to our pond in May, therefore geese will come every year in May. Is this an example of deduction or induction?

12. Research indicated that the frontal lobes of the brain involve higher cognitive functions such as conscious thought and problem-solving. Do you think these conclusions were reached primarily by induction or deduction?

13. Determine if this is an example of deduction or induction: All second-degree polynomial equations can be solved using the quadratic formula. The solutions to $3x^2 + 11x - 4 = 0$ are calculated as $x = \dfrac{-11 \pm \sqrt{(-11)^2 - 4(3)(-4)}}{(2)(3)} = \dfrac{-11 \pm 13}{6}$, $x = -4$ or $x = \dfrac{1}{3}$.

14. Consider a decision that you recently made. What kind of information did you use to help you to come to a conclusion? Did you use an inductive or a deductive thinking process?

Exercises 15-17 give examples of valid syllogisms which use deductive reasoning. Fill in the missing premise or conclusion.

15. **Major premise:** All birds have two legs.

Minor premise: Pigeons are birds.

Conclusion: _____

René Descartes

René Descartes, a 17th-century French philosopher, mathematician, and scientist, is known as the father of modern philosophy. He pioneered the use of skepticism and doubt as a method of inquiry. He famously doubted everything except his own existence in his famous statement "Cogito, ergo sum" (I think, therefore I am), and made significant contributions to mathematics, notably in developing Cartesian geometry.

16. **Major premise:** All plants are green.

Minor premise: _____

Conclusion: A cactus is green.

17. **Major premise:** _____

Minor premise: You were stung by a bee.

Conclusion: You had an allergic reaction.

18. René Descartes (1596-1650), a French philosopher and mathematician, summarized his belief in the distinction between the mind and body in the quote, "Cogito, ergo sum," translated to English as "I think, therefore I am." Complete a valid syllogism related to his statement given that the major premise is, "The act of thinking requires a conscious self."

1.2 Basic Statistical Concepts

Statistics has its roots in empiricism. The practice of empiricism requires a focus—what do you want to learn more about? This focus defines a fundamental concept in statistics, the **population**.

Population

A **population** is the total set of subjects or things we are interested in studying.

DEFINITION

The notion of a population is a very general concept. Populations are defined by what a researcher is studying and can come in all shapes and sizes. If you are researching Hank Aaron's major league batting performance, then the population would consist of all 12,364 of Aaron's major league at bats.[7] If someone is studying toucans in Brazil, then all the toucans in Brazil would constitute the population. If you are studying students at your college, then all the students attending your college represent a population.

A list of all members of a population is called a population **frame**.

Frame

A list containing all members of the population is referred to as a **frame**.

DEFINITION

According to the Census Bureau in 2020 there are about 332 million people in the United States.[8] The frame for the population of the United States would be a rather long list containing about 332 million names. Although a previous census would be a good start in developing a frame for the US population, it is doubtful that an exact frame could ever be developed at a given point in time since there is one new birth every 8 seconds, and one death every 12 seconds. There are just too many people being born, dying, and immigrating over a 10-year period to get an exact frame for the US population. But for problems that deal with smaller populations, frames are easily developed. For example, if your statistics class were the population you were studying, the final class roster would be the frame for the population.

1.2 Exercises

Basic Concepts

1. What is a population?
2. What is a frame?
3. What is a population parameter?
4. What is a sample?
5. What is a statistic?
6. What is the difference between a population parameter and a statistic?
7. Describe the relationships between populations, samples, parameters, and statistics.
8. For a given specific population frame, is the value of a specific parameter variable? Should we expect the value of a sample statistic to vary? Why or why not?

Exercises

9. A heart researcher is interested in studying the relationship between diets which are high in calcium and blood pressure in adult females. The researcher randomly selects 20 female subjects who have high blood pressure. Ten subjects are randomly assigned to try a diet which is high in calcium. The other subjects are assigned to a diet with a standard amount of calcium. After one year the average blood pressures for subjects in both groups will be measured and compared to decide if diets high in calcium decrease the average blood pressure.
   a. Identify the population.
   b. What characteristic of the population is being measured?
   c. Identify the sample.

10. A center for drug abuse is conducting a study to determine if heroin usage among teenagers has changed. Historically, they have found that about 1.3 percent of teenagers between the ages of 15 and 19 have used heroin one or more times. In a survey of 1824 teenagers, 37 indicated they had used heroin one or more times.
    a. Identify the population.
    b. What characteristic of the population is being measured?
    c. Identify the sample.

11. Heavy episodic or binge drinking is a serious problem in colleges and universities in the United States. A study reported in *The Journal of the American Medical Association* (JAMA) surveyed a total of 17,592 students selected from 140 US 4-year colleges in order to examine the extent of binge drinking.[11] The study found that 44% of the students surveyed admitted to being binge drinkers. A binge drinker was defined as consuming five or more drinks in a row for men and four or more drinks in a row for women during the two weeks prior to the survey.
    a. Identify the population.
    b. What characteristic of the population is being measured?
    c. Identify the sample.
    d. What are some problems associated with collecting the type of data described in this problem?

12. A nurse is interested in the growth curve of boys from infancy to the age of 18. One thousand boys are randomly selected, and their heights are measured at various intervals from birth until the age of 18. Based on these measurements, growth curves are constructed based on the percentage of heights observed to be at or below a certain height at each interval (this population characteristic is called a percentile and will be discussed in Chapter 4).

a. Identify the population.

b. What characteristic of the population is being measured?

c. Identify the sample.

13. A personnel director is interested in determining how effective a new reading course will be in improving the reading comprehension of her company's employees. The director randomly selects twenty employees and determines the average reading comprehension both before and after instruction in the reading course.

a. Identify the population.

b. What characteristic of the population is being measured?

c. Identify the sample.

14. A predominance of body fat, adiposity, can be associated with a myriad of human illnesses including hypertension, diabetes, stroke, heart disease, gallbladder disease, and breast cancer. A standard measure of overall adiposity is the Quetelet index, which is defined as the weight (kg) divided by the square of the height (m). In a study in the *American Journal of Epidemiology*, the Quetelet index was measured on a sample of women between the ages of 35–65 years visiting a breast screening clinic in New York City. The average value of the Quetelet index computed for the women sampled was 25.2. Assume that one of the goals of the study is to estimate the average Quetelet index for all women attending the breast screening clinic.

a. Identify the population.

b. What characteristic of the population is being measured?

c. Identify the sample.

d. What is the unknown population parameter in this problem?

e. What is the estimate of this parameter?

1.3 Descriptive versus Inferential Statistics

The science of statistics is divided into two categories, **descriptive** and **inferential**. Descriptive methods describe and summarize data and are used as a method of discovery. Inferential methods aid in making decisions and predictions about population parameters and processes for which it is impractical to obtain measurements on all population members.

Inferential Statistics

The scope of reality can be vast. Historically, the ozone layer over most of the Earth's surface is about 3 millimeters thick. A problem such as trying to assess the ozone thickness over the entire surface of the earth at a point in time would be an impossible sample space to measure. There is no way to know the actual thickness of the ozone layer except by an inductive process: namely sampling, data collection, and statistical inference.

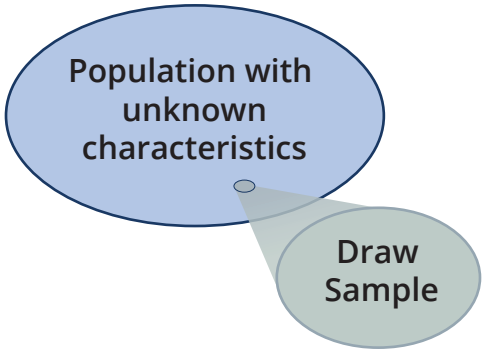


Figure 1.3.1

Inferential statistics is about estimating and making inferences (empirically supported judgments) about population parameters. It is a classical application of inductive reasoning. Increasing the sample size tends to improve the precision and reliability of the estimated population parameters.

Inferential Statistics

The objective of **inferential statistics** is to make reasonable estimates of population characteristics using sample data.

DEFINITION

If data were free, it would be preferable to have measurements of the entire population, but in most cases the required data is either not obtainable or would be much too costly to obtain. For example, to be absolutely certain that all car air bags will work satisfactorily when needed would require each new car to be crash tested. If 100 percent testing were a requirement, cars with air bags would be a scarce commodity. Fortunately for automobile manufacturers, statistical sampling techniques can reliably estimate, with a high degree of confidence, what fraction of air bags that will inflate.

Thickness of the Ozone Layer

—

The length of this line segment is 3mm, which is the approximate size of the ozone layer.

1.3 Exercises

Basic Concepts

1. Is inferential statistics a deductive or inductive process?
2. What is the difference between descriptive and inferential statistics?

What were the costs of this paper? The Wakefield paper linking the MMR vaccine to autism has been estimated to have led to a significant drop in MMR vaccine uptake in England, the U.S., and parts of Europe resulting in a large measles outbreak that caused hospitalizations and deaths. The economic costs are estimated to be in the billions of dollars. Additionally, the paper eroded public trust in vaccines and the scientific community, contributing to ongoing vaccine hesitancy and misinformation.

Could this have been avoided? The study had several significant statistical problems. The issues present in the Wakefield study were not overly complex; rather, they were concerns that a proficient statistics student could have potentially identified and questioned. These included:

1. Small sample size: The study included only 12 children.
2. Selection bias: The individuals involved in the study were not chosen at random, suggesting that they might not accurately reflect the broader population of children who were administered the MMR vaccine.
3. Lack of a control group: The study lacked a control group composed of unvaccinated children, complicating the task of contrasting autism prevalence between those who received the MMR vaccine and those who did not.
4. Confounding factors: The study did not adequately account for other factors that could be associated with both MMR vaccination and autism, such as family history of autism, which could have influenced the study's results.
5. Poor statistical analysis: The statistical analysis in the study was flawed, with inappropriate methods used to analyze the data, and selective reporting of results.

The Wakefield study and its consequences demonstrate the importance of a statistically literate society.

1.4 Exercises

Basic Concepts

1. What does it mean to be statistically literate?

2. The Wakefield MMR vaccine controversy is often cited as an example of the consequences of insufficient statistical literacy. What were the statistical problems present in the Wakefield study, and how could these have been avoided?

3. What does it mean to have an intuitive understanding of statistics? How would you develop an intuitive understanding of statistics?

4. When presented with a statistical conclusion, what types of questions should a statistically literate person ask?

1.5 Statistics and Related Fields as a Career

Note

Check out the video entitled *Statisticians in Other Fields* on the *This is Statistics* YouTube channel.

Your career will essentially be a choice of the kinds of problems you desire to solve. Because the amount of data being stored in the world is doubling every two years, we

There are several fields that are related to applied statistics. Here are a few examples:

1. **Data Science:** Data science involves the use of statistical and computational methods to extract insights and knowledge from data. It combines skills in programming, machine learning, and data visualization with a solid foundation in statistical concepts.
2. **Biostatistics:** Biostatistics involves the application of statistical methods to biological and health-related data. Biostatisticians work in areas such as clinical trials, epidemiology, and public health.
3. **Econometrics:** Econometrics involves the use of statistical methods to analyze economic data. Econometricians work in areas such as finance, market research, and policy analysis.
4. **Business Analytics:** Business analytics involves the use of statistical methods to extract insights and knowledge from business data. Business analysts work in areas such as marketing, operations, and strategy.
5. **Actuarial Science:** Actuarial science is a field of study that involves the use of mathematical and statistical methods to analyze and manage financial risks, especially in the insurance industry.

In the next chapter we will begin our journey into statistics by discussing some of the reasons there is so much data being produced.

1.5 Exercises

Basic Concepts

1. How does the chart show that there is a growing interest in the field of statistics?
2. What has been the major contributing factor in the increase in job opportunities for statisticians?
3. Why are there concerns about a shortage of highly-trained individuals with applied statistical skills?
4. Research the salaries of statisticians and data scientists. How do their salaries compare to other professions like engineering and computer science?

CR Chapter Review

Key Terms and Ideas

- Empiricism
- Statistics
- Rationality
- Deductive Reasoning
- Syllogism
- Inductive Reasoning
- Population
- Frame
- Census
- Parameter
- Metric
- Sample
- Statistic
- Descriptive Statistics
- Inferential Statistics
- Statistical Literacy

Why is this important to statisticians? Using current computing technologies, data sets that are feasible to process simple statistical models in a reasonable amount of time are on the order of terabytes.

2.1 Exercises

Basic Concepts

1. What is an exabyte?
2. What is 1000 terabytes called?
3. What are the two ways to store data?
4. What is a sensor?
5. List 3 examples of remote sensing technology.
6. Name two things that you use every day that contain sensors.
7. What is Big Data?
8. What are the four attributes of Big Data?
9. List two sources of Big Data in science.
10. List two sources of Big Data in business and industry.
11. On the Apollo mission, what device worked to ensure the safe and precise landing of the lunar module on the moon's surface? Generally speaking, what data sources was this device connected to?

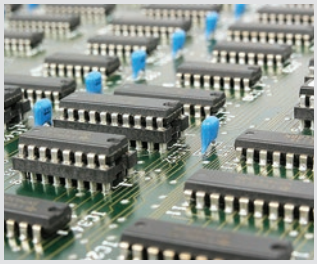
Exercises

12. Compare the size of a zettabyte to a terabyte. Which unit is larger? How much larger?

13. Compare the size of an exabyte to a yottabyte. Which unit is larger? How much larger?

14. Suppose a company collects 50 petabytes of data each year from its customers. If the total available storage capacity of the company is currently two exabytes, how many years of customer data can it store before running out of space?

15. Use Table 2.1.2 and Figure 2.1.1 to determine the year when the world's data storage capacity grew most rapidly between 2010 and 2020.



"The Most Important Master's Thesis of the 20th Century"

In 1948, Claude Shannon wrote a paper entitled "A Mathematical Theory of Communication" which was the foundational work for a field now called information theory. The paper introduced the term "bit" and demonstrated that a series of bits—1s and 0s, of which eight make a byte—could be used to represent all information. The bit/byte would become the standard unit for data storage and network communication of the future. Shannon's foundational work in information theory was not his only contribution. His master's thesis has been called the most important master's thesis of the 20th century. It showed that electrical switches could be configured to perform Boolean logic functions (i.e., digital logic). Shannon's work became the foundation of digital circuit design. Digital circuits are the fundamental component of all digital computers and without them we would not have modern computers, nor modern statistics.

Coping with Poorly Measured Concepts

In a statistical analysis, it is usually not possible to recover from poorly measured concepts or badly collected measurements. Unfortunately, during your lifetime you will be bombarded with statistics derived from poorly measured concepts, confounded measurements, and simply fictitious data. When confronted with statistical evidence of any kind, regardless of whether or not the statistical analysis is done in good faith, it is ultimately up to you to ask reasonable questions about the data and potential confounding variables in the experimental data.

2.2 Exercises

Basic Concepts

1. What is a scale?
2. Where do scales come from? Select a scale and describe its history.
3. How are scales related to measurement?
4. The best measurement scales have what four properties?
5. Describe three scales used in an automobile.
6. What is a level of measurement?
7. What are the four levels of measurement? Give an example of each.
8. What is the fundamental difference between interval and ratio data?
9. What is an arbitrary zero value? Which level of measurement has this property?
10. What is the primary difference between nominal and ordinal data?
11. For which level(s) of measurement is arithmetic appropriate?
12. What are fuzzy concepts? What are the measurement problems associated with fuzzy concepts?
13. Give an example of a tool that has been widely accepted as an instrument used to measure a fuzzy concept.
14. What is a metric?
15. How are sensors related to scale of measurement?
16. **a.** Name three metrics used in business.
    b. Name three metrics used in medicine.
    c. Name three metrics used in sports.

Exercises

17. Determine the level of measurement (ratio, interval, ordinal, or nominal) for each of the following variables.
    - **a.** The temperature (in degrees Fahrenheit) of patients with pneumonia.
    - **b.** The age at which a college student graduates.
    - **c.** Client satisfaction survey responses: Poor, Average, Good, and Excellent.
    - **d.** The region of the U.S. in which an individual lives: North, South, East, or West.

e. The number of people with a Type A personality.

f. A person's blood type.

18. Determine the level of measurement (ratio, interval, ordinal, or nominal) for each of the following variables.

a. The time it takes for a student to complete an exam.

b. Majors of randomly selected students at a university.

c. The category which best describes how frequently a person eats chocolate: Frequently, Occasionally, Seldom, or Never.

d. The number of pounds of snack food eaten by an individual in his or her lifetime.

19. Given the table below on browser usage, what is the highest level of measurement that the data could have?[60] Justify your answer.

Browser Usage Share (%)					
Month	**Year**	**Google Chrome**	**Mozilla Firefox**	**Internet Explorer/ Microsoft Edge**	**Apple Safari**
April	2022	64.3	3.4	4.1	19.2
May	2022	65.0	3.3	4.0	19.0
June	2022	65.9	3.3	4.1	18.6
July	2022	65.1	3.3	4.1	18.9
August	2022	65.5	3.2	4.3	18.8
September	2022	65.7	3.2	4.3	18.7
October	2022	65.3	3.1	4.4	19.0
November	2022	65.9	3.0	4.5	18.7
December	2022	64.7	3.0	4.2	18.3
January	2023	65.4	3.0	4.5	18.7
February	2023	65.8	2.9	4.3	18.8
March	2023	64.8	2.9	4.6	19.5

20. A researcher is studying the preferences of college students for different types of food. They survey a sample of 100 students and ask the following two questions:

- Write the number of times in the last month that you ate a lunch or dinner from each of the following types of restaurants (dine in or take out):

  _____ Asian (Chinese, sushi, pho, Thai, Indian, Korean, etc.)

  _____ Italian (spaghetti, lasagna, pasta, pizza, calzone, etc.)

  _____ Mexican (taco, burrito, enchilada, nachos, quesadilla, fajita, etc.)

  _____ American (hamburger, hot dog, mac-n-cheese, fried chicken, barbecue, etc.)

- Which one of the following is your favorite type of restaurant food?

  Mark with an X.

  _____ Asian _____ Italian _____ Mexican _____ American

  a. For the response to the first question about the number of times eating restaurant food, what level of measurement is the variable *the number of times you ate Asian food*?

  b. For the response to the second question about the favorite type of restaurant food, what level of measurement is the variable *favorite type of restaurant food*?

21. The table displays data from movies released in 2022.

Movies Released in 2022

Movie title	Rating	Runtime (minutes)	Genre	IMDb rating	Oscar nominations
Black Panther: Wakanda Forever	PG-13	161	Action	6.7	5
Avatar 2	PG-13	192	Fantasy	7.7	4
Lightyear	PG	105	Animation	6.1	0
The Fabelmans	PG-13	151	Drama	7.6	7
Scream	R	114	Horror	6.3	0
Top Gun: Maverick	PG-13	130	Action	8.3	6
Glass Onion	PG-13	139	Comedy	7.1	1

Note

IMDb is the Internet Movie Database.

a. What level of measurement is the variable *Rating*?

b. What level of measurement is the variable *Run time*?

c. What level of measurement is the variable *Genre*?

d. What level of measurement is the variable *Oscar nominations*?

22. A hotel manager is interested in getting feedback from guests. Two variables of interest to the manager are cleanliness and aesthetics of the rooms. Discuss what problems you would encounter when measuring those variables.

2.3 Exercises

Basic Concepts

1. What is a confounding variable?
2. What is the scientific method?
3. How does statistics interact with the steps in the scientific method?
4. How do you treat the problem of a confounding variable?
5. Explain the difference between the control group and the experimental group in a controlled experiment.
6. What is an explanatory variable?
7. What is a response variable?
8. What is a completely randomized design? What are the advantages of using a completely randomized design?
9. What is a before and after study?
10. What is the placebo effect? Give an example.
11. What is a double-blind study?
12. How do observational studies differ from controlled experiments?
13. What kinds of problems can be associated with an observational study?
14. What is bias? How can it be controlled?
15. Researchers use surveys for two main purposes. Name and give an example of each.

Exercises

16. Suppose you want to determine the proportion of college students in the state of Virginia that pay more than $500 per year on textbooks. Using the scientific method, how would you conduct the experiment?

17. The health and social problems associated with obesity can be a severe hindrance in attaining many of life's goals. Methods for treating obesity were compared in "One Year Behavioral Treatment of Obesity: Comparison of Moderate and Severe Caloric Restriction and the Effect of Weight Maintenance Therapy," in the *Journal of Consulting and Clinical Psychology*.[68] In the study, a group of 25 women, each of whom was at least 25 kilograms (kg) overweight, were randomly split into two groups. The first group received behavior therapy and was placed on a 1200 calorie per day diet for a period of one year. The second group received behavior therapy and was placed on a 420 calorie per day diet for the first 16 weeks of the year. Then they returned to a 1200 calorie per day diet for the remainder of the year. At the end of a 26-week period, the average weight lost was 11.86 kg for the first group and 21.45 kg for the second group. But after 52 weeks, the average weight lost was 10.94 kg for the first group and 12.18 kg for the second group.
    a. Why is this study an example of a controlled experiment?
    b. What is the explanatory variable?
    c. What is the response variable?

d. Is there a control group in the study? Explain.

e. Suppose that the data was gathered from an observational study instead of from a controlled experiment. How would this affect the conclusions that might be made from the study?

18. An article appearing in the *New England Journal of Medicine* investigated whether the academic performance of asthmatic children being treated with the drug Theophylline was inferior to a non-asthmatic group.[69] In one part of the study, 72 children were identified as being treated for asthma. For each child with asthma, a non-asthmatic sibling was also identified. (The use of sibling controls allows for control of family environment and certain genetic factors on academic achievement.) All 144 children were then given a test to measure academic achievement. There were no significant differences on the test between the two groups.

a. Why is this study an example of a controlled experiment?

b. What is the explanatory variable?

c. What is the response variable?

d. Is there a control group in the study? Explain.

e. Suppose that the data was gathered from an observational study instead of from a controlled experiment. How would this affect the conclusions that might be made from the study?

19. A small clinical pilot study was conducted by a research team from Harvard Medical School and the School of Public Health. Fifteen individuals in the early stages of Multiple Sclerosis were fed bovine myelin, a substance containing two antigens thought to be the target of the immune system's attack in Multiple Sclerosis. Another fifteen were given a placebo. In the study, fewer members of the group fed bovine myelin had major attacks of the disease.[70]

a. Which phase of the Scientific Method best describes this study?

b. Is this an observational study or a controlled experiment?

c. What is the response variable?

d. What is the explanatory variable?

e. Which group is the treatment group?

f. Which group is the control group?

20. London scientists conducted a study to determine if chocolate can trigger migraines. Twelve migraine-prone subjects were given a peppermint-laced chocolate candy and eight migraine-prone subjects were given a peppermint-laced placebo made of carob, peppermint, and vegetable fat. Five subjects from the group given chocolate developed a migraine headache within one day. No one from the group given the placebo developed a migraine in the same time period.[71]

a. Which phase of the Scientific Method best describes this study?

b. Is this an observational study or a controlled experiment?

c. What is the response variable?

d. What is the explanatory variable?

e. Which group is the treatment group?

f. Which group is the control group?

21. Jacob normally plays basketball three days a week and has begun to develop patellar tendinitis, which is inflammation in the patellar tendon and results in nagging knee pain. In an effort to relieve his knee pain, Jacob decides to take a week away from playing basketball and rest his knee. However, after about four days, his friend offers him an analgesic rub and insists that his knee will feel better in two to three days. After using the analgesic rub for a couple of days, Jacob's knee begins to feel better. Did the analgesic rub work? Explain how confounding variables might have played a role on Jacob's knee getting better.

22. The Nurse's Health Study conducted on 87,245 women at Boston's Brigham and Women's Hospital revealed that women who eat a cup of beta carotene-rich food a day have 40 percent fewer strokes and 22 percent fewer heart attacks than those who consume a quarter of a cupful per day.[72]

a. Which phase of the Scientific Method best describes this study?

b. Is this an observational study or a controlled experiment?

c. What is the response variable?

d. What is the explanatory variable?

e. Which group is the treatment group?

f. Which group is the control group?

23. A mental health research group conducted a survey with two of the questions asking "Do you practice yoga?" and "Are you happy?" After conducting the survey, the group concluded that those who practice yoga are generally happier than those that do not practice yoga. Do you think practicing yoga makes one happier? Describe how confounding variables could play a role with the conclusion drawn by the research group.

24. A survey was conducted by an investment firm asking participants the following questions: "Are you financially secure?" and "Do you independently make decisions about your investments?" After analyzing the data from the survey, the firm concluded that people who make investment decisions independently tend to be not as financially secure as those who make decisions with the help of an investment advisor. What confounding variables could have played a role in this conclusion?

2.4 Data Classification

Since the kind of data available affects the types of analyses that can be performed, it is important to recognize data attributes. Data or variables can be categorized in several ways:

- structured or unstructured
- qualitative or quantitative
- discrete or continuous

Years of Employment is a quantitative variable which is a measurement of time. While the data values are reported to one decimal place, the variable is continuous because time is measured along an interval of values.

Annual Salary is a quantitative variable reported in the unit measurement of US dollars leading to a discrete data classification.

Both *Years of Employment* and *Annual Salary* are classified as ratio variables because they fit the criteria. These numeric values have measurement units of equal size. Additionally, a measurement of zero years or $0 has a meaningful interpretation. Finally, a ratio of these variables results in a meaningful proportion; a person who is employed for ten years has worked at the company twice as long as a person with five years of employment.

2.4 Exercises

Basic Concepts

1. Describe the difference between structured and unstructured data. Give an example of each.
2. What is a significant benefit of structured data?
3. What is qualitative data? Give an example.
4. What is quantitative data? Give an example.
5. Which levels of measurement are associated with qualitative data? Which levels are associated with quantitative data?
6. What is the difference between discrete and continuous data?

Exercises

7. The results of a study investigating the nutritional status of mid-nineteenth century Americans were reported in "The Height and Weight of West Point Cadets: Dietary Changes in Antebellum America," in the *Journal of Economic History*.[78] The data is based upon physical examination lists for West Point applicants from 1843 to 1894. Some of the information obtained from each cadet were his height, weight, the state from which the cadet was appointed, the occupation of the father, the income of the parents, and the type of home residence (city, town, or rural) of the cadet.
   a. List the different variables measured on the cadets.
   b. Which variables are quantitative and which are qualitative?
   c. Give the levels of measurement for these variables.
   d. Give an example of unstructured data that may appear in the student record of a 21st century West Point cadet.

8. The major television networks regularly conduct polls in order to ascertain the feelings of Americans on current political issues. In May of 1993, such a poll was conducted by ABC concerning United States involvement in Bosnia.[79] The respondent's gender, political affiliation, and opinion (approve, disapprove, or no

opinion) on how President Clinton was handling the situation in Bosnia represented some of the information supplied by the respondent on the survey. Each respondent was also asked to rate the job that the news media had done (excellent, good, not so good, poor) in covering the situation in Bosnia.

a. List the different variables measured on the respondents.

b. Which variables are quantitative and which are qualitative?

c. Give the levels of measurement for these variables.

d. What are some problems associated with collecting data in polls such as the one described in this exercise?

9. Identify the following variables as discrete or continuous.

a. The number of doctors who wash their hands between patient visits.

b. The amount of liquid consumed by the average American each day.

c. The weight of a newborn baby at a local hospital.

d. The time it takes a person to react to a stimulus.

e. The number of voters who favor a particular candidate.

10. Identify the following variables as discrete or continuous.

a. The number of on-time flights between 1 and 5 pm at the Hartsfield-Jackson International Airport in Atlanta.

b. The height of skyscrapers in New York City.

c. The price of General Electric's common stock.

d. The temperature of US cities.

e. The number of alcoholics who are men.

11. A researcher is studying the preferences of college students for different types of food. They survey a sample of 100 students and ask the following two questions:

- Write the number of times in the last month that you ate a lunch or supper from each of the following types of restaurants (dine in or take out):

  _____ Asian (Chinese, sushi, pho, Thai, Indian, Korean, etc.)

  _____ Italian (spaghetti, lasagna, pasta, pizza, calzone, etc.)

  _____ Mexican (taco, burrito, enchilada, nachos, quesadilla, fajita, etc.)

  _____ American (hamburger, hot dog, mac-n-cheese, fried chicken, barbecue, etc.)

- Which one of the following is your favorite type of restaurant food? Mark with an X.

  _____ Asian _____ Italian _____ Mexican _____ American

a. Is the variable "number of times you ate Mexican restaurant food last month" qualitative or quantitative?

b. Is the variable "number of times you ate Mexican restaurant food last month" discrete, continuous, or neither?

c. Is the variable "favorite type of restaurant food" qualitative or quantitative?

d. Is the variable "favorite type of restaurant food" discrete, continuous, or neither?

Cross-Sectional Data

Cross-sectional data are measurements created at approximately the same period of time.

DEFINITION

For example, consider the life expectancy at birth in 2020 for selected countries given in Table 2.5.2.[85]

Table 2.5.2 - Life Expectancy at Birth 2020

Country	Life Expectancy
Afghanistan	66
Australia	84
Botswana	70
Egypt	73
Guatemala	75
Japan	85
Kenya	67
Sierra Leone	56
Spain	84
Sri Lanka	78
Sweden	83
United Kingdom	82
United States	79

The data in table 2.5.2 represents cross-sectional measurements since the measurements were made in the same time period (2020). People in Japan are expected to live on average until age 85—about 6 years longer than the average for Americans. Developed countries such as Australia and the United States generally have higher life expectancies than developing countries such as Sierra Leone and Kenya. But according to the World Health Organization, life expectancies in developing countries are on the rise due to medical interventions based on advanced technology and drugs. In fact, developing countries are expected to experience a massive increase in their elderly populations over the next 25 years. Most of the statistical methods developed in this text are devoted to cross-sectional data.

2.5 Exercises

Basic Concepts

1. What are time series measurements?
2. What problems are associated with the concept of population when studying time series data?
3. What is a stationary process?
4. What is a nonstationary process?

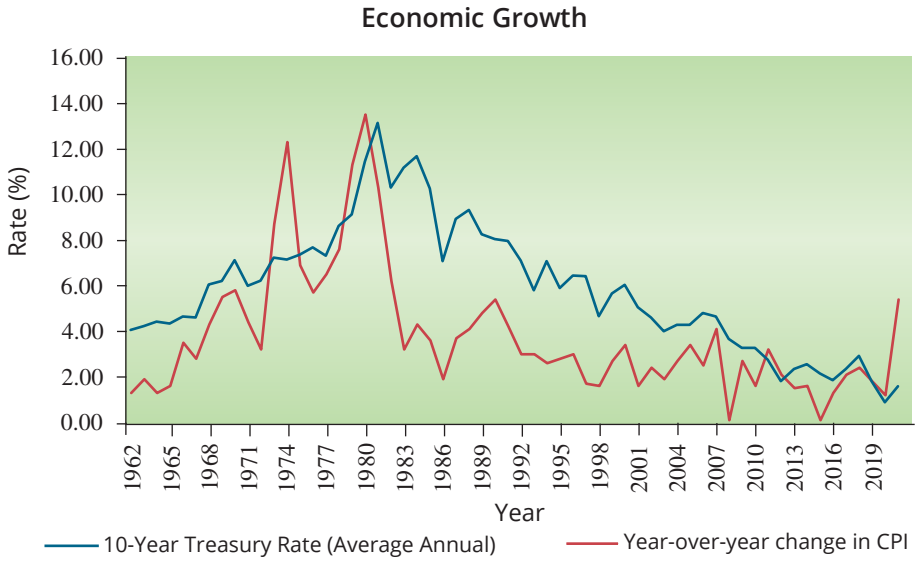
5. What is a trend? If a time series has an 'upward trend,' what does this mean?
6. What is cross-sectional data?
7. What is the difference between cross-sectional data and time series data?

Exercises

Data

The data set is available on stat.hawkeslearning.com under **Discovering Statistics and Data, Fourth Edition > Data Sets > Long Term Interest Rates**.

8. Consider the following graph of long-term interest rates (10-year treasury notes) and inflation rates reported as the year over year change in the consumer price index:[86]

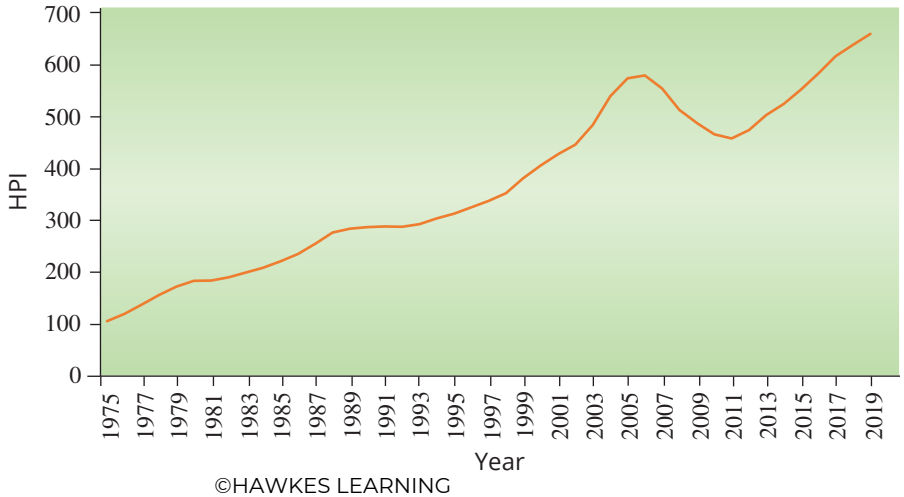


   a. Are the long-term interest rates presented above time series or cross-sectional data?
   b. Are the inflation rates (change in CPI) presented above time series or cross-sectional data?
   c. For each of parts **a.** and **b.**, if the data is time series data, does it appear to be stationary or nonstationary?

Data

The data set is available on stat.hawkeslearning.com under **Discovering Statistics and Data, Fourth Edition > Data Sets > Housing Price Index**.

9. The Housing Price Index (HPI) is calibrated using appraisal values and sales prices for mortgages bought or guaranteed by Fannie Mae and Freddie Mac. The HPI values reflect the base year being used (annual appreciations are the same), with the index value having a base of 100 when first recorded in 1975.[87]



a. Is the data time series or cross-sectional?

b. If the data is time series data, does it appear to be stationary or nonstationary?

10. The following table shows the annual average crude oil price from 1946 through 2023.[88] Prices are adjusted for inflation to February 2023 prices using the Consumer Price Index (CPI-U) as presented by the Bureau of Labor Statistics. Inflation adjusted prices were at an all-time high in 1980, reaching $136.79 dollars per barrel. Crude oil prices reached an all-time low in 1998 (lower than the price in 1946!) when the price per barrel dipped to $21.99. Using the data in the table, discuss if the data set contains time series or cross-sectional data. Also, discuss the data and make some inferences. That is, can you explain some of the fluctuations in the oil prices? [**Note:** The nominal price is the actual price (in dollars) in the specified year.]

Annual Average Domestic Crude Oil Prices ($ per Barrel)

Year	Nominal	Inflation Adjusted Price (Feb 2023)	Year	Nominal	Inflation Adjusted Price (Feb 2023)	Year	Nominal	Inflation Adjusted Price (Feb 2023)
1946	$1.63	$24.73	1972	$3.60	$25.90	1998	$11.91	$21.99
1947	$2.16	$29.06	1973	$4.75	$31.92	1999	$16.56	$29.83
1948	$2.77	$34.67	1974	$9.35	$56.96	2000	$27.39	$47.83
1949	$2.77	$35.00	1975	$12.21	$68.22	2001	$23.00	$39.09
1950	$2.77	$34.64	1976	$13.10	$69.27	2002	$22.81	$38.12
1951	$2.77	$32.11	1977	$14.40	$71.45	2003	$27.69	$45.29
1952	$2.77	$31.39	1978	$14.95	$68.99	2004	$37.66	$59.93
1953	$2.92	$32.77	1979	$25.10	$103.06	2005	$50.04	$77.01
1954	$2.99	$33.50	1980	$37.42	$136.79	2006	$58.30	$86.98
1955	$2.93	$32.86	1981	$35.75	$118.45	2007	$64.20	$92.98
1956	$2.94	$32.56	1982	$31.83	$99.30	2008	$91.48	$127.41
1957	$3.14	$33.60	1983	$29.08	$87.86	2009	$53.48	$74.86
1958	$3.00	$31.27	1984	$28.75	$83.28	2010	$71.21	$98.24
1959	$3.00	$30.96	1985	$26.92	$75.28	2011 (Partial)	$87.04	$116.42
1960	$2.91	$29.63	1986	$14.44	$39.62	2012	$86.46	$113.31
1961	$2.85	$28.68	1987	$17.75	$46.99	2013	$91.17	$117.73
1962	$2.85	$28.34	1988	$14.87	$37.88	2014	$85.60	$108.73
1963	$2.91	$28.61	1989	$18.33	$44.62	2015	$41.85	$53.12
1964	$3.00	$29.10	1990	$23.19	$53.24	2016	$36.34	$45.50
1965	$3.01	$28.72	1991	$20.20	$44.62	2017	$43.97	$53.96
1966	$3.10	$28.74	1992	$19.25	$41.27	2018	$57.77	$69.20
1967	$3.12	$28.15	1993	$16.75	$34.90	2019	$50.01	$58.84
1968	$3.18	$27.46	1994	$15.66	$31.78	2020	$32.25	$37.45
1969	$3.32	$27.24	1995	$16.75	$33.07	2021	$60.84	$67.41
1970	$3.39	$26.28	1996	$20.46	$39.22	2022	$87.40	$89.90
1971	$3.60	$26.75	1997	$18.64	$34.94	2023 (Partial)	$70.58	$70.78

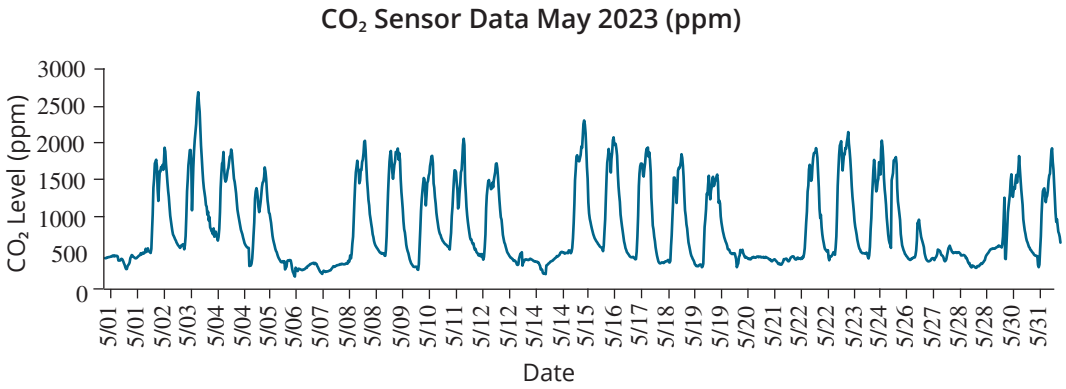
Data

The data set is available on stat.hawkeslearning.com under **Discovering Statistics and Data, Fourth Edition > Data Sets > Domestic Crude Oil Prices**.

Data

The data set is available on stat.hawkeslearning.com under **Discovering Statistics and Data, Fourth Edition > Data Sets > CO_2 Sensor Data**.

11. The data depicted below is one month of CO_2 sensor data measured in parts per million (ppm) levels at an office complex. Is this data stationary or non-stationary? What is the level of measurement of the data?



CR Chapter Review

Key Terms and Ideas

- Big Data
- Volume
- Variety
- Velocity
- Veracity
- Scale
- Measurement
- Level of Measurement
- Ratio Data
- Interval Data
- Ordinal Data
- Nominal Data
- Fuzzy Concepts
- Confounding Variable
- Peer Review
- The Scientific Method
- Experimental Design
- Controlled Experiment
- Control Group
- Experimental Group
- Treatment
- Explanatory Variable
- Response Variable
- Completely Randomized Design
- Before and After Study
- The Placebo Effect
- Double-Blind Study
- Single-Blind Study
- Observational Study
- Simpson's Paradox
- Surveys
- Bias
- Structured Data
- Unstructured Data
- Qualitative Data
- Quantitative Data
- Discrete Data
- Continuous Data
- Time Series Data
- Stationary Process
- Nonstationary Process
- Trend
- Cross-Sectional Data

Example 3.1.5

Creating a Cumulative Relative Frequency Distribution of Survey Results

Determine the cumulative relative frequency for the July customer satisfaction survey data from Example 3.1.1.

July Customer Satisfaction Survey			
Response	**Frequency**	**Cumulative Frequency**	**Cumulative Relative Frequency**
1–Very Dissatisfied	0	0	$\frac{0}{20}=0$
2–Somewhat Dissatisfied	5	5	$\frac{5}{20}=0.25=25\%$
3–Neutral	7	12	$\frac{12}{20}=0.6=60\%$
4–Somewhat Satisfied	6	18	$\frac{18}{20}=0.9=90\%$
5–Very Satisfied	2	20	$\frac{20}{20}=1=100\%$

From the cumulative relative frequency distribution it is apparent that 60% of customers from the July survey do not have a favorable opinion of the product.

Example 3.1.6

Creating a Cumulative Relative Frequency Distribution for the Heart Rate Data

Determine the cumulative relative frequency distribution for the heart rate data.

Solution

Heart Rate Cumulative Relative Frequency		
Heart Rate	**Relative Frequency**	**Cumulative Relative Frequency**
57–66	0.04	0.04
67–76	0.20	0.24
77–86	0.64	0.88
87–96	0.10	0.98
97–106	0.02	1.00

From the cumulative relative frequency distribution it is easy to see that 88% of the students had heart rates less than or equal to 86 beats per minute.

3.1 Exercises

Basic Concepts

1. What does it mean for a set of data to have a distribution?
2. Describe the purpose of a frequency distribution.

3. What are the basic questions to ask when examining the structure of a data set?
4. What are the three steps to constructing a frequency distribution?
5. In the construction of a frequency distribution, what are the two requirements that the classification categories must meet?
6. What are the fundamental decisions in constructing frequency distributions for quantitative data?
7. Describe the general guidelines for selecting the number of classes for a quantitative frequency distribution.
8. What is a good starting point for determining the class width?
9. What is a relative frequency distribution? How do you calculate relative frequencies from raw frequencies?
10. What is a cumulative frequency distribution? How do you calculate a cumulative frequency distribution?
11. What is a cumulative relative frequency distribution? What are two ways to determine a cumulative relative frequency distribution?

Exercises

12. Parkinsonism is an affliction of the aged and is frequently caused by Parkinson's disease, Alzheimer's disease, or other illnesses. The results from a recent study on Parkinsonism were reported in "Prevalence of Parkinsonian Signs and Associated Mortality in a Community Population of Older People," *New England Journal of Medicine*.[2] A sample of 467 people, all 65 years of age or older, was selected from East Boston, Massachusetts. Each person was clinically evaluated and various signs of Parkinsonism, if any, were noted. The following table is a frequency distribution for some of the signs of Parkinsonism.

Signs of Parkinsonism	
Sign	**Frequency**
Reduced arm swing	210
Prolonged turning	153
Right leg rigidity	141
Left leg rigidity	154
Slow finger taps	197
Shuffling gait	83

a. What level of measurement does the data possess?

b. What percent of the sample suffered from left leg rigidity? Round your answer to two decimal places.

c. Add up the frequencies. Why does the sum of the frequencies exceed the total sample size of 467?

d. Suppose 30 people suffer from both left leg rigidity and right leg rigidity. How many people in the sample suffer from rigidity in at least one of their legs?

e. Would it be reasonable to create a cumulative distribution for this data? Explain.

13. A small commuter airline in the West keeps records of complaints received from its customers. Complaints for March and July are listed in the following table.

Customer Complaints		
Type of Complaint	**March**	**July**
Tickets cost too much	11	15
Stewardess did not provide blankets	8	3
Schedules not convenient	12	17
Plane often late	17	16
Seats too stiff	3	3
Airplane too hot	6	20
Airplane too cold	8	5
Poor reservation system	5	5
Plane interior looks shabby	5	6

a. Classify the items by the following categories: comfort, price, service, and schedule, and develop a qualitative frequency distribution.

b. Classify the items by the following categories: plane, personnel, building/equipment, and other, and develop a qualitative frequency distribution.

c. Would another person necessarily assign the same items to the same categories as you have?

d. Do the categories chosen in parts **a.** and **b.** meet the requirement that categories be mutually exclusive and exhaustive? Discuss.

14. Before purchasing a specific product, a consumer reviews the ratings provided by previous customers. The following table displays the product ratings given by customers on a scale of 1 star (worst) to 5 stars (best).

Number of Stars	Frequency
1	9
2	11
3	30
4	17
5	8

a. What level of measurement does the data possess?

b. How many total customers rated the product?

c. What percentage of customers rated the product with 4 stars?

Data

stat.hawkeslearning.com under **Discovering Statistics and Data, Fourth Edition > Data Sets > Mount Pleasant Real Estate Data**

15. Using the *House Style* variable from the Mount Pleasant Real Estate data set from the web resource, consider the following.

a. What level of measurement does the data possess?

b. Is the data qualitative or quantitative?

c. Use technology to construct a frequency distribution for the *House Style* variable.

16. The following data represents the area, in square feet, of thirty homes listed for sale in Washington County, Wisconsin.

1152	1200	1216	1331	1408	1474	1479	1492
1508	1560	1647	1652	1654	1654	1670	1843
1905	1920	1924	2030	2165	2169	2202	2207
	2314	2498	2710	3234	3251	3440	

a. What level of measurement does the data possess?

b. Determine the minimum number of classes necessary if a class width of 400 is used.

c. Construct a frequency distribution with five classes.

d. Construct a frequency distribution with eight classes.

e. Which frequency distribution do you think best displays the data? Explain.

17. A business magazine was conducting a study into the amount of travel required for mid-level managers across the U.S. Seventy-five managers were surveyed for the number of days they spent traveling each year.

Mid-Level Manager Travel						
Days Traveling	0–6	7–13	14–20	21–27	28–34	35 and above
Frequency	15	21	27	9	2	1

a. Construct a relative frequency distribution.

b. Construct a cumulative frequency distribution.

18. Every year, the average temperatures of 100 selected US cities are published by the National Oceanic and Atmospheric Administration (NOAA). The average temperature (°F) for the month of October for 15 randomly selected cities from the list of 100 are listed in the following table.[3]

Average Temperatures (°F)				
68.5	50.9	67.5	57.5	56.0
47.1	50.1	65.8	51.5	49.5
75.2	56.0	62.3	53.0	46.1

a. Construct a frequency distribution for the average temperatures for the month of October.

b. Construct a relative frequency distribution for the average temperatures for the month of October.

c. Construct a cumulative frequency distribution for the average temperatures for the month of October.

19. The average temperatures (°F) for the month of January for forty randomly selected cities are listed in the following table.

Average January Temperature (^{0}F)	Number of Cities
14-22	1
23-31	4
32-40	11
41-49	10
50-58	8
59-67	4
68-76	2

a. What level of measurement does the data possess?

b. Choose the interval that contains the average January temperature for Boston, Massachusetts which is 36°F.

c. What percentage of cities have an average January temperature that is less than 32°F?

d. What percentage of cities have an average January temperature of 50°F or greater?

e. Explain how to use the relative frequency to determine the answer in part **c**.

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > California DDS Expenditures

20. Using the California DDS Expenditures data set from the web resource, perform the following.

a. Construct a frequency distribution with 8 classes for the *Expenditures* variable.

b. Construct a relative frequency distribution for the expenditures using the frequency distribution from part **a**.

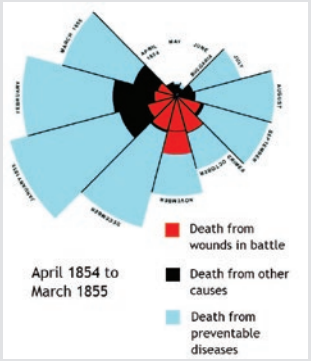
c. Construct a cumulative frequency distribution for the expenditures.

3.2 Displaying Qualitative Data Graphically

Graphical analysis presents a trade-off: although we lose sight of the individual observations (the raw data), it allows us to create a representation of the data distribution that "speaks to the eyes." The trade is almost always beneficial since a well-designed graph gives our visual processing system the kind of image it processes best, a picture.

Because a set of data can be graphically represented in many different ways, selecting and creating graphical displays requires a certain amount of artistic judgment.

Several types of graphs and tabular displays will be discussed in this chapter. Bar charts, stacked bar charts, and pie charts are effective, visually appealing methods of graphically displaying qualitative data. A quick look at publications such as *Time*, *USA Today*, *The Wall Street Journal*, *Scientific American*, and *Forbes* provides convincing evidence of the frequent and beneficial usage of these data visualization techniques.



A Passion for Compassion

In the 19th century, statistics was not widely seen as an applicable skill. That is, until Florence Nightingale came onto the scene. When she arrived at the front line of the Crimean War, she was appalled by the situation. The mortality rate was too high, and the hospitals were in complete disarray. She immediately set about organizing what little records were kept and started to gather a lot of new data. Upon analyzing this new data, she discovered that the majority of deaths that were occurring in British military hospitals were due to preventable diseases. Using this new information, Nightingale was able to present a case to Parliament for improving the sanitary practices in British hospitals. She utilized data analysis and visualization to literally save thousands of lives, and in the process, her "rose diagram", also known as a "coxcomb chart", became an iconic data visualization.

A pie chart showing where government revenues originate is given in Figure 3.2.11. In order to determine from the pie chart how much was received in a particular category, multiply the total amount by the percentage given for that category in the pie chart. For example, to find the amount of government receipts contributed by individual income taxes, multiply the total amount of government receipts ($4.738 trillion) by the percentage contributed by individual income taxes (49.3%). This means that $4.738 \cdot 0.493 = \$2.3358$ trillion of all government receipts come from individual income taxes.

Although pie charts are useful displays of categorical data, they have their limitations. Once the number of categories rises above ten, the information conveyed by a pie chart is less meaningful. As Figure 3.2.12 and Figure 3.2.13 demonstrate, when the number of employees increases from 5 to 40, using a pie chart to show the percentage of sales for each employee is not as informative.

Percentage of Sales
(5 Employees)

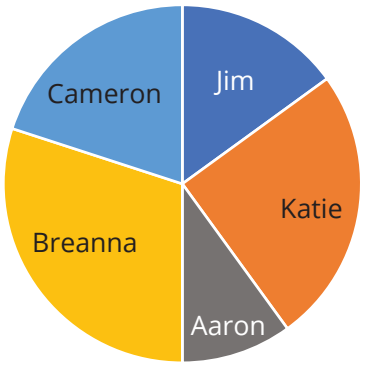


Figure 3.2.12

Percentage of Sales
(40 Employees)

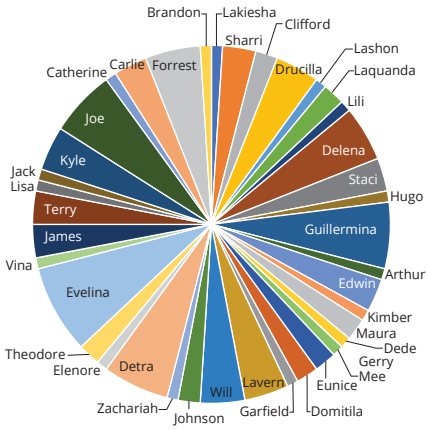


Figure 3.2.13

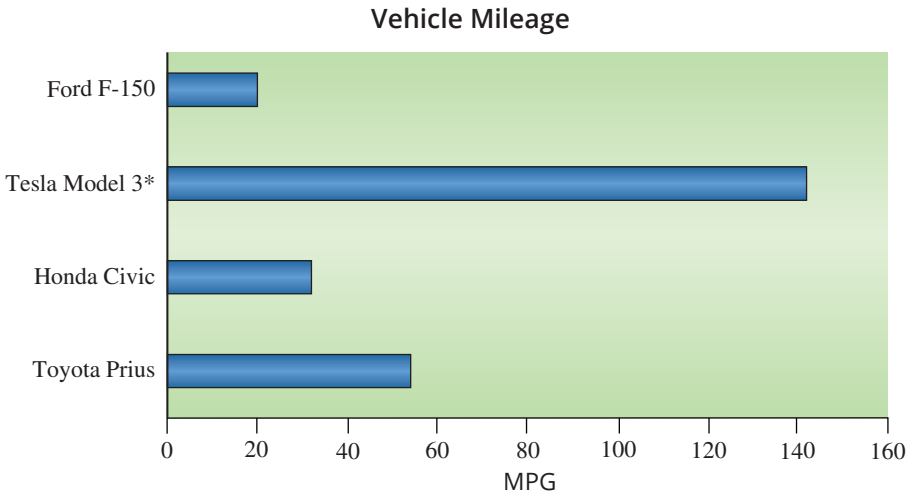
3.2 Exercises

Basic Concepts

1. What are some benefits of graphing?
2. What is the major disadvantage of graphing?
3. Describe the types of data that a bar chart would be useful in displaying.
4. Where should miscellaneous categories be displayed in a bar chart?
5. Explain how axis scales on bar charts can be misleading.
6. What is a stacked bar chart?
7. In what circumstances would a stacked bar chart be preferred over a standard bar chart?
8. What is a pie chart?
9. In Figure 3.2.7, the displayed data is not normalized. How could you normalize this data?

Exercises

10. A consumer magazine uses bar charts to compare four popular brands of automobiles. This particular bar chart represents a comparison of the miles per gallon (mpg) for the four brands.



*MPGe, or miles per gallon equivalent, for electric vehicles

a. What is wrong with this picture?

b. Evaluate the bar chart using the guidelines suggested in the section on the aesthetics of bar chart construction.

11. The following frequency distribution is for a portion of the Employee Satisfaction data set.

Data

stat.hawkeslearning.com under **Discovering Statistics and Data, Fourth Edition > Data > Employee Satisfaction**

Frequency Distribution of Employee Salaries

Department	High Salary	Medium Salary	Low Salary
Accounting	74	358	335
Management	225	180	225
Marketing	80	402	376
Sales	269	2099	1772
Technical	201	1372	1147
Total	**849**	**4411**	**3855**

a. Construct a bar chart for the number of employees with high salaries by department.

b. Construct a bar chart for the number of employees with medium salaries by department.

c. Construct a bar chart for the number of employees with low salaries by department.

d. Construct a stacked bar chart for the number of employees with low, medium, and high salaries by department.

e. Construct a pie chart for the number of employees with high salaries.

f. Construct a pie chart for the number of employees with medium salaries.

g. Construct a pie chart for the number of employees with low salaries.

12. The Centers for Disease Control and Prevention (CDC) reports information about the leading cause of death for the United States each year. The frequency table displays the data from 2021.[6]

Leading Causes of Death in the U.S., 2021	
Cause of Death	**Frequency**
Heart disease	695,547
Cancer	605,213
COVID-19	416,893
Accidents (unintentional injuries)	224,935
Stroke (cerebrovascular diseases)	162,890
Chronic lower respiratory diseases	142,342
Alzheimer's disease	119,399
Diabetes	103,294
Chronic liver disease and cirrhosis	56,585
Nephritis, nephrotic syndrome, and nephrosis	54,358

a. Construct a bar chart for the leading causes of death in the United States in 2021.

b. Did you create a Pareto chart? Explain your response.

c. What did you learn from the chart?

13. Consider the following data regarding professions with the highest projected percent of change in employment for the years 2021 through 2031.[7]

Occupation Growth Rates	
Occupation	**Projected Increase 2021–2031**
Nurse practioners	46%
Wind turbine service technicians	44%
Ushers, lobby attendants, ticket takers	41%
Motion picture projectionist	40%
Cooks, restaurant	37%
Data scientist	36%
Athletes and sports competitors	36%
Information security analysts	35%

a. Construct a bar chart for the projected growth rates of the various occupations.

b. What did you learn from the chart?

14. Consider the following data regarding the payment methods which consumers reported using during October 2020. BANP: bank account number payment. OBBP: Online banking bill pay. Other includes Paypal, account-to-account transfers, mobile payments, and deductions from income.

Payment Methods	
Method of Payment	**Relative Frequency**
Cash	19%
Check	7%
Credit Card	27%
Debit Card	28%
Pre-paid Card	2%
BANP	7%
OBBP	5%
Other	5%

a. Construct a bar chart for the relative frequencies of the various methods of payment.

b. Construct a pie chart for the relative frequencies of the various methods of payment.

c. Comment on any information about the relative frequencies of the various methods which you were able to ascertain by examining the charts.

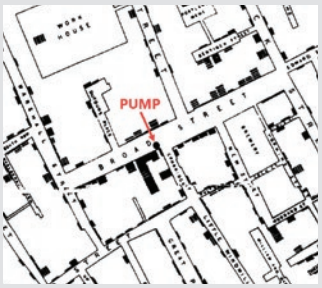
3.3 Displaying Quantitative Data Graphically

Quantitative graphs are powerful tools used in data visualization to represent and analyze numerical data. They come in various forms, including bar graphs, line graphs, histograms, scatter plots, dot plots, pie charts, and choropleth maps, each with its unique utility. These graphs help to summarize large data sets, display patterns, trends, and correlations, and make complex information more comprehensible. Examining raw data as compared to viewing a well-designed graph of that data is akin to hearing a symphony described in words versus actually listening to the music; the latter offers an immediate, rich understanding that simple description can never capture.

Histograms

A **histogram** is a common graphical method that reveals the distribution of a set of data. Histograms are often constructed based on frequency distributions of quantitative data. They look similar to bar graphs but are used to analyze quantitative data rather than qualitative data.

Histograms literally show the shape of the data by graphically displaying a frequency distribution.



Absence of Evidence is Not Evidence of Absence

During the London cholera outbreaks of the mid-1800s, thousands of people died within a relatively short period. At the time, the prevailing theory regarding how cholera was spread was called the miasma theory. It stated that the disease was spread through "bad air" that emanated from rotting organic matter. However, Dr. John Snow suspected that unsanitary water from the River Thames was the true culprit. Unfortunately, germ theory had not been developed yet, so Dr. Snow didn't fully understand how the alternative transmission method worked. In 1854, Dr. Snow utilized sampling and data visualization to illustrate that most of the cholera outbreaks happening at the time were occurring in houses that were close to the water pump on Broad Street. Still, the skeptics endured. However, even though his examination of the water was absent of evidence for harmful microbes, that does not mean that the microbes themselves were absent. Over a decade later, Louis Pasteur would officially propose germ theory, vindicating the work of Dr. Snow.

Table 3.3.6 - Percentage of Obese Adults by US County

FIPS Code	Percentage of Obese Adults, 2016 Normalized Values
1001	30.5
1003	26.6
1005	37.3
1007	34.3
1009	30.4
. . .	

Using our new data, we generate the following map.

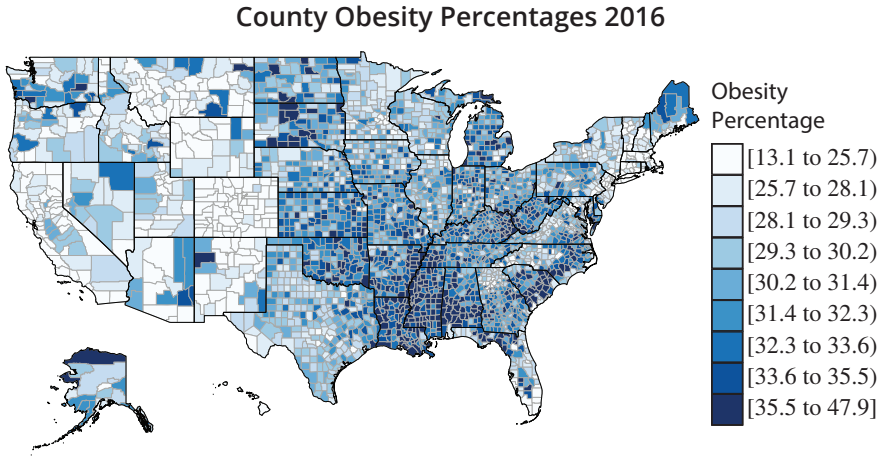


Figure 3.3.14

There are over 3000 counties in the U.S. The incredible aspect of these graphs is you can literally see the 3000+ data points and their geographic distribution. If we use Figure 3.3.13 to come to a conclusion about the geographic distribution of obesity in the U.S., we would have assumed that the Southwest and Northeast regions are the areas in the country that struggle with obesity. Comparing Figure 3.3.14 to Figure 3.3.13 produces an entirely new set of conclusions. Once we normalize the obesity variable by making it a ratio of the total county population, Figure 3.3.14 suggests that it is actually the Southeast and the Midwest that have the greatest struggle with obesity. In many instances, the data you start with will need to undergo some form of transformation to make it valuable for your intended purposes or goals.

3.3 Exercises

Basic Concepts

1. What is the main characteristic of data that a histogram reveals?
2. Describe the type of data that could be usefully described with a histogram.
3. True or false: A frequency distribution contains all of the information needed to construct a histogram.
4. List the important features to look for when studying a histogram.

5. Explain why the stem-and-leaf plot is sometimes called a "hybrid graphical method."
6. Identify the advantages of a stem-and-leaf plot.
7. Consider the following data value: 39. What would be the stem and the leaf for this value if we identified the stem as the tens digit? What would be the stem and the leaf if we identified the stem as the hundreds digit?
8. When constructing a stem-and-leaf plot, how do you determine which part to make the stem and which part to make the leaf?
9. What is an ordered array?
10. What are some advantages of the ordered array?
11. Dot plots are most useful for what types of data?
12. What are some advantages of using a dot plot?
13. How can the most frequently occurring value be identified by studying a dot plot?
14. Why is it important to plot time series data?
15. The time variable is always graphed on which axis?
16. What is a choropleth map?

Exercises

17. The weights (in pounds) for the players on an NFL football roster are shown below.

153	183	185	189	190	190	195	196	198	200	202	204
205	207	209	212	213	215	220	220	225	225	228	233
235	236	238	243	244	245	246	247	248	254	254	255
255	256	257	265	268	280	298	302	305	308	310	311

   a. Construct a frequency distribution for the weights of players on the roster.

   b. Construct a histogram for the weights of players on the roster.

18. Using the OECD Better Life 2022 data set from the web resource, create a frequency distribution and histogram for the variable *Voter Turnout* and use them to answer the following questions. For the histogram and frequency distribution, use 40% for the minimum value and use class widths of 10%.

   a. What is the level of measurement of the variable?

   b. Construct a relative frequency distribution for voter turnout.

   c. How many of the countries have a voter turnout of 70% or greater?

   d. What percent of the countries have less than a 50% voter turnout?

   e. What percent of the countries have a voter turnout of 80% or greater?

Data

stat.hawkeslearning.com under **Discovering Statistics and Data, Fourth Edition > Data Sets > OECD Better Life Index 2022**

19. A chemist is interested in knowing the amount of alcohol contained in American-brewed beers. To study this, the chemist uses data containing information about several different kinds of American-brewed beers, and evaluates the alcohol by volume for each. Using the Beers and Breweries data set from the web resource, perform the following:

   a. Construct a frequency distribution for the alcohol by volume (ABV) variable. Use 0.001 for the minimum value and 0.130 for the maximum value. Use bin widths of 0.010.

Data

stat.hawkeslearning.com under **Discovering Statistics and Data, Fourth Edition > Data Sets > Beers and Breweries**

b. Construct a relative frequency distribution for the ABV. Round the relative frequencies to four decimal places.

c. Construct a histogram of the relative frequency distribution.

d. Comment on any information about the alcohol by volume in American-brewed beers which you were able to ascertain by examining the distributions and the histogram.

20. Some final heat times (in seconds) for the boys 100-meter races at the 2023 Georgia state track meet are listed below.[17]

10.72	10.75	10.78	10.79	10.80	10.83	10.86	10.94
10.97	10.97	10.99	11.00	11.00	11.02	11.03	11.04
11.05	11.05	11.08	11.09	11.11	11.14	11.17	11.22
11.23	11.24	11.25	11.27	11.40	11.40	11.46	11.60

a. Construct a frequency distribution for the 100-meter race times.

b. Construct a relative frequency distribution for the 100-meter race times.

c. Construct a histogram of the relative frequency distribution.

d. Comment on any information about the 100-meter race times which you were able to ascertain by examining the distributions and the histogram.

21. To attend a friend's wedding this summer, you'll fly from the Hartsfield-Jackson Airport in Atlanta, GA to Denver, CO. The following data shows the cost (in dollars) of several flights that are compatible with your travel plans.

209	198	188	198	227	246	256	220
198	205	246	198	227	199	198	194
198	188	198	209	188	198	231	205

a. What level of measurement does the data possess?

b. Construct a stem-and-leaf display for the data using the tens digits as the stems.

c. Comment on the shape of the distribution.

22. The following data shows the credit scores of recent lease applicants at an apartment complex.

645	756	668	590	713	647	811	725	806
675	632	740	583	689	739	791	826	670
782	654	619	672	689	702	717		

a. Construct a stem-and-leaf display for the data using the hundreds digits as the stems.

b. Construct a histogram using the classes 500-599, 600-699, 700-799, 800-899.

c. Comment on the shape of the distribution. What do you notice about the shapes of the stem-and-leaf display and the histogram of the credit score data?

d. Applicants with a credit score above 700 can remit a reduced security deposit. What percentage of these applicants have a credit score above 700?

e. Individuals with a credit score below 660 are considered "subprime" and will likely not qualify for the lease. What percentage of applicants have a credit score below 660?

23. Daily high temperatures (°F) in June 2022 for two US cities are shown in the side-by-side stem-and-leaf display shown below.

Charleston, SC	Stem	Milwaukee, WI
	6	1 3 5 6 7 9
2	7	4 4 4 5 5 6 6 6 8 9 9
2 2 2 2 2 2 4 4 4 4 4 6 6 6 6 6 8 8 8	8	0 0 0 5 7 8 8
0 0 0 1 1 3 3 3 5 5	9	2 4 4 5 9
0	10	0

a. What level of measurement does the data possess?

b. Based upon the stem-and-leaf display, compare the June temperatures for the two cities. State several observations about the ways that the distributions are similar and different.

c. Suppose that someone does not like extremely warm summer temperatures. What percentage of days had temperatures less than 90 degrees in Charleston? in Milwaukee?

d. What was the most frequent temperature in Charleston?

24. *Fortune* magazine publishes a list of the top 100 best companies to work for. For the top 10 companies on this list, the average annual employee salaries are given in the following table (in thousands of dollars).

Average Salaries (Thousands of Dollars)									
121	122	136	74	118	101	114	61	95	132

a. Construct a stem-and-leaf display for the data using the tens digits as the stems.

b. Comment on any information about the average annual salaries (in thousands of dollars) of the top 10 companies which you were able to ascertain by examining the stem-and-leaf display.

c. Construct an ordered array of the average annual salaries in rank order.

d. Does the ordered array provide any additional insight into the nature of the data?

25. The pH level of drinking water obtained from a well was regularly tested. The recorded data is as follows:

7.1	6.5	6.9	6.6	7.0	7.3	7.1	6.8	6.8	6.5
6.7	6.9	6.8	6.5	6.3	6.6	6.7	7.0	7.2	6.9
6.8	6.6	6.9	6.8	7.1	6.8	6.7	6.5	6.8	7.0

a. Construct a dot plot of the data.

b. Which data value occurs most often?

c. The recommended pH levels for drinking water should fall in the range of 6.5 to 8.5. Is that true of this data?

d. Does this well water tend to be acidic or alkaline? Explain your response.

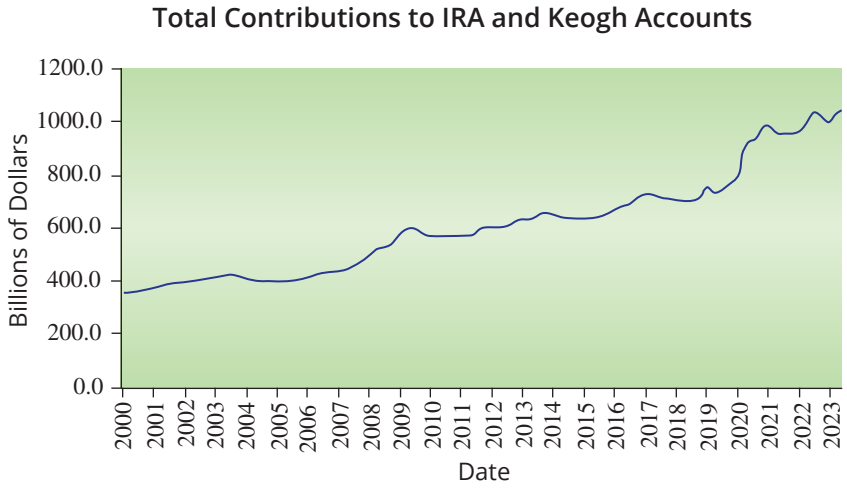
26. Listed in the following table is the number of passing attempts per game by Super Bowl champion Patrick Mahomes in the 2022 NFL regular season.[18]

a. Construct a dot plot of the data.

b. Which data value occurs most often?

Passing Attempts by Patrick Mahomes								
39	35	35	37	43	40	34	68	35
34	42	27	42	41	28	42	26	

27. The following line graph displays the total IRA and Keogh accounts (in billions of dollars) in the U.S., charted from January 2000 to May 2023.[19]

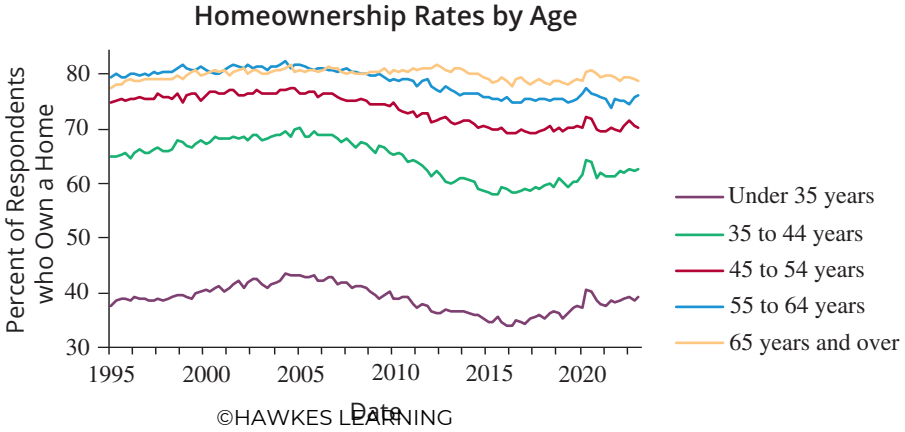


a. What conclusions can you make regarding the total contributed to the accounts?

b. Is the data time series data?

c. If the data is time series data, is the series stationary or nonstationary?

28. The following chart displays the homeownership rate data collected by the United States Census Bureau for January 1995 through January 2023. The percentage of owner-occupied housing units is reported for various age ranges.[20]



a. Examine the graph and discuss the data. What conclusions can you make?

b. If the data is time series data, is it a stationary or nonstationary time series? Explain your reasoning.

29. The unemployment rate is a key economic indicator that measures the percentage of the labor force that is unemployed and actively seeking employment. The unemployment rates for the state of North Carolina from January 2010 to January 2022 are given in the table below.[21,22]

North Carolina Unemployment Rate							
Year	2010	2011	2012	2013	2014	2015	2016
Percent	11.2	10.4	9.7	9.4	6.4	5.7	5.3
Year	2017	2018	2019	2020	2021	2022	
Percent	4.9	4.2	4.0	3.8	5.6	3.6	

a. What is the level of measurement of the unemployment rate data?

b. Construct a time series plot for the data.

c. What conclusions can you make from the plot?

3.4 Analyzing Graphs

Graphs that help us visualize data can either be enlightening, in the sense that they give us insight and understanding of a set of data, or misleading, either intentionally or unintentionally. When you see graphs in the media, you need to be cautious to ensure the data has been accurately represented by the graph. This section will help you analyze graphs for accuracy and appropriate presentation of the given information. Here are a few key ideas to consider when interpreting information displayed in graphical form.

Graph Labeling

Every graph should be properly labeled with an appropriate title that tells you what type of information is being displayed. Also, if the graph has a horizontal and vertical axis, these should be labeled and should include the unit of measurement when necessary for the understanding of the data. For example, in Figure 3.4.1, the title does not provide enough information about the data. Why were those countries chosen? Do they have relatively high or low prison populations compared to the rest of the world? Furthermore, we do not know whether this information is relevant to modern times. Is this data for a specific year? The countries are labeled along the horizontal axis but note that the vertical axis is just labeled *Population*. We have no idea what the values along the vertical axis represent. Is the prisoner population in units of thousands, millions, or billions? In fact, this chart shows the countries with the top ten highest prisoner populations for the year 2021. The unit for the vertical axis should be thousands, which means that the United States had a prison population of approximately 1690 thousand, or 1.690 million, in the year 2021. Without these seemingly small pieces of information, the graph is not very informative. It is also good practice to use the largest possible unit for the scale of an axis, which in this case is correctly chosen to be thousands.

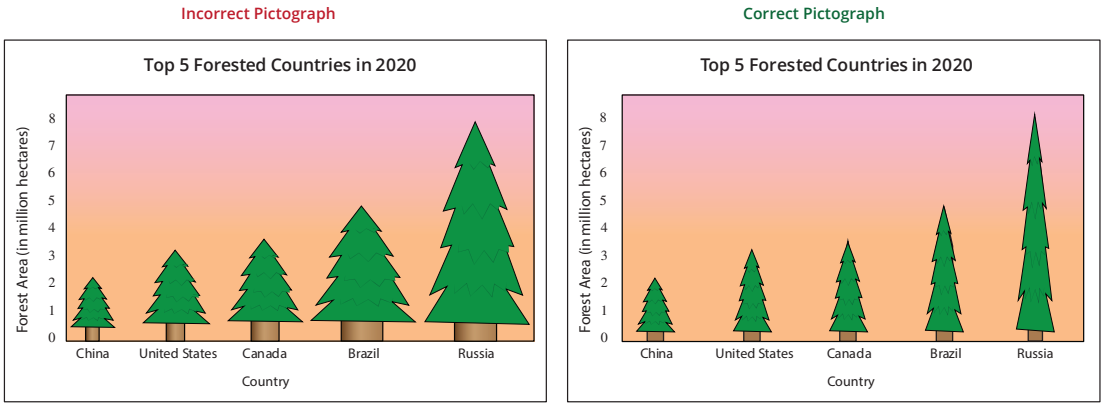


Figure 3.4.10

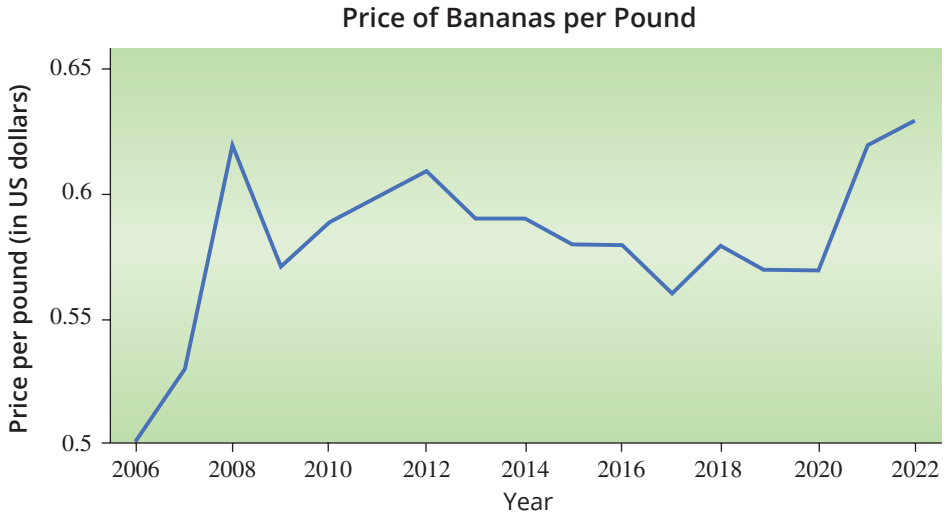
3.4 Exercises

Basic Concepts

1. Why is it important to label and title graphs properly?
2. What types of sources are generally reliable?
3. Why is the scaling of a graph important?
4. Why are data transformations useful?

Exercises

5. Do you see any issues with the scales used on the axes of the graph depicting banana prices per pound in July?[26] Why or why not?



6. Using the San Francisco Salaries 2014 data set from the web resource, create a histogram for the variable *TotalPayBenefits* and answer the following:

a. Does the distribution of the data in the histogram look bell-shaped, skewed right, or skewed left?

b. Construct a new histogram for the variable *LogTotalPayBenefits*, which is a log transformation of the variable *TotalPayBenefits*.

c. Does the distribution of the data in the log transformed histogram look bell-shaped, skewed right, or skewed left?

Data

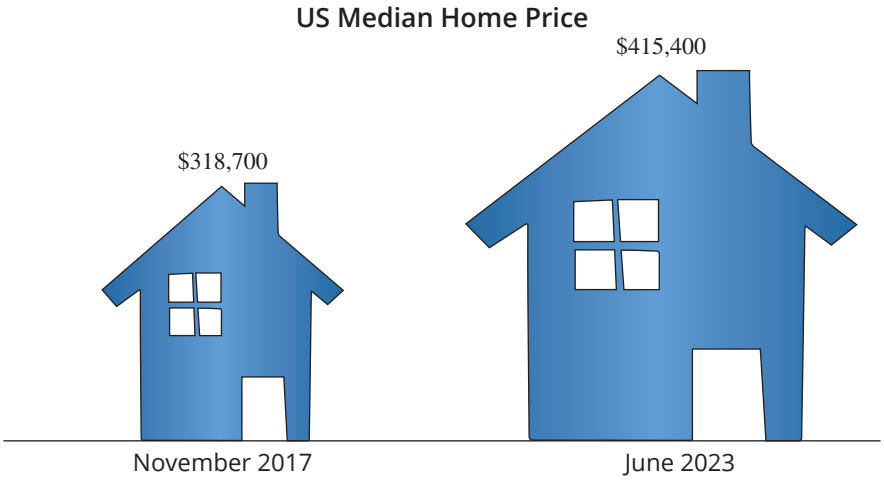
stat.hawkeslearning.com under **Discovering Statistics and Data, Fourth Edition > Data Sets > San Francisco Salaries 2014**.

7. The US median home price increased from $318,700 in November 2017 to $415,400 in June 2023, as shown in the following pictograph.[27]

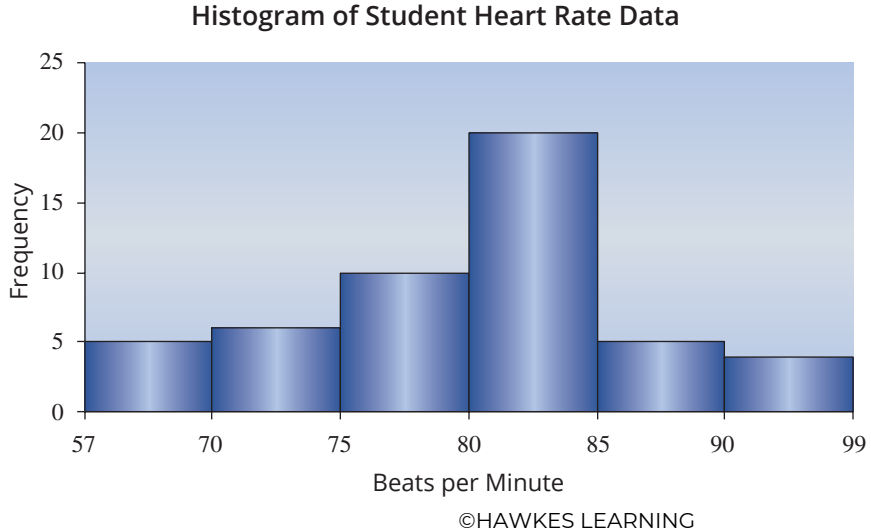
a. What was the percentage increase in US median home price between November 2017 and June 2023?

b. Is the pictograph shown an accurate depiction of this increase? Why or why not?

c. How could you improve the pictograph so that it accurately represents the information?



8. The following histogram of the heart rate data presented earlier in the chapter has a different number of classes than were used in the histogram in Figure 3.1.3. What errors can you find in the histogram?

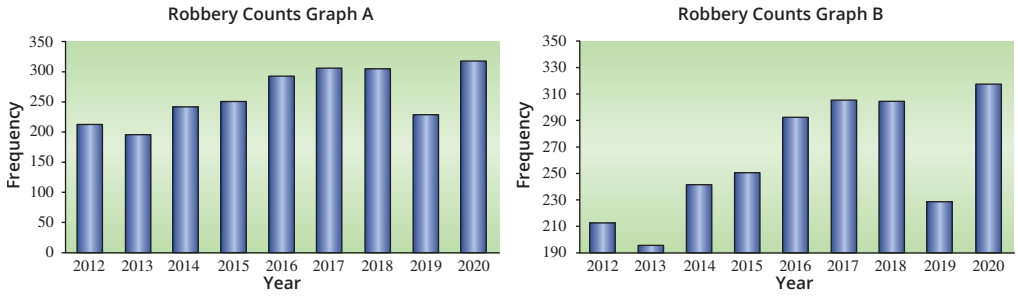


9. The number of robberies in North Charleston, SC is depicted in two different graphs below.[28] Use these graphs to answer the following.

a. Which graph do you feel better represents the data? Why?

b. What transformation could you apply to the robbery data that would normalize it?

c. Approximately how many times taller is the 2020 bar compared to the 2012 bar in Graph B? How many times more robberies were there actually in 2020 compared to 2012?



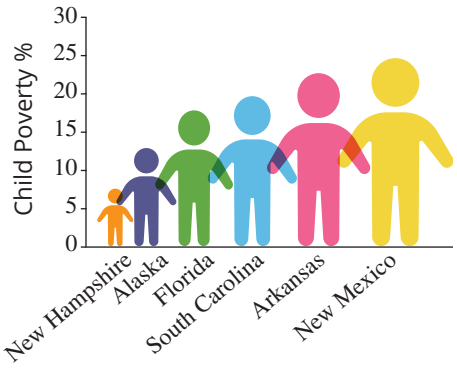
10. Use the pictograph on child poverty in six states of the United States to answer the following questions.

a. What state in the graph has the highest child poverty rate? Does this surprise you?

b. How much larger does the New Mexico child poverty rate appear compared to the New Hampshire child poverty rate?

c. Does the pictograph represent the child poverty percentage accurately? Why or why not?

State	Child Poverty %
Alaska	13.0
Arkansas	22.1
Florida	17.7
New Hampshire	7.1
New Mexico	24.9
South Carolina	19.7



Data

stat.hawkeslearning.com

Discovering Statistics and Data, Fourth Edition > Data Sets > COVID-19 Deaths by Country

11. Using the COVID-19 Deaths by Country data set from the web resource, create a histogram for the variable *Deaths*.

a. Is the *Deaths* variable in the data normalized?

b. Construct a new histogram for the variable *LogDeaths*, which is a log transformation of the variable *Deaths*.

c. Does the distribution of the data in the log transformed histogram look bell-shaped, skewed right, or skewed left?

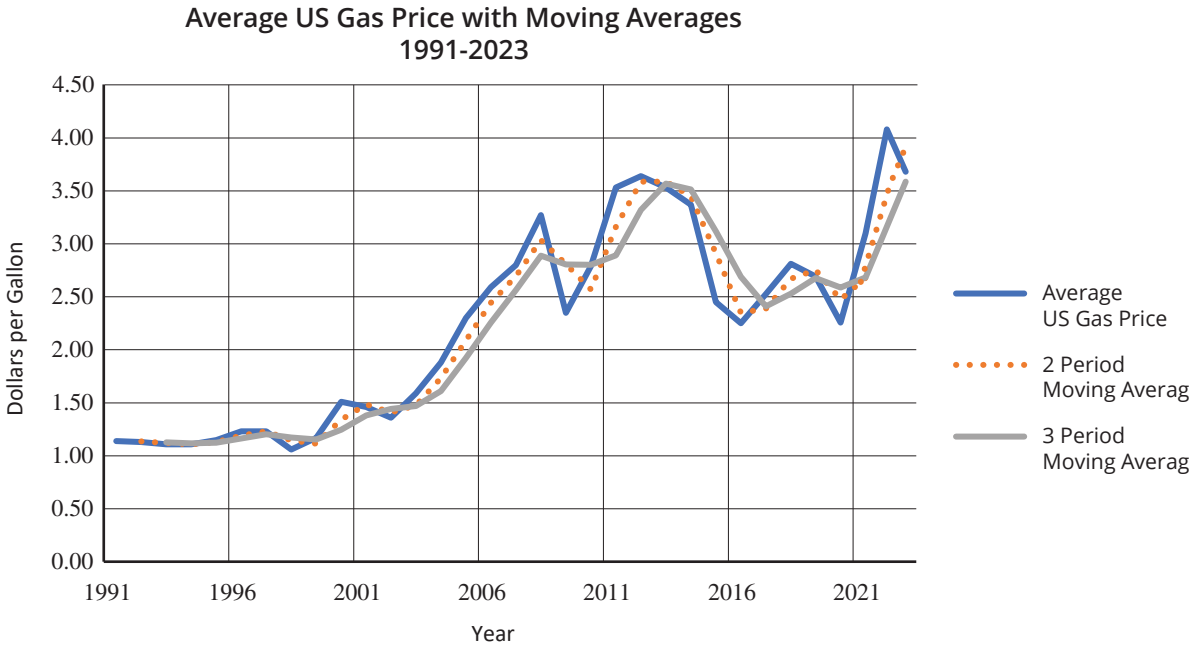


Figure 4.1.7

4.1 Exercises

Basic Concepts

1. Describe the difference between statistics and parameters.
2. Describe three major attributes used in summarizing a data set.
3. What are numerical descriptive statistics and why are they important?
4. Identify and describe five measures of location. List the advantages and disadvantages of each.
5. What is a resistant measure?
6. Describe a situation in which using the weighted mean as a measure of location would be appropriate.
7. What does it mean if we say that a data set is positively skewed? Negatively skewed?
8. Explain why the mean should not be calculated for a nonstationary time series.
9. What is a moving average? When is it useful?
10. When you compute the mean of a data set are you thinking inductively or deductively? How about the median, mode, and trimmed mean?
11. How does determining a statistic relate to empiricism?

Exercises

12. Calculate the mean, median, 10% trimmed mean, and the mode for the following data.

25	29	31	38	39	40	41	44	46	47
50	53	55	57	62	62	67	71	74	78

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > US Violent Crime 2020

13. Using the US Violent Crime data set found on the companion website, determine the mean, median, mode, and 20% trimmed mean for the year 2019. Round your answers to 2 decimal places.

14. Use the US Violent Crime rate data set from the previous problem.

a. Remove Washington, D.C. (District of Columbia) from the data and then determine the mean, median, mode, and the 20% trimmed mean for the year 2019. Round your answers to 2 decimal places.

b. Compare your results with those in exercise #13. Order the four measures of center from the measure with the least amount of change to the one with the greatest amount of change. Which measure of center was the most sensitive to the outlier (Washington, D.C.)?

15. Determine the mean, median, the 10% trimmed mean, and the mode for the following data.

2	22	6	18	10	14	12	12	16	8

16. Discuss the usefulness of each of the measures of central tendency with respect to the following situations.

a. A company is considering a move into a regional market for specialty soft drinks. In analyzing the size of the containers that its competitors are currently offering, would the company be more interested in the mean, median, or mode of their containers?

b. The creative director for an advertising agency is trying to target an ad campaign that will be shown in one city only. Would he be more interested in the mean or median family income in the city?

c. A young economist was assigned the task of comparing the interest rates on ninety-day certificates of deposit (CDs) in three major cities. Should she compare the mean, median, or modal interest for the banks in the three cities?

d. A telephone company is interested in knowing how customers rate their service: excellent, good, average, or poor. Would the company be more interested in studying the mean, median, or mode of the customer service ratings?

17. Discuss the usefulness of each of the measures of central tendency with respect to the following situations.

a. A doctor is interested in analyzing the increase in systolic blood pressure caused by a certain antibiotic. Would the doctor be more interested in studying the mean, median or mode of the systolic blood pressures?

b. A car manufacturer is trying to decide in what colors it should offer its new sports coupe. In analyzing the preferred colors of other sports coupes, would the manufacturer be more interested in the mean, median, or mode of the colors?

c. A manufacturer of chocolate bars is interested in knowing how people rate its chocolate: the best, above average, average, below average, or the worst. Would the company be more interested in the mean, median, or mode of the ratings?

d. A realtor is interested in studying the prices of recent home sales in an area which has many diverse neighborhoods. Would the mean, median, or mode of the prices of recent home sales be the best measure of central tendency?

18. Using the Amazon Stock Price data set from the companion website, perform the following calculations. Round your answers to 2 decimal places.

a. Compute and graph a 3-day moving average of the daily closing stock price for the years 1997 to 2017.

b. Describe any trends that you see in the data from the graph in part **a.**

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Amazon Stock Price

19. A stroke patient has been monitoring her blood pressure daily for 20 days since she was discharged from the hospital. The systolic blood pressure readings are recorded below.

Systolic Blood Pressure Readings									
100	104	117	95	98	111	105	106	115	101
101	102	115	116	113	103	104	119	127	132

a. What level of measurement does the data possess?

b. Compute the mean, 10% trimmed mean, and the 20% trimmed mean.

c. Which measure computed in part **b.** best describes the typical systolic blood pressure of the patient?

d. Would conclusions drawn from this data be an example of empiricism or rationalism?

Technology

To learn how to filter data, please visit stat.hawkeslearning.com and navigate to **Discovering Statistics and Data, Fourth Edition > Technology Instructions > Data Manipulation > Filtering**.

20. Using the CO_2 Emissions data set from the companion website, answer the following questions.

a. What level of measurement does the data possess?

b. For the United States, create a time series plot of CO_2 emissions with year on the horizontal axis. Does the data have trend?

c. How should we summarize this data, as a time series or with traditional summary measures of central tendency? Explain.

d. Would conclusions drawn from this data be an example of empiricism or rationalism?

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > CO_2 Emissions

21. Consider the following monthly sales for a small clothing store in a resort community.

Monthly Sales			
Month	**Sales ($)**	**Month**	**Sales ($)**
January	100,500	**July**	200,000
February	120,000	**August**	185,000
March	133,000	**September**	175,000
April	145,000	**October**	120,000
May	160,000	**November**	180,000
June	180,000	**December**	330,000

a. Draw a line graph of the data.

b. Compute the two-period moving averages for the data.

c. Compute the three-period moving averages for the data.

d. Add line graphs for the two-period moving averages and three-period moving averages to the graph which you constructed in part **a.**

e. Which series of data (the original sales data, the two-period moving averages, or the three-period moving averages) do you think best represents sales for the year? Why?

22. Consider the following average monthly balances for one bank customer for January through March. Compute the weighted average balance for the three-month period. Note that each average monthly balance must be weighted by the number of days in that month on a non-leap year. Round to the nearest cent.

Average Monthly Balances for a Bank Customer (January through March)	
Month	**Average Monthly Balance**
January	$1885.67
February	$1312.92
March	$2001.53

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > US County Data

23. Using the US County Data set on the companion website, determine the percentage of the total US population that has at least a high school diploma by utilizing the weighted mean. Use the *At.Least.High.School.Diploma* variable, which is the percentage of a county population with at least a high school diploma, as the data values, and use the *Total.Population* variable as the weights.

24. The following data represent the blood types of 20 patients on the surgical ward at a local hospital.

Blood Type of Surgical Patients			
Patient	**Blood Type**	**Patient**	**Blood Type**
1	O+	11	A+
2	O+	12	O−
3	A+	13	A−
4	B+	14	O+
5	O+	15	A+
6	A+	16	B−
7	B+	17	O+
8	A+	18	A+
9	AB+	19	O+
10	O+	20	O+

a. What level of measurement does the data possess?

b. What is the modal blood type of patients on the surgical ward?

4.2 Measures of Dispersion

Suppose all people looked alike, all cars looked alike, everyone wore the same kind of clothes, and there was only one kind of hamburger (plain). Without diversity it would be a boring world and a world in which statistics would be of little value. Since much of statistics is devoted to describing, analyzing, and explaining variability, understanding how variability is measured is essential to understanding statistics.

The concept of **variability** (also referred to as **dispersion** or **spread**) is as vague as the concept of central tendency. And vague concepts lead to different measurement ideas. The same issues that are important in evaluating location measures are meaningful in evaluating measures of dispersion.

Many of the good measures of dispersion use the concept of deviation from the mean. If the mean is a focal point or base, use it as a common basis from which to measure variation. The distance that a point is from its mean is called a **deviation from the mean**. A data set and its deviations from the mean are calculated in Table 4.2.1.

Table 4.2.1 – Calculating Deviations from the Mean	
Mean = 10	
Data Values	**Deviations from the Mean (Data − Mean = Deviation)**
3	$3 - 10 = -7$
12	$12 - 10 = 2$
20	$20 - 10 = 10$
15	$15 - 10 = 5$
0	$0 - 10 = -10$

4.2 Exercises

Basic Concepts

1. Describe three measures of variation. Discuss the strengths and weaknesses of each.
2. What does the standard deviation measure?
3. Why are the variance and standard deviation more commonly used as measures of variability than the MAD?
4. Explain how the variance can be construed as an average.
5. True or False: The variance and standard deviation are resistant measures.
6. When is it appropriate to calculate the variance of a time series?
7. Why is a "bell-shaped" distribution associated with the empirical rule?
8. What is the empirical rule? When is it appropriate to use? What statistical techniques would be useful in establishing appropriateness?
9. What is Chebyshev's Theorem?
10. Suppose you had a data set with 100,000 SAT scores and you wished to describe an equidistant interval around the mean with 75% of the scores. Aside from the empirical rule or Chebyshev's theorem, how could you define the interval? Hint: Think empirically.

Exercises

11. Since Super Bowl football games are a sample of all NFL football games, use the Super Bowl data set from the companion website to determine the following.
    a. Determine the sample variance of the variable *Winner_Rush Attempts*. Round your answer to three decimal places.
    b. Determine the sample standard deviation of the variable *Winner_Rush Attempts*. Round your answer to three decimal places.
    c. Determine the range of the variable *Winner_Rush Attempts*.
    d. What are some of the factors which might contribute to the variation in the observations?

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Super Bowl

12. Since Super Bowl football games are a sample of all NFL football games, use the Super Bowl data set from the companion website to determine the following.
    a. Determine the sample variance of the variable *Loser_Rush Attempts*. Round your answer to three decimal places.
    b. Determine the sample standard deviation of the variable *Loser_Rush Attempts*. Round your answer to three decimal places.
    c. Determine the range of the variable *Loser_Rush Attempts*.
    d. What are some of the factors which might contribute to the variation in the observations? Try to use different factors than the ones used in the previous exercise.

13. The interest rates on 30 year mortgages offered by seven randomly selected banks in a large metropolitan area are recorded below.

7.5% 8.0% 7.0% 7.25% 8.5% 8.25% 7.75%

a. Determine the sample variance of the interest rates.

b. Determine the sample standard deviation of the interest rates.

c. Determine the range of the interest rates.

d. What are some of the factors which might contribute to the variation in the observations?

14. A researcher has hypothesized that sophomore college students are more disciplined than freshman college students. The researcher believes that a reasonable measure of discipline is the performance on a statistics test in terms of both absolute score and consistency of scores. Seven freshman statistics students and seven sophomore statistics students are randomly selected and their scores on a statistics test are observed.

Statistics Test Scores							
Freshman	65	100	75	45	85	73	95
Sophomore	75	80	95	85	82	72	49

a. Determine the average test score for freshman students and sophomore students separately.

b. Determine the variance of the test scores for freshman students and sophomore students separately.

c. Determine the standard deviation of the test scores for freshman students and sophomore students separately.

d. Do you think that the data tend to support the hypothesis that sophomore college students are more disciplined than freshman college students based on the researcher's measurement?

e. What do you think about this particular measurement of discipline?

15. An initial investment of $135,000 was placed in each of 10 stocks and the stocks were divided into two portfolios. The market values of the stocks in each portfolio at the end of year one of the investment are presented below.

Market Values ($)					
Portfolio A	150,000	155,000	145,000	160,000	140,000
Portfolio B	130,000	175,000	100,000	150,000	195,000

a. What level of measurement does the data possess?

b. What statistical criteria might you use to select the better portfolio? Justify your answer.

c. Compute the statistics you proposed in part **b.**

d. Intuitively, which portfolio has the least amount of risk? Why?

16. Given the data set shown below.

81	99	97	81	85	86	99	93	96	83	82	91

a. Compute the mean and standard deviation for the original data.

b. Add 20 to each of the data points in the original data. Compute the mean and standard deviation for this adjusted data set and compare them to the mean and standard deviation of the original data. What do you notice?

c. Subtract 10 from each of the data points in the original data. Compute the mean and standard deviation for this adjusted data set and compare them to the mean and standard deviation of the original data. What do you notice?

d. Make an empirical generalization about the effect of adding or subtracting a constant value to each member of a data set on the mean and standard deviation of the data.

17. Given the data set shown below.

8	14	6	10	20	12	4	10	18	2

a. Compute the mean and standard deviation for the original data.

b. Multiply each of the data points by 10. Compute the mean and the standard deviation for the adjusted data set and compare to the mean and the standard deviation of the original data. What do you notice?

c. Divide each of the data points by 2. Compute the mean and standard deviation for the adjusted data set to the mean and the standard deviation for the original data. What do you notice?

d. Make an empirical generalization about the effect of multiplying or dividing a constant value from each member of a data set on the mean and standard deviation of the data.

18. The average score on a pre-employment test is 26 with a standard deviation of 7. Using Chebyshev's Theorem, state the range in which at least 88.89% of the data will reside.

19. The daily average number of phone calls to a call center is 972 with a standard deviation of 127. Using Chebyshev's Theorem, state the range in which at least 75% of the data will reside.

20. The length of a full-term human pregnancy is often defined as 40 weeks. From empirical data the length of time from conception to birth is on average 266 days with a standard deviation of 16 days. Use the empirical rule to determine the following ranges for the length of a human pregnancy.

a. What time interval (in days) will contain the lengths of approximately 68% of human pregnancies?

b. What time interval (in days) will contain the lengths of approximately 95% of human pregnancies?

c. What time interval (in days) will contain the lengths of approximately 99.7% of human pregnancies?

d. What assumption did you make in order to use the empirical rule in this scenario?

21. For people with diabetes, it is important to determine the level of glucose in the bloodstream, avoiding amounts that are too high or too low, to prevent negative health consequences. The measurements of glucose levels in the blood can be obtained through a finger stick with a lancet or by wearing a device with a subcutaneous probe attached to the skin. Both methods result in a tremendous amount of data used by medical professionals to monitor the person's health and prescribe treatment. Over a 4-week span of measurements, the blood glucose levels of one patient had a bell-shaped distribution with a mean of 114 mg/dl and a standard deviation of 24 mg/dl. Use the empirical rule to determine the following.

a. The percentage of measurements that the patient had blood glucose levels "in range", which is the desired target interval between 90 and 138 mg/dl.

b. The percentage of measurements that the patient had blood glucose levels less than 66 mg/dl, which is seriously low, categorized as hypoglycemia.

c. The percentage of measurements that the patient had blood glucose levels greater than 138 mg/dl, which is considered elevated and should be treated with an insulin dosage.

22. A management consulting firm is evaluating the salary structure for a large insurance company. The goal of the study is to develop salary ranges for each of the possible job grades within the company. The company and the firm have agreed that a reasonable salary range for each job grade can be determined by finding the salary range in which 95% of the current salaries for that job grade fall. The average salary and the standard deviation of the salaries are listed in the table below for three of the job grades.

Salary ($)			
Job Grade	**25**	**33**	**40**
$\bar{x}$	40,000	55,000	70,000
s	3,000	2,000	5,000

a. Determine the appropriate salary ranges for the three job grades.

b. What assumption did you make about the salaries in each of the job grades in answering part **a.**?

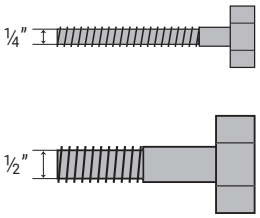
23. A consumer interest group is interested in comparing two brands of vitamin C. One brand of vitamin C advertises that its tablets contain 500 mg of vitamin C. The other brand advertises that its tablets contain 250 mg of vitamin C. Tablets for each brand are randomly selected and the milligrams of vitamin C for each tablet are measured with the following results.

Vitamin C Content (mg)		
	Brand A (500 mg)	**Brand B (250 mg)**
$\bar{x}$	500	250
s	10	7

a. Compute the coefficient of variation for Brand A.

b. Compute the coefficient of variation for Brand B.

c. Which brand more consistently produces tablets as advertised? Explain.



24. A manufacturer of bolts has two different machines. One machine is used to produce 1/4 inch bolts; the other machine is used to produce 1/2 inch bolts. It is very important that the machines consistently produce bolts of the correct diameters, or the bolts will not fit on the corresponding nuts. In order to compare the two machines, management randomly selects bolts produced from each machine and computes the average diameter of the bolts and the standard deviation of the bolts. The results of the study are shown in the table below.

Bolt Diameter		
	Machine X $\left(\frac{1}{4}"\right)$	**Machine Y** $\left(\frac{1}{2}"\right)$
$\bar{x}$	0.25"	0.50"
s	0.03"	0.05"

a. Compute the coefficient of variation for Machine X.

b. Compute the coefficient of variation for Machine Y.

c. Relative to their mean diameters, which machine more consistently produces bolts of the correct diameter? Explain.

4.3 Measures of Relative Position, Box Plots, and Outliers

Suppose you want to know where an observation stands in relation to other values in a data set. For example, on many standardized tests such as the SAT, GMAT, and ACT, the test scores themselves are rather meaningless unless they are associated with some measure that tells you how well you did relative to others taking the same test. There are two principal methods of communicating relative position: **percentiles** and ***z*-scores**. Both of these methods are data transformations which change the scale of the data in some way.

Percentiles

The P^{th} percentile is a data transformation. It transforms a data value into its relative position in the data set. In fact, we have already discussed the 50^{th} percentile; it is the median. In data sets that do not contain significant quantities of identical data, the 30^{th} percentile is a value such that about 30 percent of the values are below it, and around 70 percent are above it.

P^{th} Percentile

Given a set of data $x_1, x_2, \ldots, x_n$, the **P^{th} percentile** is a value, say x, such that approximately P percent of the data is less than or equal to x and approximately $(100 - P)$ percent of the data is greater than or equal to x.

DEFINITION

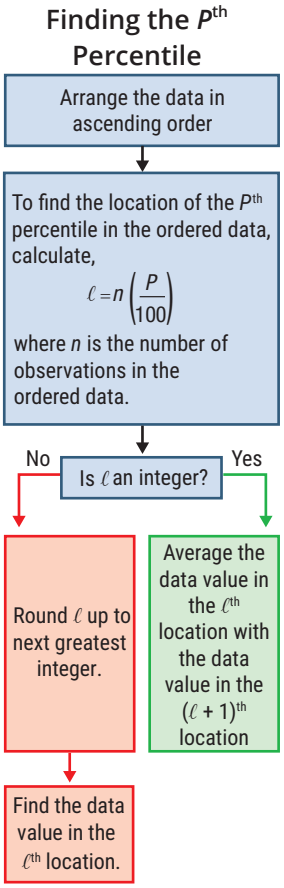


Figure 4.3.1

Example 4.3.4

Determining *z*-Scores for Biology and Psychology Test Scores

Suppose you scored an 86 on your biology test and a 94 on your psychology test. The mean and standard deviations of the two tests are given in the following table.

Test Scores		
Course	**Mean**	**Standard Deviation**
Biology	74	10
Psychology	82	11

What are the z-scores for your two tests? On which of the tests did you perform relatively better?

Solution

The z-score for the biology test is $z = \dfrac{86 - 74}{10} = 1.20$.

The z-score for the psychology test is $z = \dfrac{94 - 82}{11} \approx 1.09$.

On the biology test you scored 1.2 standard deviations above the mean, compared to only 1.09 standard deviations above the mean for the psychology test. Even though the raw score on the psychology test is larger than the raw score on the biology test, relative to the means and variability in the data sets, the performance on the biology test was slightly better. Once again, changing the scale of the data (to standard deviation units) has beneficial effects. It enables the comparison of two measurements that are drawn from different populations.

Properties of a *z*-Score

- If a z-score is negative, the corresponding data value is less than the mean.
- Conversely, if a z-score is positive, the corresponding data value is greater than the mean.
- The z-score is a unit-free measure. That is, regardless of the original units of measurement (centimeters, meters, or kilometers), an observation's z-score will be the same.

PROPERTIES

4.3 Exercises

Basic Concepts

1. What are two methods for describing relative position?
2. If a data value is determined to be the 72nd percentile, what does this mean?
3. Describe how to find the percentile of a particular value.
4. What are quartiles? Are they equivalent to percentiles? If so, how?
5. What is the interquartile range? What does it measure?
6. What are the advantages of using a box plot to display a data set?
7. What are the key calculations needed in order to construct a box plot?

8. What is an outlier? How can outliers be identified?
9. What is a z-score? Why is it useful?

Exercises

10. According to the U.S. Constitution in Article 2, the minimum age to be elected president is 35 years old. To date, the youngest president was Theodore Roosevelt who was elected in 1900 at the age of 42. The oldest president is Joe Biden at the age of 78. The following data shows the age (in years) at inauguration of all United States presidents from 1900 to 2022.

42 43 46 47 51 51 51 52 54 54 55
55 56 56 60 61 62 64 69 70 78

   a. Use the data shown to determine the median age for the presidents elected from 1900 to the present.
   b. Determine the 80th percentile for the data. Interpret this value in the context of the presidents' ages.
   c. When Barack Obama was elected in 2008, he was 47 years old. Determine the percentile for his age. What percentage of most recent presidents were younger than him?
   d. Ronald Reagan was the oldest president to date at his inauguration in 1980. Determine the percentile for his age, 69 years. What percentage of most recent presidents were older than him?

11. The *safety score* variable in the OECD Better Life Index 2022 data set is a measure of how safe a person in this country feels walking alone at night. Using the *safety score* variable, answer the following questions.

Data
stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > OECD Better Life Index 2022

   a. What level of measurement does the data possess?
   b. Determine the 20th percentile.
   c. Determine the 95th percentile.
   d. Interpret the meaning of each of these percentiles.

12. Use the *safety score* variable from the OECD Better Life Index data set from the previous problem to determine the following. Round your answers to the nearest whole number.
   a. Determine the percentile rank for Ireland's safety score.
   b. Determine the percentile rank for Chile's safety score.
   c. Based on the OECD Better Life Index, in which of these two countries would you feel safer to walk alone at night?

13. Using the *Adult.smoking* variable in the US County Data, answer the following questions.

Data
stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > US County Data

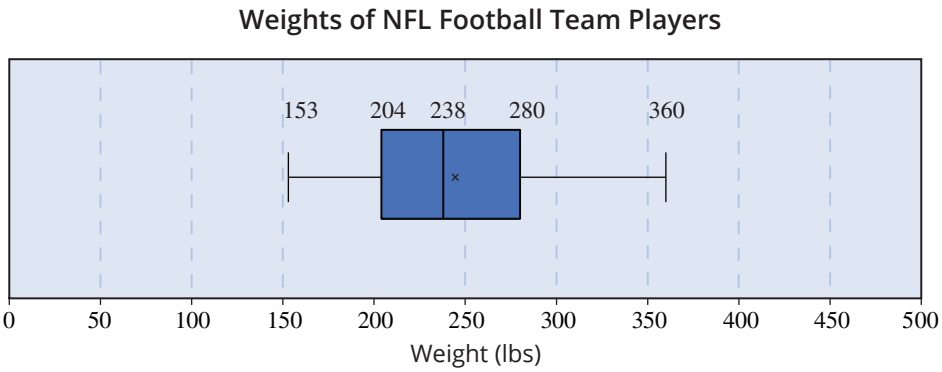
   a. What level of measurement does the data possess?
   b. Determine the 20th percentile. Round your answer to three decimal places.
   c. Determine the 95th percentile. Round your answer to three decimal places.
   d. Interpret the meaning of each of these percentiles.

14. Use the US County Data from the previous problem to find the following.

a. Determine the percentile rank for Lee County, Kentucky's *Adult.smoking* percentage. Round to the nearest whole number.

b. Determine the percentile rank for Ozaukee County, Wisconsin's *Adult.smoking* percentage. Round to the nearest whole number.

15. A team in the National Football League (NFL) has 55 players on the active roster. The 5-number summary of the players' weights is given in the box plot.



a. What is the median weight of the players?

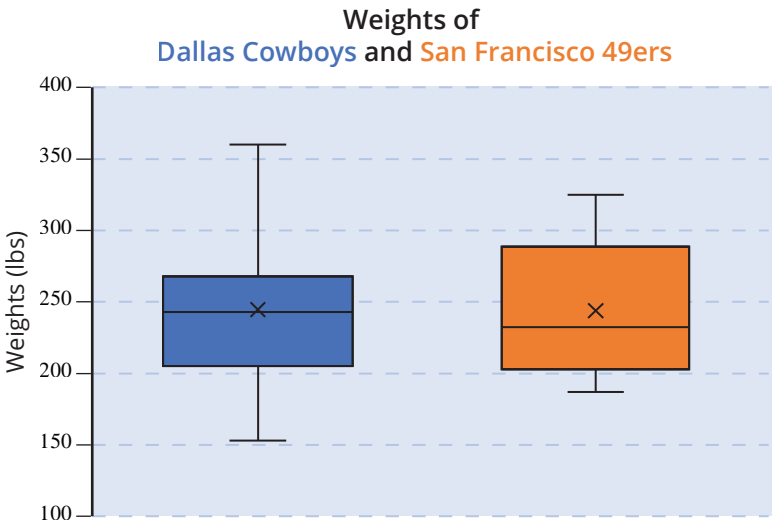
b. What is the interquartile range of the players' weights?

c. Approximately, what is the percentage of players weighing greater than or equal to 204 pounds?

d. How many players weigh between 238 and 280 pounds?

16. The following is a 5-number summary and box plot of the player weights of the Dallas Cowboys and the San Francisco 49ers.[13] Compare the weights of the two teams and note any differences that you observe.

Football Players' Weights (lbs)		
	Cowboys	49ers
Min	153	187
Q1	206	204.25
Median	243	232.5
Q3	268	288.75
Max	360	325
IQR	62	84.5
Mean	244.6032	243.9667
Std Dev.	46.03296	43.30049



17. Subjects in a marketing study were shown a product video and at the end of the video were given a test to measure their recall. The scores are listed below.

26	27	28	31	31	35	38	40	45	49	57
57	58	61	61	61	62	64	68	72	78	81
84	85	86	87	92	95	97	100			

a. What level of measurement does the data possess?

b. Determine Q_1, the first quartile.

c. Determine Q_2, the second quartile.

d. Determine Q_3, the third quartile.

e. Explain the meaning of these percentiles in the context of the marketing study.

f. Determine the interquartile range.

g. Construct a box plot for the test scores. Are there any outliers?

h. Compute the z-score for a test score of 81.

i. Compute the z-score for a test score of 62.

j. Explain what the z-scores in parts **h.** and **i.** are measuring.

18. Use the marketing study data from the previous problem to find the following.

a. Determine the percentile rank for the subject who scored 49.

b. Determine the percentile rank for the subject who scored 95.

19. Using the on-base percentage *(OBP)* variable from the Moneyball data set, answer the following questions.

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Moneyball

a. What level of measurement does the data possess?

b. Determine Q_1, the first quartile. Round your answer to three decimal places.

c. Determine Q_2, the second quartile. Round your answer to three decimal places.

d. Determine Q_3, the third quartile. Round your answer to three decimal places.

e. Explain the meaning of these percentiles in the context of the on-base percentages.

f. Determine the interquartile range.

g. Construct a box plot for the on-base percentages. Are there any outliers?

h. Compute the z-score and percentile for an on-base percentage of 0.280. Round the z-score to three decimal places and the percentile to the nearest whole number.

i. Compute the z-score and percentile for an on-base percentage of 0.355. Round the z-score to three decimal places and the percentile to the nearest whole number.

j. Explain what the z-scores in parts **h.** and **i.** are measuring.

20. Using the on-base percentage (*OBP*) variable from the previous problem, find the following.

a. Determine the percentile rank for the Chicago Cubs in the year 2012. Round to the nearest whole number.

b. Determine the percentile rank for the New York Yankees in the year 2012. Round to the nearest whole number.

21. Consider a set of data in which the sample mean is 64 and the sample standard deviation is 21. For the following specific values, compute the z-score and interpret the results.

a. $x = 80$ **b.** $x = 64$ **c.** $x = 40$

22. A statistics student scored a 78 on the first exam of the semester and an 87 on the second exam of the semester. The average score and standard deviation of scores for the two exams are given in the following table. On which exam did the student perform relatively better?

Test Scores		
Statistic	**First Exam**	**Second Exam**
μ	74	85
σ	10	6

23. A hospital measures babies' lengths when they are born in both inches and centimeters. Eight babies are randomly selected and the following lengths are recorded in both inches and centimeters.

Newborn Lengths								
Baby	1	2	3	4	5	6	7	8
Inches	17.75	18.50	19.25	19.75	20.25	20.50	20.50	20.75
Centimeters	45.09	46.99	48.90	50.17	51.44	52.07	52.07	52.71

a. Determine the mean length in inches and centimeters for the babies.

b. Determine the standard deviation of the lengths of the babies in both inches and centimeters.

c. Prior to calculating the z-score for the length of a particular baby in inches and the z-score for the length of that same baby in centimeters, what does your intuition tell you about the values of the z-scores?

d. Determine the z-score for the length of Baby 3 measured in inches.

e. For Baby 3, determine the z-score for its length measured in centimeters.

f. Consider the z-scores determined in parts **d.** and **e.** Are the z-scores as you expected them to be? Explain.

4.4 Exercises

Basic Concepts

1. Describe the purpose of data subsetting.
2. Describe a data set where data subsetting should be implemented. What are the disadvantages of not subsetting the data?
3. What is Simpson's paradox?

Exercises

4. Suppose you are a craft beer lover taking a trip to Denver on business and you want to be sure to stop at one of the local breweries while you are there. Using the Beers and Breweries data set from the companion website, subset the data to only show beers brewed in Denver, Colorado, and answer the following questions.
   - **a.** What level of measurement do each of the variables represent?
   - **b.** What variables other than *City* could be used to subset the data?
   - **c.** How many craft breweries are in Denver?
   - **d.** Which craft beer has the highest *Alcohol by Volume (ABV)* of the beers brewed in Denver? Give the name of the beer and the brewery.
   - **e.** For the Renegade Brewing Company, how many different IPA styles of beer do they make? What are they?
   - **f.** What is the mean and standard deviation of the *ABV* values for the craft beers made by the Wynkoop Brewing Company?
   - **g.** Compute the coefficient of variation of the *ABV* values for both the Renegade and Wynkoop breweries. Which brewery has more consistent ABV values?

5. Suppose you are looking for a house in Mount Pleasant, SC, which is near Charleston, and you have limited your search to three subdivisions: Park West, Dunes West, and Carolina Park. Using the Mount Pleasant Real Estate Data from the companion website, answer the following questions.
   - **a.** What level of measurement do each of the variables represent?
   - **b.** Which variables could be used to subset the data?
   - **c.** How could you subset the data using quantitative variables such as *List Price* and *Acreage*?
   - **d.** How many different house styles are represented in these three subdivisions? What are the styles?
   - **e.** How many of the houses were built in the years 2015 to 2017? Which subdivision has the most homes built during this timeframe?
   - **f.** What is the average price of homes built in the years 2015 to 2017 in Carolina Park? Round your answer to the nearest whole dollar.
   - **g.** For all homes built in the years 2015 to 2017 in the three subdivisions, what is the minimum and maximum priced homes and in which subdivision are they?
   - **h.** What is the price per square foot of the two homes in part **g.**?
   - **i.** What variables do you think may contribute to the high price of the house with the maximum price?

Technology

To learn how to perform calculations on a filtered data set, please visit stat.hawkeslearning.com and navigate to **Discovering Statistics and Data, Fourth Edition > Technology Instructions > Data Manipulation > Subset Calculations.**

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Beers and Breweries

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Mount Pleasant Real Estate Data

4.5 Exercises

Basic Concepts

1. When analyzing grouped data, are the measurements exact? Why or why not?
2. What calculations are required in order to analyze grouped data?

Exercises

3. A client of a commercial rose grower has been keeping records on the shelf-life of a rose. The client sent the frequency distribution to the grower. Determine the sample mean and sample variance for the shelf-life given the following frequency distribution.

Rose Shelf-Life	
Days of Shelf-Life	**Frequency**
1 – 6	2
7 – 12	3
13 – 18	9
19 – 24	6
25 – 30	3
31 – 36	1

4. A frequency distribution of Alcohol by Volume (ABV) for the Beers and Breweries data set from the companion website is shown below. Use the frequency distribution to perform the following.

Data

Discovering Statistics and Data, Fourth Edition > Data Sets > Beers and Breweries

ABV Frequencies	
ABV	**Frequency**
0.0010 – 0.017	1
0.0175 – 0.033	6
0.0335 – 0.049	402
0.0495 – 0.065	1228
0.0655 – 0.081	565
0.0815 – 0.097	146
0.0975 – 0.113	45
0.1135 – 0.129	3

   a. Compute the population mean ABV of all beers based on the frequency distribution. Round your answer to three decimal places.

   b. Compute the population variance of the ABVs of the different beers based on the frequency distribution. Round your answer to four decimal places.

   c. Compute the population standard deviation of the ABVs of the different beers based on the frequency distribution. Round your answer to three decimal places.

5. An article in Business Week discussed the large spread between the federal funds rate and the average credit card rate. The table below is a frequency distribution of the credit card rate charged by the top 100 issuers. Note that at the time these figures were published, the average federal funds rate was well below 5%.

Credit Card Rates

Credit Card Rate	Frequency
19% – 24%	36
18% – 18.9%	8
17% – 17.9%	15
16% – 16.9%	12
15% – 15.9%	29

a. Determine the sample mean of the credit card rate charged by the top 100 issuers based on the frequency distribution.

b. Determine the sample variance of the credit card rate charged by the top 100 issuers based on the frequency distribution.

c. Determine the sample standard deviation of the credit card rate charged by the top 100 issuers based on the frequency distribution.

6. The frequency distribution for the response time (in minutes) for EMTs after a 911 call is shown below. Calculate the population mean and variance for the response time.

Response Time for EMTs

Time (in minutes)	Frequency
7.0 – 7.9	16
8.0 – 8.9	25
9.0 – 9.9	38
10.0 – 10.9	43
11.0 – 11.9	33
12.0 – 12.9	19

Example 4.6.2

Determining a Batting Average

Suppose you have been playing softball and have kept records on each plate appearance. According to your records you have batted 216 times. Of those 216 plate appearances, you have walked 24 times, gotten out on a sacrifice hit 7 times, and reached base on a hit 64 times. Let's compute your batting average, which is a proportion. The batting average is the proportion of times you reached base on a hit, excluding walks and sacrifice hits. In this case the number in the group of at bats we will consider is

$$\begin{aligned} N &= \text{Plate appearances} - \text{Walks} - \text{Sacrifice Hits} \\ &= 216 - 24 - 7 \\ &= 185 \text{ at bats.} \end{aligned}$$

The proportion of times you got a hit (excluding walks and sacrifice hits) is

$$p = \frac{64}{185} \approx 0.346.$$

Hence your batting average is 0.346.

Sabermetrics

In baseball and softball, statistics are plentiful. Using these metrics to improve performance, plan a game strategy, or evaluate a player is part of the field of sabermetrics. Some other proportions which are related to Batting Average include OBP and SLG. They give more information about the offensive output of a player.

On Base Percentage, OBP, uses the number of times a batter reaches base, including when walked (BB) per plate appearance.

Slugging Percentage, SLG, measures the offensive productivity of a batter by calculating a weighted average for the number of bases achieved per hit, i.e., a double counts twice as much as a single, etc. A player who generates many extra-base hits would have a higher SLG value.

The most commonly used metric to gauge a player's offensive ability is the statistic OPS, On-base Plus Slugging, which is the sum of OBP and SLG.

Source: www.mlb.com/glossary/standard-stats

For the softball example, you would not convert 0.346 to a percentage because batting averages are always reported as a proportion. In Major League Baseball (MLB) statistics, the zero in front of the decimal point is usually omitted in batting averages. So, if you were in MLB your batting average would be reported as .346.

This chapter has been devoted to summarizing data. Yet with the exception of the mode, none of the summary methods discussed should be applied to nominal data. Using proportions is one of the few summary methods available for analyzing qualitative data.

4.6 Exercises

Basic Concepts

1. What is a proportion?
2. What is the difference in notation between a population and a sample proportion?
3. Other than the mode, proportions are one of the few summary methods available to analyze what type of data?

Exercises

4. *Science News*, Vol. 143 reported some "depressing news for low-cholesterol men."[14] A study conducted at the University of California, San Diego found that among men age 70 and older in the low cholesterol group (concentration of less than 160 mg of cholesterol per deciliter of blood), nine of 75 reported symptoms of mild depression. Calculate the sample proportion of men age 70 and older in the low cholesterol group who reported symptoms of mild depression.

5. A study conducted at Virginia Commonwealth University in Richmond indicates that many older individuals can shed insomnia through psychological training.[15] A total of 23 insomnia sufferers averaging age 67 years old completed eight weekly sessions of cognitive-behavior therapy. After the therapy, 13 participants enjoyed a substantially better night's sleep. Compute the sample proportion of insomnia sufferers who enjoyed a better night's sleep after the therapy.

6. A study was conducted to explore the relationship between smoking and depression. Researchers interviewed 995 smokers and asked them if they had ever experienced severe depression. Of those surveyed, 250 said that they had experienced severe depression. Compute the sample proportion of smokers who experienced severe depression.

7. Researchers conducted a study of the relationship between baldness and heart disease.[16] Of 600 men age 21–54 who had just suffered their first heart attack, 50 were found to have vertex scalp baldness (hair loss from the top of the head). Compute the sample proportion of men age 21–54 who had just suffered their first heart attack and also experienced vertex scalp balding.

8. Using the Beers and Breweries data set from the companion website, consider the following questions. Round your answers to three decimal places.

Data

Discovering Statistics and Data, Fourth Edition > Data Sets > Beers and Breweries

a. What proportion of beers are brewed in Colorado?

b. What proportion of beers are brewed in California?

c. What is the overall proportion of beers brewed in Colorado or Texas?

9. According to a study administered by the National Bureau of Economic Research, half of Americans would struggle to come up with $2000 in the event of a financial emergency.[17] The majority of the 1900 Americans surveyed said they would rely on more than one method to come up with emergency funds if required. In the survey, 532 people said that they "certainly" would not be able to cope with an unexpected $2000 bill if they had to come up with the money in 30 days, and 418 people said they "probably" would not be able to cope.

a. What percentage of Americans "certainly" would not be able to produce $2000 in the event of an emergency according to the study?

b. What percentage of Americans would "probably" not be able to pay a $2000 bill in 30 days if required?

c. What does this say about the savings habits of Americans?

10. What college football conference has the right to brag about putting players in the NFL? The following table displays the conference affiliations of the first round draft picks for the NFL from 2012 to 2022. Use the data to answer the following questions. Round answers to three decimal places where appropriate.

First Round NFL Draft Picks by Conference 2012–2022	
Football Conference	**Number of Players**
SEC	117
Big Ten	55
Pac-12	48
ACC	57
Big 12	25
AAC	15
Independent	12
Other	22

a. What proportion of first round draft picks are from the SEC?

b. What proportion of first round draft picks are from the "Other" conferences category?

c. Is it true that a player in the SEC has a better chance of being drafted in the NFL than a player from a different conference? Explain.

11. It is no secret that Wall Street firms compete aggressively to attract clients to their firms. Having more clients translates into fees and revenues that turn into profits. A survey of 125 clients were asked what attracted them to their respective Wall Street firm. The following table shows the results.

Client Response	
Perk Received	**Client Response**
Lucrative Golf Outings	12
Lavish Dinners	33
Free Private Jet Use	8
Prime Seats at Sports Events	20
Other	22
No Perk Received	30

a. Which type of perk appears to be most successful in attracting clients?

b. What proportion of clients were attracted to a Wall Street firm by the perk identified in part **a.**?

c. What proportion of clients did not receive a perk at all?

d. Given that these perks aren't inexpensive, what conclusion can you make about providing perks to clients? Explain.

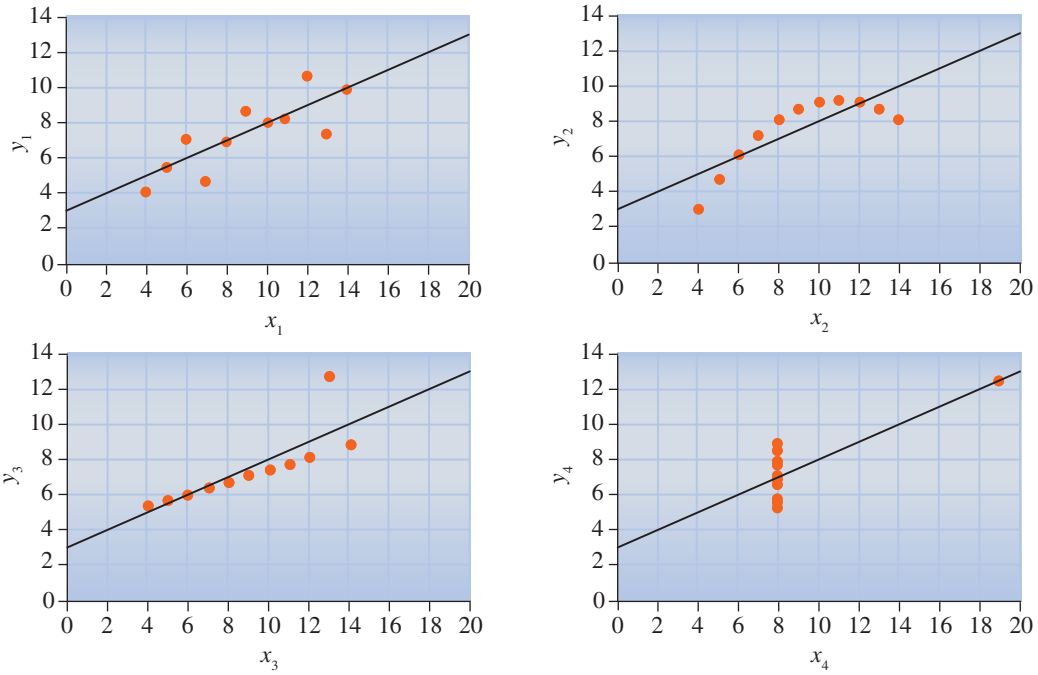


Figure 5.1.21

The Anscombe Quartet brings out an important data analytic principle. No matter what kind of numeric data is being analyzed, it is critically important to plot it before making any conclusions. For single variables, an analyst should at least be looking at histograms and box plots. When we begin to look for relationships in bivariate or multivariate data, creating scatterplots to display the relationship between the variables is more important than calculating summary measures of linear relationship, i.e., the correlation coefficient.

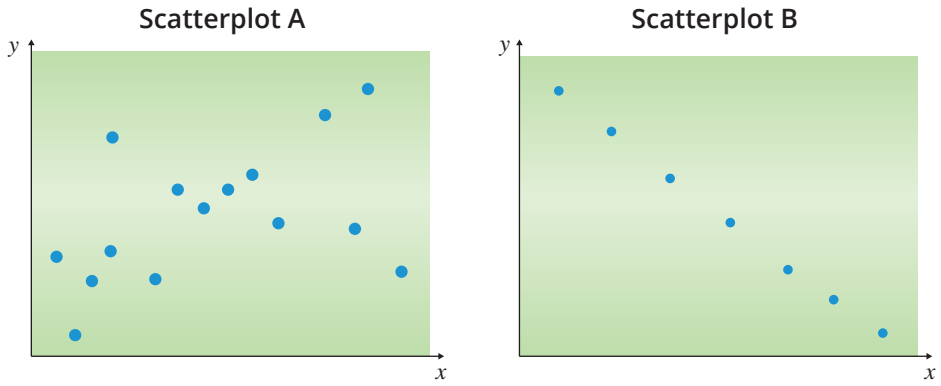
5.1 Exercises

Basic Concepts

1. Give an example of a situation in which knowledge of a relationship between two variables is desired.
2. If a relationship can be uncovered, what are the potential benefits?
3. What is bivariate data? How is bivariate data different from univariate data?
4. What graphical tool is often used in the discovery of relationships?
5. What sort of questions should you ask when studying a graphical representation of bivariate data?
6. If bivariate data exhibits an inverse relationship, what does that mean?
7. How do you measure the exact relationship between two variables?
8. In what range is the value of r when bivariate data exhibits a positive relationship? A negative relationship?
9. If the value of r is small, does this always mean that no relationship exists? Explain.
10. What is confounding? Why is confounding a problem?

Exercises

11. Answer the following questions regarding the overall pattern of the data for each of the scatterplots.



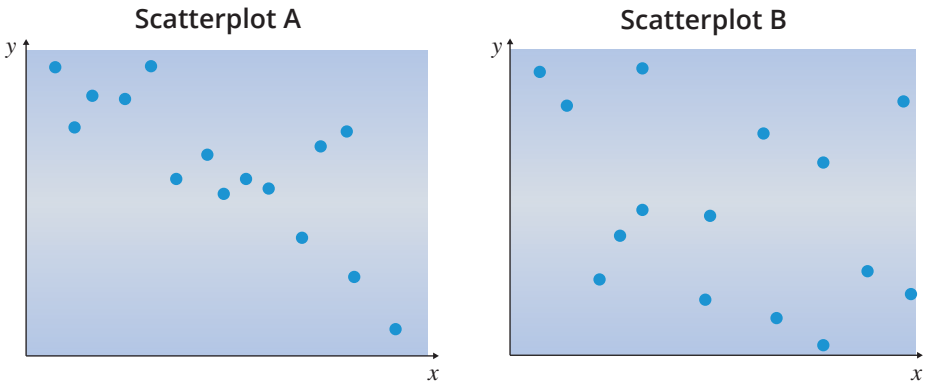
a. Does the pattern roughly follow a linear pattern?

b. Is the pattern upward sloping or downward sloping?

c. Are the data values tightly clustered in the pattern or widely dispersed?

d. Are there significant deviations from the pattern?

12. Answer the following questions regarding the overall pattern of the data for each of the scatterplots.



a. Does the pattern roughly follow a linear pattern?

b. Is the pattern upward sloping or downward sloping?

c. Are the data values tightly clustered in the pattern or widely dispersed?

d. Are there significant deviations from the pattern?

13. In the story of Moneyball, Billy Beane and Paul DePodesta initially calculated that 95 wins in a season was necessary for the Oakland A's to make it into the playoffs.[7] Next, they wanted to understand the relationship between overall season run differential (the number of runs scored in a season minus the number of runs allowed) and the number of games won in a season. Go to the companion website and download the Moneyball data set. Beane and DePodesta performed this analysis in 2002, so they only had data up to the year 2001. Subset the data set to only include data for the years 1962–2001.

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Moneyball

a. Analyze the data collected for the study by answering the following questions:

i. What questions might Beane and DePodesta be trying to answer?

ii. Do the variables selected in the data set seem appropriate for answering the question?

iii. What biases or errors might be present in the data?

iv. What level of measurement (nominal, ordinal, interval, ratio) does each variable possess?

v. How is the data collected – through observation or controlled experiment?

b. Plot the data points for the variables *RD* (run differential) and *W* (wins) on a scatterplot.

c. Based on the scatterplot in part **b.**, answer the following questions regarding the overall pattern of the data.

i. Does the data roughly follow a linear pattern?

ii. Is the pattern upward sloping or downward sloping?

iii. Are the data values tightly clustered in the pattern or widely dispersed?

iv. Are there significant deviations from the pattern?

14. A pharmacist is interested in studying the relationship between the amount of a particular drug in the bloodstream (in nanograms per milliliter ng/ml) and reaction time (in seconds) of subjects taking the drug. Ten subjects are randomly selected and administered various doses of the drug. The reaction times (in seconds) are measured 15 minutes after the drug is administered with the following results.

Reaction Time of a Drug										
Amount of Drug (ng/ml)	1	2	3	4	5	6	7	8	9	10
Reaction Time (in sec)	0.5	0.7	0.6	0.7	0.8	0.8	0.9	0.6	0.9	1.0

a. Analyze the data collected for the study by answering the following questions.

i. Do the variables selected for measurement seem appropriate for the study of interest?

ii. What biases or errors might be present in the data? What confounding variables could impact the conclusion?

iii. What level of measurement (nominal, ordinal, interval, ratio) does the data possess?

iv. How is the data collected—through observation or controlled experiment?

b. Plot the data points on a scatterplot.

c. Based on the scatterplot in part **b.**, answer the following questions regarding the overall pattern of the data.

i. Does the pattern roughly follow a linear pattern?

ii. Is the pattern upward sloping or downward sloping?

iii. Are the data values tightly clustered in the pattern or widely dispersed?

iv. Are there significant deviations from the pattern?

15. Illustrate, using a scatterplot, a data set that would have a correlation coefficient of 1.

16. Illustrate, using a scatterplot, a data set that would have a correlation coefficient of -1.

17. Describe the relationships indicated by the correlation coefficients as tightly clustered in a positive linear fashion, tightly clustered in a negative linear fashion, loosely clustered in a positive linear fashion, loosely clustered in a negative linear fashion, or no linear relationship.

a. $r = 0.9$

b. $r = 0.5$

c. $r = -0.9$

d. $r = -0.5$

e. $r = 0$

f. What assumption did you make about the scatterplots in answering **a.** through **e.**?

18. Describe the relationships indicated by the correlation coefficients as tightly clustered in a positive linear fashion, tightly clustered in a negative linear fashion, loosely clustered in a positive linear fashion, loosely clustered in a negative linear fashion, or no linear relationship.

a. $r = 0.8$

b. $r = 0.4$

c. $r = -0.8$

d. $r = -0.4$

e. $r = 0.1$

f. What assumption did you make about the scatterplots in answering **a.** through **e**.?

19. Sometimes the following descriptions are assigned to the correlation coefficient:

$r = 0$	no linear relationship
$-0.5 < r < 0$	weak negative linear relationship
$0 < r < 0.5$	weak positive linear relationship
$-0.8 < r \leq -0.5$	moderate negative linear relationship
$0.5 \leq r < 0.8$	moderate positive linear relationship
$-1.0 < r \leq -0.8$	strong negative linear relationship
$0.8 \leq r < 1.0$	strong positive linear relationship
$r = 1$	exact positive linear relationship
$r = -1$	exact negative linear relationship

Describe the relationships indicated by the correlation coefficients below using the descriptions defined above.

a. $r = 0.9$

b. $r = 0.5$

c. $r = -0.9$

d. $r = -0.5$

e. $r = 0$

f. What assumption did you make about the scatterplots in answering **a.** through **e.**?

20. Describe the relationships indicated by the correlation coefficients below using the descriptions defined in problem 19 above.

a. $r = 0.8$

b. $r = 0.4$

c. $r = -0.8$

d. $r = -0.4$

e. $r = 0.1$

f. What assumption did you make about the scatterplots in answering **a.** through **e.**?

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Super Bowl

21. Using the Super Bowl data set from the companion website, consider the following:

a. Construct a scatterplot using the variables *Winner_First Downs* and *Winner_Total Yards*.

b. Does there appear to be a negative or positive relationship between the variables?

c. Compute the correlation coefficient.

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > OECD Better Life Index 2022

22. Using the OECD Better Life Index 2022 data set from the companion website, consider the following:

a. Construct a scatterplot using the variables *Safety Score* and *Life Satisfaction*.

b. Determine the correlation coefficient.

c. Describe the relationship indicated by the correlation coefficient and the scatterplot.

23. If the following variables have high negative linear correlations, is it reasonable to conclude that an increase in one variable causes a decrease in the other variable? Explain what could be causing this apparent relationship.

a. Time spent studying and number of followers on Instagram

b. Amount of caffeine consumed and academic performance

c. Amount of spending money and time to spend with friends

24. If the following variables have high positive linear correlations, can we conclude that an increase in one variable causes an increase in the other variable? Explain what could be causing this apparent relationship.

a. Sale of air conditioners and sale of tomatoes

b. Sale of greeting cards and sale of chocolates

c. Number of wrecks on a local highway and absenteeism from work

Continued

Then, they set about recruiting the players within their budget that helped them achieve their goal. The A's are still considered one of the most efficient teams when it comes to recruiting. The Oakland A's went to the playoffs in 2000–2003, 2006, 2012–2014, and 2018–2020. The improvement of the Oakland A's performance is a case where empiricism triumphed over belief, conjecture, and speculation.

Interpretation of b_0 and b_1

The intercept coefficient, b_0, is the average value of the dependent variable, y, when the independent variable, x, is equal to zero.

The slope coefficient, b_1, is the average change in the dependent variable, y, for a one unit change in the independent variable, x.

DEFINITION

5.2 Exercises

Basic Concepts

1. What is the value of a model?
2. What is regression analysis?
3. What is the difference between a dependent and an independent variable?
4. What is $\hat{y}$? How does this differ from y?
5. What is the technique used to estimate the linear regression coefficients?
6. What is the relationship between scatterplots and linear regression?
7. Why is it often difficult to accurately describe real world situations using a linear regression equation?
8. What is the sum of squared errors and what does it measure?
9. Explain why the best line is referred to as the least squares line.
10. What measure should be minimized in order to find the least squares line?
11. What is the equation for finding the slope of the least squares line?
12. What is the equation for finding the intercept of the least squares line?
13. Interpret the intercept coefficient, b_0.
14. Interpret the slope coefficient, b_1.

Exercises

15. Suppose that a company wishes to predict sales volume based on the amount of advertising expenditures. The sales manager thinks that sales volume and advertising expenditures are modeled according to the following linear equation. Both sales volume and advertising expenditures are in thousands of dollars.

$$\textit{Estimated Sales Volume} = \hat{y} = 49.25 + 0.51\ \textit{Advertising Expenditures}$$

   a. What is the dependent variable in this model? Explain.
   b. What is the independent variable in this model? Explain.
   c. What is the estimated sales volume for this company when the marketing department spends $40,000 on advertising?
   d. If the company had a target sales volume of $100,000, how much should the sales manager allocate for advertising in the budget?

e. What is the sales manager forgetting to account for when using this linear equation to determine sales volume? What kinds of problems could this cause for the company?

16. Suppose the following estimated regression equation was determined to predict salary based on years of experience.

$$\textit{Estimated Salary} = \hat{y} = 25689.10 + 2148.35\ \textit{Years of Experience}$$

a. What is the dependent variable? What is the independent variable?

b. What is the value that estimates b_0 in this particular equation?

c. What is the value that estimates b_1 in this particular equation?

d. What is the estimated salary for an employee with 15 years of experience?

17. Plot the following lines.

a. $y = 2 + 3x$

b. $y = 4 + 8x$

c. $y = 9 - 2x$

d. $y = x$

18. Plot the following lines.

a. $y = 100 + 50x$

b. $y = 0.5 + 0.7x$

c. $y = 20 - 5x$

19. Consider the following estimated regression equation.

$$\hat{y} = 10x - 5$$

a. Complete the following table.

Predicted Values					
x	2	5	7	9	10
y					

b. Do these two variables appear to have a positive or negative relationship?

c. For these two variables, what sign would you expect the correlation coefficient to have? Explain.

20. Consider the following data.

Observed Values					
x	0	1	5	6	8
y	2	4	9	7	8

a. Draw a scatterplot of the data.

b. Draw a line which you believe fits the data.

c. Suppose that $\hat{y} = 3 + 0.8x$ is a line that fits the data reasonably well. Complete the following table.

Observed and Predicted Values				
Observed *x*	**Observed *y***	**Predicted *y***	**Error**	**Squared Error**
0	2			
1	4			
5	9			
6	7			
8	8			

d. What is the sum of squared errors for this data?

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Mount Pleasant Real Estate Data

21. Using the Mount Pleasant Real Estate data set from the companion website, consider the following:

a. Suppose we want to predict *List Price* based on *Square Footage*. Write the estimated regression equation in terms of *List Price* and *Square Footage*. (Assume the parameters of this model have not been estimated.)

b. Create a scatterplot using the *List Price* and *Square Footage* variables and draw a least squares regression line.

c. Suppose we determine that an equation that fits the data reasonably well is $\textit{Estimated List Price} = \hat{y} = -144{,}193 + 256.3753\ \textit{Square Footage}$. Using this equation, complete the following table. Round values to the nearest whole number.

Housing Prices and Square Footage				
Observed Selling Price	**Observed Square Footage**	**Predicted Selling Price (Thousands of Dollars)**	**Error**	**Squared Error**
$375,000	1797			
$423,600	2135			
$448,315	1895			
$515,250	2423			
$556,400	2800			
$600,500	3045			
$583,620	3115			
$683,025	3210			
$635,250	3143			
$615,300	2730			
$731,410	3340			
$860,750	3521			
$835,000	4236			
$815,500	3841			

d. Compute the sum of squared errors for the data in the table in part **c.** Round your answer to the nearest whole number.

e. Is there a variable other than *Square Footage* that might predict *List Price* more accurately? Create a scatterplot using that variable and *List Price*, and compare it to the plot in part **b.** Is the grouping tighter or more dispersed?

22. In the story of Moneyball, Billy Beane and Paul DePodesta determined that 95 wins out of 162 games was necessary in order for the Oakland A's to make it to the MLB playoffs. They also determined that there was a strong linear relationship between season run differential and the number of wins in a season. Using the Moneyball data set from the companion website, subset the data to only include data prior to the year 2002 (1962-2001), and answer the following questions.

Data

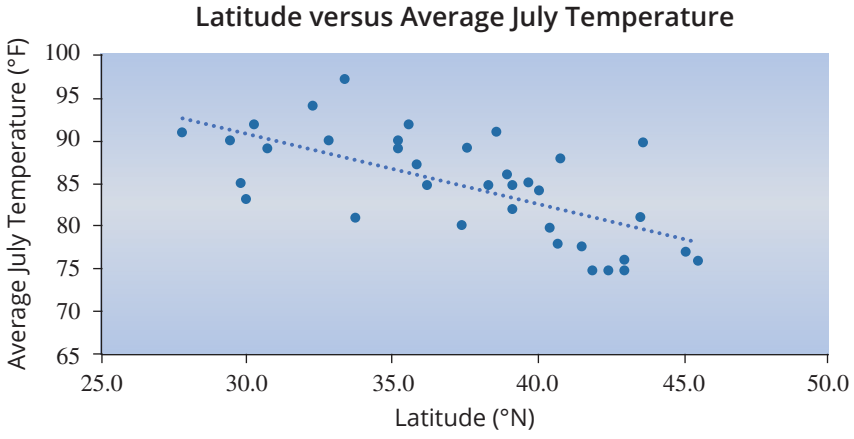
stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Moneyball

a. What is the regression equation for predicting wins (W) using run differential (RD)?

b. What is the run differential necessary in order to win 95 games? Use the regression equation from part **a.** to calculate your answer. Do not round any values until the final answer. Round your final answer to the nearest whole number.

23. The summary statistics for the latitudes and average July temperature of 36 cities in the United States are shown in the table. The sample data is also displayed in a scatterplot with the regression line included.

	Mean	Standard Deviation
Latitude (°N)	37.4	4.89
Average July Temperature (°F)	85	6.04
Correlation	−0.66	



a. What is indicated by the negative value of the correlation coefficient?

b. Use the summary statistics to determine the equation of the regression line that predicts the average July temperature when given the latitude of a certain U. S. city.

c. St Louis, Missouri is located at 38.6°N latitude. What is the average July temperature that is predicted by the regression model?

24. Consider the following data.

x	−2	−1	0	3	5
y	1	3	5	4	8

a. Plot the data points on a scatterplot.

b. Determine the least squares line. Use x as the independent variable.

c. Plot the least squares line on the scatterplot.

d. Use the model to compute the error for each data point.

25. Suppose a linear regression analysis produced the following equation relating an individual's salary to the current value of his or her home.

$$\textit{Estimated Current Value of Home} = \hat{y} = 72{,}331 + 3.14\ \textit{Annual Salary}$$

a. Which of the variables in the model is the dependent variable?

b. Which of the variables in the model is the independent variable?

c. What would be the predicted current value of home for someone earning a salary of $52,000?

d. If a person earned $5000 additional income, how much of an increase in home value would be predicted?

e. In terms of the problem, interpret the estimate of the slope in the model.

f. In terms of the problem, interpret the estimate of the intercept in the model.

g. Do you believe annual salary is a causal factor in explaining the price of someone's home? Explain.

26. Suppose a linear regression analysis produced the following equation relating a basketball player's total points scored to the number of minutes played in a season.

$$\textit{Estimated Points Scored} = \hat{y} = -97.2 + 0.645\ \textit{Minutes Played}$$

a. Which of the variables in the model is the dependent variable?

b. Which of the variables in the model is the independent variable?

c. What would be the predicted value of total points scored for a basketball player who plays 500 minutes in a season?

d. If a basketball player played an additional 100 minutes, how much of an increase in total points scored would be predicted?

e. In the model, which of the coefficients is the slope?

f. In the model, which of the coefficients is the intercept?

g. Do you believe the number of minutes played is a causal factor in explaining the total points scored? Explain.

27. Suppose you were studying the educational level of husbands and wives (measured in number of years of education). You have randomly selected 10 couples and have obtained the data in the following table.

Education Level										
Husband	12	16	16	18	20	17	23	14	12	16
Wife	14	16	14	16	16	18	18	12	16	20

a. Suppose you wanted to predict the husband's years of education based on the wife's. Use the data to estimate the appropriate model.

b. Use the model in part **a.** to predict the husband's educational level if married to a woman with 16 years of education.

c. Suppose you wanted to predict the years of education for the wife based on the husband's years of education. Use the data to create the appropriate model. Did you get the same model as in part **a.**?

d. Use the model created in part **c.** to predict the wife's educational level if married to a husband with 16 years of education.

e. Do you believe there is a causal relationship between the two variables? If so, which direction is the causality? Does the husband's education cause the wife to have more or less education, or vice versa?

28. There were ten homes sold in the Seaside Subdivision in April 2021. The data shown below shows the square footage and sale price for each of the houses. A table of summary statistics for each variable and the correlation coefficient is shown as well.

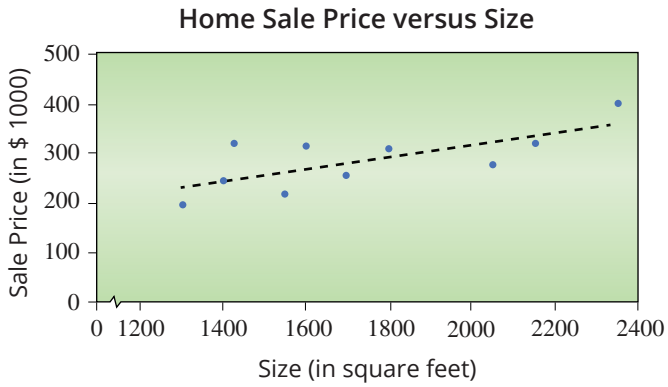
a. Determine the equation of the least squares regression line.

b. Interpret the value of the slope coefficient in the context of the data.

c. Calculate the error for the house with 1550 square feet.

Size (sq. ft.)	1400	1600	2050	1800	1300	1550	2350	2150	1425	1700
Sale Price (in $1000)	245	312	279	308	199	219	405	324	319	255

Variable	Mean	Standard Deviation
Explanatory - Size	1732.5	350.4
Response - Price	286.5	60.2
Correlation Coefficient	0.735	



5.3 Evaluating the Fit of a Linear Model

The Importance of Errors

The usefulness of the regression model depends on the magnitude of the prediction errors you expect the model to produce. The Jeep Cherokee Limited model is

$$Asking\ Price = \hat{y} = b_0 + b_1\ Age + \text{error}.$$

Yet we ignored the error component in the previous section when we predicted the prices of the Jeep Cherokee Limited for different ages (in Table 5.2.3). For instance, when we predicted the price of a two-year-old Jeep Cherokee Limited, we found

$$Estimated\ Asking\ Price = \hat{y} = \$47{,}030.83 - \$3{,}846.09 \cdot (2) = \$39{,}338.65.$$

5.3 Exercises

Basic Concepts

1. Why is the magnitude of the prediction errors important when estimating a regression model?
2. What is the mean error for a least squares model?
3. Describe what the magnitude of the variation in the error terms tells us about the reliability of the regression model.
4. What is the variance of the error terms?
5. How many degrees of freedom are associated with the error term in a simple linear regression model?
6. What is the square root of the variance of the error term known as?
7. Describe where the summary statistics for the standard error and mean square error are found in a standard regression summary output in Microsoft Excel.
8. Is there a universal rule on how large is *large* with regard to standard error in a model?
9. What is estimated by the variance of the error term and what is estimated by the standard error?
10. What is the total sum of squares?
11. How are the total sum of squares and the sample variance related?
12. Define error in terms of a regression model.
13. What part of the simple linear regression model captures the unexplained variation?
14. Describe the total sum of squares in terms of explained and unexplained variation.
15. What is the sum of squares of regression?
16. Express SSR in terms of the total sum of squares and the sum of squared errors. Interpret this in terms of model variation.
17. Why will there be errors in virtually all regression models?
18. What is the coefficient of determination? What kinds of values can the coefficient of determination take?
19. Suppose that regression analysis is performed and the resulting model has an R^2 value of 0.856. Interpret this value.
20. How is the coefficient of determination related to the correlation coefficient?

Exercises

21. Consider the following summary output.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.911653228
R Square	0.831111609
Adjusted R Square	0.79733393
Standard Error	0.253142413
Observations	7

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	1.576737452	1.576737	24.60535
Residual	5	0.320405405	0.064081	
Total	6	1.897142857		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	4.021621622	0.181401491	22.16973	3.47E-06
X Variable 1	-0.22297297	0.044950802	-4.96038	0.004247

a. What is the variance of the error for the data?

b. What is the standard error of the model?

Data

stat.hawkeslearning.com

Discovering Statistics and Data, Fourth Edition > Data Sets > US County Data

22. Using the US County Data from the companion website, use the variables *Diabetes.percent* and *Adult.obesity.percent* to perform the following.

a. Calculate the regression equation to predict *Diabetes.percent* using *Adult.obesity.percent*. Round values to 5 decimal places.

b. What are b_0 and b_1? Round your answers to 5 decimal places.

c. Using the information from parts **a.** and **b.**, complete the following table. Round Predicted *Diabetes.percent* and Error to 3 decimal places, and round Squared Error to 5 decimal places.

Observed versus Predicted Values				
Observed *Adult.obesity.percent*	**Observed *Diabetes.percent***	**Predicted *Diabetes.percent***	**Error**	**Squared Error**
0.408	0.173			
0.275	0.084			
0.375	0.115			
0.349	0.156			
0.312	0.070			
0.382	0.210			

d. Compute the sum of squared errors for the table in part **c.** Round your answer to 5 decimal places.

e. Compute the variance of the error term for the table in part **c.** Round your answer to 5 decimal places.

f. Compute the standard error of the table in part **c.** Round your answer to 5 decimal places.

g. Do you believe the estimates of b_0 and b_1 provide a reliable estimated regression equation for the data? Explain.

23. In the previous section, we used the Moneyball data set (1962-2001) to determine that a season run differential of about 135 runs was necessary for the Oakland A's to make it to the MLB playoffs. However, Coach Billy Bean and statistician Paul DePodesta needed to figure out how to make that run differential a reality. They found that two of the most statistically significant variables that contributed to the number of runs scored were on-base percentage and slugging percentage. Use the Moneyball data set from the companion website, subsetted to only include the years 1962-2001 since that was the only data available to Beane at the time, and perform the following.

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Moneyball

a. Calculate the regression equation to predict runs scored (*RS*) using on-base percentage (*OBP*). Round values to 5 decimal places.

b. Calculate the regression equation to predict runs scored (*RS*) using slugging percentage (*SLG*). Round values to 5 decimal places.

c. Calculate the regression equation to predict runs allowed (*RA*) using opponent on-base percentage (*OOBP*). *OOBP* is only measured from 1999 on, so use data for the years 1999-2001 to estimate the equation. Round values to 5 decimal places.

d. Calculate the regression equation to predict runs allowed (*RA*) using opponent slugging percentage (*OSLG*). *OSLG* is only measured from 1999 on, so use data for the years 1999-2001 to estimate the equation. Round values to 5 decimal places.

e. Using the regression equation from part **a.**, complete the following table. Round values to the nearest whole number.

Runs Scored using On-Base Percentage

Observed *RS*	Observed *OBP*	Predicted *RS*	Error	Squared Error
687	0.319			
897	0.350			
724	0.320			
923	0.354			
642	0.323			

f. Calculate the sum of squared errors and the standard error for the table in part **e.** Round answers to 3 decimal places.

g. Using the regression equation from part **d.**, complete the following table. Round your answer to the nearest whole number.

Predicted Runs Allowed using Opponent Slugging Percentage				
Observed *RA*	Observed *OSLG*	Predicted *RA*	Error	Squared Error
713	0.398			
806	0.440			
627	0.378			
968	0.494			
766	0.437			

h. Calculate the sum of squared errors and the standard error for the table in part **g.** Round answers to 3 decimal places.

24. A digital marketing company has been experimenting with the effect of price on sales. Five different product prices have been sent to different sets of customers. They have carefully tracked the customers from each group and have recorded the proportion from each price category that purchased the product. The results are given in the following table.

Product Price Marketing Experiment Results					
Proportion That Purchased Product	0.032	0.028	0.026	0.015	0.009
Price of Product ($)	29.95	34.95	39.95	44.95	49.95

a. What level of measurement do the two variables in the table possess?

b. Specify the model that the marketing manager would be interested in estimating.

c. Which of the variables is the dependent variable in the model?

d. Which of the variables is the independent variable in the model?

e. Draw a scatterplot of the data.

f. Use the data in the table to estimate the model.

g. Predict the proportion that will buy the product if the price is $35.00.

h. Compute the mean error for the model you estimated in part **f.**

i. Determine the variance of the error term.

j. What is the coefficient of determination? Interpret this value in terms of the problem.

25. An economist is studying the relationship between income and savings. He has randomly selected seven subjects and obtained income and savings data from them. He wishes to use a simple linear regression model to predict savings based on annual income.

Income and Savings							
Income (Thousands of Dollars)	28	25	34	43	48	39	74
Savings (Thousands of Dollars)	0.2	0	0.8	1.2	3.1	2.1	8.3

a. What level of measurement do the two variables in the table possess?

b. Which of the variables is the dependent variable in the model?

c. Which of the variables is the independent variable in the model?

d. Draw a scatterplot of the data. Does the scatterplot suggest that a linear model is appropriate? Explain.

e. Use the data to estimate the appropriate model.

f. Predict the savings for someone who earns fifty thousand dollars annually.

g. Interpret the meaning of the slope coefficient in the problem.

h. What fraction of the variation in savings is explained by income?

26. Since 2009, the average term for a new-car loan was nearly 64 months. This leaves the buyer vulnerable to owing more on the car than it is worth. When applying for an automobile loan, it is oftentimes recommended to sign up for the shortest term you can afford. It is believed that along with one's credit rating, the length of the loan will help the buyer get a favorable interest rate. The following table contains interest rates and lengths of loans for 20 randomly selected auto purchases. Using the data in the table, answer the following questions.

Lengths of Loans and Interest Rates

Months Financed	Interest Rate (%)	Months Financed	Interest Rate (%)
12	4.00	48	6.51
24	4.40	48	6.68
36	5.24	60	7.13
12	3.43	60	7.48
24	4.40	72	8.31
36	5.79	60	7.85
36	5.98	72	8.07
48	6.58	72	8.48
36	5.31	48	6.12
36	5.91	72	8.07

a. Using statistical software, estimate the coefficients of the least squares regression equation.

b. Interpret the meaning of the slope and the intercept in part **a.**

c. Predict the interest rate for a person interested in a four-year auto loan.

d. Should you use the model to predict interest rates for an eight-year loan? Justify your answer.

e. Determine the coefficient of determination and explain its meaning in terms of the problem.

f. Calculate the correlation coefficient for this model. What does it mean?

g. What interest rate would one expect to get if they were planning to apply for a five-year auto loan?

Nassim Taleb

Nassim Nicholas Taleb is a prominent statistician, author, and stock trader. He is widely recognized for coining the term "black swan" to describe rare, unpredictable events that can have a major impact on our lives, society, and predictive models. He has written several influential books including "The Black Swan" and "Antifragile." [11,12]

The COVID-19 pandemic can be considered a black swan event due to its unprecedented nature, global impact, and the significant disruptions it caused across various sectors. Black Swan events can cause severe inaccuracies in the predictive ability of regression models, as you can see if you try to predict S.C. unemployment for 2021.

If we look at the data for South Carolina from January 2010 through January 2020, we can see a clear downward trend of the unemployment rate over time, with some minor ups and downs which are normal due to economic cycles.[10] Using January 2010 as Month 1 and January 2020 as Month 121, we will model the trend by fitting a linear regression model to the data in order to predict the unemployment rate using time (*Month*) as our only independent variable.

The least squares equation is

$$\textit{Estimated Unemployment Rate} = \hat{y} = 11.3805 - 0.07944\ \textit{Month}.$$

The computer output below tells us we have a good model of the unemployment rate using *Month* as the independent variable. Notice the very high R^2 value of almost 97%.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	11.3805	0.0908	125.35	0.000	
Period	-0.07944	0.00129	-61.51	0.000	1.00

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.496235	96.95%	96.92%	96.85%

Use the model to predict the unemployment rate in South Carolina for April 2020.

Solution

Using January 2010 as Month 1, then April 2020 would be Month 124. Replacing this value into our model we get:

$$\textit{Estimated Unemployment Rate} = \hat{y} = 11.3805 - 0.07944(124) = 1.52994.$$

This value would align with the downward trend we observed in the graph. If we also consider our model is explaining almost 97% of the variability in the unemployment rate, then we would be very confident that the unemployment rate in South Carolina in April 2020 will be close to 1.53%. However, if we look at the actual value for April 2020, we find that the unemployment rate was actually 11.7%. That is a large error. What happened? As you may know, in the beginning of 2020 we had an unprecedented global pandemic, which forced a lot of companies to lay off many of their employees, resulting in skyrocketing unemployment rates. No model could have predicted this.

5.4 Exercises

Basic Concepts

1. Why is the mean not a reasonable descriptor for nonstationary time series data?
2. What is a linear time trend?
3. What is the independent variable in a linear trend model?
4. Is there a difference between the way the best fit line is determined for time series data and the way it is determined for other types of data?

Exercises

5. Using the CO_2 Emissions data set from the companion website, look at the CO_2 emissions per capita over time for Chile. Use the data to answer the following.

a. Looking at the data for Chile, do you believe the trend line will slope upward or downward?

b. Suppose we are interested in constructing a linear trend model for the data. Identify the independent and dependent variables for this model.

c. Write the general equation for the time trend model in terms of year and CO_2 emissions per capita.

d. Use statistical software to estimate the least squares model for the data.

e. Use this model to predict the CO_2 emissions per capita for Chile in 2020.

f. Can we determine the accuracy of this prediction? Explain.

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > CO_2 Emissions

6. Consider the following monthly sales data for an up-and-coming technology company.

Sales Data

Month	Sales (Thousands of Dollars)	Month	Sales (Thousands of Dollars)
1	321	7	698
2	542	8	710
3	540	9	799
4	581	10	821
5	641	11	833
6	700	12	850

a. Identify the independent and dependent variables for the linear time trend model.

b. Using statistical software, the following summary output was produced. Write the estimated regression equation.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.949341195
R Square	0.901248704
Adjusted R Square	0.891373575
Standard Error	51.20789475
Observations	12

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	239318.1818	239318.1818	91.26449427
Residual	10	26222.48485	2622.248485	
Total	11	265540.6667		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	403.7575758	31.51628057	12.81107949	1.57569E-07
Month	40.90909091	4.282219283	9.553245222	2.41268E-06

c. What is the mean square error for this model? The standard error?

d. Using this model, predict the company's sales for the 13th month.

e. What percent of the variation in sales is explained by the linear time trend model? Does this model seem to accurately fit the data?

7. Consider the following data on beer production in the United States from 2010 to 2021.[13]

Beer Production in the U.S., 2010 to 2021	
Year	Production (in million barrels)
2010	195.14
2011	192.72
2012	195.74
2013	191.60
2014	192.56
2015	191.00
2016	190.46
2017	185.57
2018	183.28
2019	179.98
2020	179.95
2021	180.89

a. Identify the independent and dependent variables for a linear time trend model.

b. Using statistical software, determine the estimated regression equation for the data.

c. Based on the coefficient of determination, is this model a good fit to the data? Explain.

d. Predict the beer production for the next time period (*Year* = 2022).

e. What might be some external factors that would cause beer production to increase?

f. What might be some external factors that would cause beer production to decrease?

8. Consider the following time series data collected for three variables: Mean Sea Level, Air Pollution from Nitrogen Oxides, and Active Facebook Users Worldwide.

Sea Level, Air Pollution, Facebook Users			
Time Period	**Mean Sea Level (meters)**	**Air Pollution NOx(1,000 tons)**	**Active Facebook Users Worldwide (millions)**
1	0.227	26,883	1936
2	0.058	26,377	2006
3	0.065	27,079	2072
4	0.111	25,757	2129

Sea Level, Air Pollution, Facebook Users			
Time Period	Mean Sea Level (meters)	Air Pollution NOx(1,000 tons)	Active Facebook Users Worldwide (millions)
5	0.189	25,165	2196
6	0.157	24,697	2234
7	0.126	22,335	2271
8	0.079	20,261	2320
9	0.139	14,750	2375
10	0.170	11,563	2414
11	0.148	7985	2449
12	0.173	7645	2498

a. Plot each of the three variables against the *Time Period* variable.

b. What might be some confounding variables that could affect the predictive accuracy of each linear model?

c. Based on the plots in part **a.**, which set of data do you think is the best candidate to model with a linear time trend model? Explain your reasoning.

d. Fit a linear model to the data you selected in part **c.** and calculate the coefficient of determination.

e. For the data selected in part **c.**, plot the data again adding a linear trend line and predict the value for the next observation (*Time Period* = 13).

5.5 Scatterplots for More Than Two Variables

As displayed in Minard's graph of Napoleon's march in Chapter 3, it is often desirable to show more than two variables within the same graphic to determine if a relationship exists. However, a certain degree of creativity is required in order to figure out how to depict the desired variables within the spatial or dimensional limits. As each variable is added to the graph, a new method of reference must be associated with it. Scatterplots can be employed to portray relationships between more than two variables. Suppose we are interested in creating a graph that compares countries based on air pollution and rooms per person. Using data from the Organisation for Economic Cooperation and Development (OECD) Better Life Index for 2022, we created the graph in Figure 5.5.1.[14]

Data

The full OECD Better Life Index 2020 data set can be found on stat.hawkeslearning.com under **Discovering Statistics and Data, Fourth Edition > Data Sets > OECD Better Life Index 2022**.

5.5 Exercises

Basic Concepts

1. What are some methods of displaying measurements other than traditional axes?
2. What are the limitations of multivariable graphs?
3. What is Gapminder?

Exercises

4. Navigate to the Gapminder Trendalyzer tool in a web browser to answer the following questions. Make sure the x-axis is set to *Adjusted GDP per capita*, and the y-axis is set to *Life Expectancy*.
   - **a.** By looking at the graph for the year 2015, do you notice any patterns in the data?
   - **b.** The American Civil War is arguably the most discussed event in American history. It is, by far, the bloodiest war the United States has ever experienced. An estimated 650,000 to 750,000 soldiers died during the Civil War, more than the American casualties from both World Wars, Korea, Vietnam, Iraq, and Afghanistan combined. Between the years 1860 and 1864, how much did the life expectancy in the United States decrease?
   - **c.** In 1918, the entire world experienced a notable change. Describe what happens on the Gapminder graph in the year 1918. What was the cause of the worldwide event? Hint: Use a search engine to do some research.
   - **d.** In the early 1990s, oil reserves were discovered in Equatorial Guinea which lead to a massive boost to the national GDP per capita, and it has remained one of the highest in Africa ever since.
     - **i.** From 1990 to 2008, how much did Equatorial Guinea's GDP per capita increase?
     - **ii.** How much did its life expectancy increase from 1990 to 2008?
     - **iii.** In 2008, how does the life expectancy of Equatorial Guinea compare to countries with a similar GDP per capita?

Technology

To learn how to use the Gapminder Trendalyzer tool, please visit stat.hawkeslearning.com and navigate to **Discovering Statistics and Data, Fourth Edition > Technology Instructions > Gapminder Trendalyzer**.

Lloyd's of London

This very modern looking building is the home of one of the world's largest commercial insurers, Lloyd's of London founded in 1688. At Lloyd's, like all other insurers, risk is measured in probabilities, which are frequently subjective. Lloyd's differs from other insurers in the kinds of policies they write. Lloyd's has written policies on nuclear reactors, space shuttle cargo, oil tankers, art treasures, kidnap and ransom, as well as the legs of ballerinas and football players.

Insurance has a very important place in commerce, and without it, many business activities would not be possible. If a shipping company could not insure its ships, raising the money to buy them would be virtually impossible. Insurance is big business. In addition to sizable revenues, Lloyd's employs about 2,000 people. Lloyd's is a market, rather than an entity. It houses underwriters who evaluate insurance risk for the syndicates they represent. A syndicate is a group of individuals, called Names, who individually assume a small amount of risk in return for a commensurate portion of the premium. For large policies, like an ocean cargo vessel, even a syndicate does not usually underwrite the entire policy; more often groups of syndicates each take a small percentage–thus further diluting each individual's risk.

Probability, Statistics, and Business

Most of the time, when working with samples, statisticians try to deduce from the samples the population parameters (means, proportions, variances, etc.) of certain variables. This process of making judgments about population parameters is called **statistical inference**.

Statistical Inference

Statistical inference involves the use of sample data to form generalizations or inferences about a population. Using sample data to estimate the values of population parameters is one form of statistical inference.

DEFINITION

Because samples are random, there is no guarantee that the sample will be representative of the population. If the sample is not representative, then using the sample mean as an estimate (inference) of the population mean would not be very wise. Probability is used to assess the quality of our inference. Since statistical inference is an inductive process, all statistical conclusions must be endowed with a degree of uncertainty. Because probability is used to assess the reliability of sample inferences, it is a fundamental element of all inferential statistics.

The probability concept also has many direct applications in business. When a manager wonders whether dropping a bid price by 5% will increase the probability of winning the bid, he or she is thinking about chance. Probability is also used as a criteria in designing and evaluating product reliability, insurance valuation, inventory management, project management, and in the study of queuing theory (a probabilistic analysis of waiting lines).

Probability theory emerged from the need to better understand a game of chance. Business decisions, like games, have uncertain outcomes. In an effort to make better decisions, businesses spend considerable amounts of money trying to quantify uncertainty. This means trying to turn uncertainty into a probability. Insurance companies have historically done a good job quantifying uncertainty. In fact, a special kind of statistician, called an actuary, has emerged to assist in the development of insurance models which quantify uncertainty and aid in business decisions.

The next time you watch a 30-second commercial during the Super Bowl, consider the fact that a company has just spent on average $6.5 million for the airtime plus a substantial amount of money developing the advertisement. Without knowing the effect of the advertisement in advance, extensive amounts of money are put at risk with an uncertain outcome. The manager making the decision to invest in the advertisement uses subjective probability to assess the risk and reward.

6.1 Exercises

Basic Concepts

1. Describe randomness.
2. What is probability?
3. List the necessary conditions for a random experiment.

4. What is an outcome? What is an event?
5. What are the two approaches to objective probability?
6. What are some of the problems associated with the relative frequency approach?
7. What is the Law of Large Numbers?
8. Describe the classical approach to probability.
9. Which of the concepts of probability would be most closely associated with the term "empirical probability"?
10. What is statistical inference?
11. Discuss the relationship between probability and statistics.

Exercises

12. Consider the following random experiment. A doctor is interested in determining whether or not his patients think that he listens attentively to what they are saying. He randomly selects several patients and administers an anonymous survey that asks which of the following categories best describes his attentiveness: Very Attentive, Somewhat Attentive, Not Attentive.
    a. Determine the sample space for the above experiment.
    b. Determine all possible outcomes for the event $A = \{$the doctor is not described as Very Attentive$\}$.

13. A gambler has made a weighted die. In order to decide which of the six sides is most likely to turn up, he tosses the die 33 times and notes the number of dots on the uppermost surface. The results of the experiment (sorted) add space are shown in the following table.

Rolls of a Weighted Die										
1	1	1	1	1	1	1	1	1	1	1
1	1	1	2	2	2	2	2	2	2	3
3	3	3	3	4	4	5	5	5	6	6

    a. Using the relative frequency approach, what is the probability of observing each side?
    b. Assuming all outcomes have an equal payoff, which side do you think the gambler will bet on when the die is tossed?
    c. Were your conclusions regarding the probability of each outcome of the roll of the die obtained empirically or deductively?

14. Assume there are two red, two yellow, and two blue buttons in a hat. A button is randomly drawn out of the hat, the color is noted, and the button is returned. This is repeated fifty times. The results are listed in the following table.

Button Drawing				
Yellow	Yellow	Red	Yellow	Red
Red	Red	Blue	Red	Blue
Blue	Red	Red	Yellow	Red
Red	Blue	Yellow	Red	Yellow
Yellow	Blue	Red	Blue	Red
Red	Red	Red	Red	Yellow
Blue	Yellow	Yellow	Blue	Red
Yellow	Red	Red	Red	Yellow
Red	Yellow	Yellow	Yellow	Red
Red	Red	Blue	Red	Blue

a. Using the relative frequency approach, what is the probability of drawing each color?

b. Were the probabilities in part **a.** determined deductively or inductively?

15. Twenty-five small business owners in an urban area are randomly selected and asked if they own a handgun. Twenty-two of those surveyed said that they do own a handgun. If a small business owner is randomly selected from the sample, estimate the probability that the person will own a handgun.

16. Thirty high school teachers are randomly selected and asked if they favor standardized testing. Twenty of those surveyed said that they did favor standardized testing. If a high school teacher is randomly selected from the sample, estimate the probability that the teacher will favor standardized testing.

17. Fifty chief executive officers (CEOs) of publicly traded companies are randomly selected and their salaries are determined. Forty-five of the CEOs selected have salaries in excess of $500,000. If a CEO from one of the selected publicly traded companies is randomly selected from the sample, find the probability that the CEO will have a salary in excess of $500,000.

18. Forty emergency calls to which a local police department responded were randomly selected. Of the forty emergency calls fifteen were categorized as burglaries.

a. Estimate the probability that the next emergency call to which the local police department responds will be a burglary.

b. Which probability interpretation are you using in this problem?

19. For the following situations, decide which probability interpretation is most reasonable to use: relative frequency, subjective, or classical.

a. Whether or not you will have a wreck on your next trip to the mall.

b. Whether or not a car coming off the Ford assembly line will have a defect.

c. The probability that you will graduate from college in four calendar years.

d. Whether a person will be in an automobile accident during the next year.

e. The probability that you will be dealt a full house from a well-shuffled deck of cards.

20. For the following situations, decide which probability interpretation is most reasonable to use: relative frequency, subjective, or classical.

a. Suppose you have purchased a lottery ticket. Describe your chances of winning the lottery.

b. The probability you will enjoy a vacation trip to Mexico.

c. The probability your company's sales will exceed seven million dollars this year.

d. One hundred people receive keys to a new car in a radio contest. Only one key actually fits the car. The probability that key number 25 will open the car door.

e. The probability that you will get a ticket if you drive 70 mph on the interstate between work and home this coming Tuesday.

f. The probability that the S&P 500 will increase or decrease by at least 25 points in one day.

21. Consider a student who is taking a multiple choice examination where there are five possible answers for each question. Since the student has not studied or attended any of the classes, the student decides to randomly guess at each question.

a. Find the probability that the student will answer the first question correctly.

b. Find the probability that the student will answer the first question incorrectly.

22. A game show contestant has to choose one of three doors to win a prize. Behind one door the prize is a trip to Hawaii; behind another door, the prize is a large flatscreen TV; behind the final door, the prize is a bag of potatoes. If a contestant randomly selects a door,

a. Find the probability that the contestant will win a trip to Hawaii.

b. Find the probability that the contestant will not win a trip to Hawaii.

c. Which probability interpretation are you using in this problem?

23. For marketing purposes, an online pet supplies company classifies its customers by state of residence and whether or not they own a pet. The research department has gathered data from a random sample of 682 customers. The data is summarized in the table.

a. What is the probability that a customer lives in Florida?

b. What is the probability that a customer owns a pet?

c. Which probability interpretation are you using in this problem?

State of Residence and Pet Ownership of Customers

State	Owns a Pet	Does Not Own a Pet
Arizona	66	42
Colorado	57	70
Florida	143	118
South Carolina	125	61

6.2 Exercises

Basic Concepts

1. What laws must probability obey, regardless of the methodology used to derive the probabilities?
2. Suppose you are taking a test next week. Interpret each of the following statements.
   a. $P(\text{receiving an A on the test}) = 0$
   b. $P(\text{receiving an A on the test}) = 1$
3. What is a compound event?
4. Draw a Venn diagram to represent the intersection of three events.
5. Define the following set operations: union, intersection, and complement.
6. If you know the probability of two events, what else must you know in order to determine the probability of *one event or the other*?
7. If two events A and B are mutually exclusive, what is $P(A \cap B)$?

Exercises

8. Determine if the following values could be probabilities. If the value cannot be a probability, explain why.
   a. 0
   b. $\frac{36}{25}$
   c. $\frac{7}{8}$
   d. -0.4
   e. 0.23
9. Determine if the following values could be probabilities. If the value cannot be a probability, explain why.
   a. 1
   b. $\frac{15}{16}$
   c. $\frac{4}{3}$
   d. 0.99
   e. -0.05
10. Interpret the following probabilities with respect to the occurrence of some event.
    a. $P(\text{event}) = 0$
    b. $P(\text{event}) = 1.0$
    c. $P(\text{event}) = 0.45$
    d. $P(\text{event}) = 65\%$
    e. $P(\text{event}) = -1.0$
11. Find the following probabilities.
    a. The probability of an event that must happen.
    b. The probability of an event that cannot happen.
    c. The probability of rolling an even number in a single toss of a six-sided die.
    d. The probability of rolling a two and a five in a single toss of a six-sided die.

12. Find the following probabilities related to odds.

a. If the odds in favor of an event A occurring is 3:5, what is the probability of event A?

b. If the odds against an event A occurring is 3:5, what is the probability of event A?

13. The annual premium amounts charged by life insurance companies to their clients are set very carefully. If the amount is too high, the client will take his or her business to another company. If it is too low, the insurance company may not make enough profit to stay in business. In order to properly determine a premium, the company often relies on life tables. These tables allow one to compute the probabilities of death at various ages. They are constructed only after collecting and reviewing extensive data on age at death from a large group of people. A life table is normally constructed assuming that 100,000 people are alive at age 0. This number is simply a reference value used to make comparisons throughout the table. Other numbers could be used. The table then gives the number of people of the original 100,000 that are alive at the beginning of various years of life. In order for the insurance company to optimally set premiums, a separate table should be constructed for the different genders and races. The following abbreviated life table is valid only for females.

Life Table

Year	0	1	5	10	15	20
Number Alive	100,000	99,090	98,912	98,815	98,716	98,477
Year	25	30	35	40	45	50
Number Alive	98,204	97,897	97,500	96,958	96,097	94,766
Year	55	60	65	70	75	80
Number Alive	92,623	89,449	84,565	77,772	68,200	55,535

a. What is the probability that a newborn female lives until the age of 40?

b. What is the probability that a newborn female dies before she reaches the age of 50?

14. A health care provider classifies its customers by their housing situation and whether they have health insurance coverage. The market research department has gathered data from a random sample of 759 customers.

Health Care Consumers

Have Health Insurance Coverage	**Housing Situation**	
	Rent	**Own**
Yes	196	298
No	92	173

a. What is the probability that a customer rents their home?

b. What is the probability that a customer owns their home?

c. What is the probability that a customer has health insurance coverage and rents their home?

d. What is the probability that a customer owns their home and does not have health insurance coverage?

e. What is the probability that a customer has health insurance coverage and rents their home or does not have health insurance coverage and owns their home?

f. What is the probability that a customer does not have health insurance coverage?

g. What approach to probability did you use to calculate your answers?

h. Are the events {rents their home} and {owns their home} mutually exclusive? Explain.

15. A large life insurance company is interested in studying the insurance policies held by married couples. In particular, the insurance company is interested in the amount of insurance held by the husbands and the wives. The insurance company collects data for all of its 1000 policies where both the husband and the wife are insured. The results are summarized in the following table.

Life Insurance Coverage

		Amount of Life Insurance on Husband ($)			
		0–249,999	250,000–499,999	500,000–999,999	1,000,000 or more
Amount of Life Insurance on Wife ($)	0–249,999	400	200	50	50
	250,000–499,999	50	50	30	30
	500,000–999,999	20	10	25	25
	1,000,000 or more	20	10	15	15

a. For a randomly selected policy, what is the probability that the husband will have between $250,000 and $499,999 of insurance?

b. For a randomly selected policy, what is the probability that the wife will have between $500,000 and $999,999 of insurance?

c. For a randomly selected policy, what is the probability that the wife will have $1,000,000 or more of insurance or the husband will have $1,000,000 or more of insurance?

d. For a randomly selected policy, what is the probability that the wife will have between $0 and $249,999 of insurance and the husband will have between $0 and $249,999 of insurance?

e. For a randomly selected policy, what is the probability that the wife will not have between $0 and $249,999 of insurance?

f. For a randomly selected policy, what is the probability that the husband will have $250,000 or more of insurance?

g. What approach to probability did you use to determine your answers?

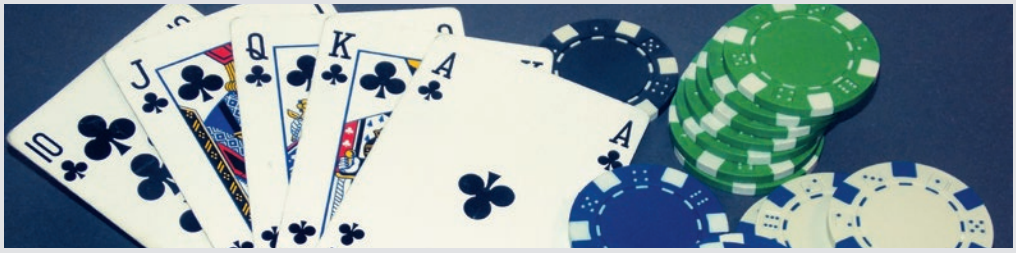
h. Are the events {the wife has $1,000,000 or more in insurance} and {the husband has between $250,000 and $499,999 of insurance} mutually exclusive? Explain.

$$P(B \mid A) = \frac{\left(\begin{array}{c}\text{number of non-spades in the unknown cards remaining} \\ \text{in the deck given the } 4^{\text{th}} \text{community card was a non-spade}\end{array}\right)}{\text{total number of unknown cards remaining in the deck}} = \frac{36}{44} \approx 0.818182.$$

Therefore,

$$P(\text{neither CC is a spade}) = P(A \cap B) = P(A)\,P(B|A) \approx 0.822222 \cdot 0.818182 \approx 0.6727.$$

There is approximately a 67.27% chance of getting non-spades in the next two CCs. Would you call her bet?



Texas Hold'em

Here is a brief overview of how the game is played.

1. The game begins with each player being dealt two cards face down, which are called "hole cards" or "pocket cards."
2. The first round of betting begins with the player to the left of the dealer, who has the option to "call" (match the current bet), "raise" (increase the current bet), or "fold" (discard their hand and end their participation in the current hand). All of the bets throughout the rounds comprise the "pot."
3. After the first round of betting, three community cards are dealt face up in the center of the table, which are called the "flop." These cards can be used by all players to make their best five-card hand.
4. A second round of betting takes place, beginning with the player to the left of the dealer.
5. A fourth community card is dealt face up in the center of the table, which is called the "turn."
6. A third round of betting takes place, beginning with the player to the left of the dealer.
7. A fifth and final community card is dealt face up in the center of the table, which is called the "river."
8. A final round of betting takes place, beginning with the player to the left of the dealer.
9. If more than one player is still in the hand after the final betting round, the players reveal their hole cards and the player with the best five-card hand wins the "pot."

6.3 Exercises

Basic Concepts

1. Define conditional probability.

2. How do you calculate the conditional probability $P(A|B)$?

3. Explain the difference between dependent and independent events.

4. Are mutually exclusive events dependent or independent? Explain your answer.

5. If events A and B are independent, what is $P(A|B)$ equal to?
6. What is the product rule?
7. In the case *People v. Collins* an appeals court overturned the conviction. What flaws did the appeals court detect in the case against the accused assailants?
8. What does it mean to sample with replacement?

Exercises

9. A health care provider classifies its customers by their housing situation and whether they have health insurance coverage. The market research department has gathered data from a random sample of 759 customers.

Health Care Consumers

Have Health Insurance Coverage	Housing Situation	
	Rent	Own
Yes	196	298
No	92	173

   a. Given that the customer rents their home, what is the probability that the customer does not have health insurance?

   b. Given that the customer does not have health insurance, what is the probability that the customer rents their home?

   c. Given that the customer owns their home, what is the probability that the customer has health insurance?

   d. Given that the customer has health insurance, what is the probability that the customer owns their home?

10. The following table was given in Section 6.2, Exercise 15.

Life Insurance Coverage

		Amount of Life Insurance on Husband ($)			
		0–249,999	250,000–499,999	500,000–999,999	1,000,000 or more
Amount of Life Insurance on Wife ($)	0–249,999	400	200	50	50
	250,000–499,999	50	50	30	30
	500,000–999,999	20	10	25	25
	1,000,000 or more	20	10	15	15

   a. Given the wife has between $500,000 and $999,999 of insurance, what is the probability that the husband has $1,000,000 or more of insurance?

   b. Given the wife has between $0 and $249,999 of insurance, what is the probability that the husband has between $0 and $999,999 of insurance?

   c. Given that the husband has between $0 and $250,000 of insurance, what is the probability that the wife will have $1,000,000 or more of insurance?

   d. Given that the husband has $1,000,000 or more of insurance, what is the probability that the wife will have $1,000,000 or more of insurance?

11. A computer software company receives hundreds of support calls each day. There are several common installation problems, call them A, B, C, and D. Several of these problems result in the same symptom, *lock up* after initiation. Suppose that the probability of a caller reporting the symptom *lock up* is 0.7 and the probability of a caller having problem A and a *lock up* is 0.6.

a. Given that the caller reports a lock up, what is the probability that the cause is problem A?

b. What is the probability that the cause of the malfunction is not problem A given that the caller is experiencing a lock up?

12. A television advertising representative has determined the following probabilities based on past experience. The probability that an individual will watch an ad during the Super Bowl is 0.10. Given that the individual watches the ad, the probability that the individual will buy the product is 0.005. It is also known that the probability that an individual would buy the product is 0.02. Given that an individual buys the product, find the probability that the individual watched the television ad during the Super Bowl.

13. Medical researchers have determined that there is a 2% chance that an individual will have a gene which gives him a predisposition for heart disease. Given that an individual has the gene, the probability that heart disease will develop is 25%. It is also known that the probability that an individual has heart disease is 12%.

a. Find the probability that an individual will have the gene and develop heart disease.

b. Given that a person has heart disease, what is the probability that they have the gene?

14. Use the table given in Exercise 9.

Are the events {customer rents their home} and {customer owns their home} independent? Explain.

15. Use the table given in Exercise 10.

Are the events {the husband has $1,000,000 or more in insurance} and {the wife has $250,000 or more in insurance} independent? Explain.

16. Suppose you were flipping a coin. What is the probability that you would observe a head:

a. on two consecutive flips?

b. on three consecutive flips?

c. on four consecutive flips?

d. on 100 consecutive flips?

17. Suppose an atomic reactor has two independent cooling systems. The probability that Cooling System A will fail is 0.01 and the probability that Cooling System B will fail is 0.01. What is the probability that both systems will fail simultaneously?

18. Mandy is 30, and the probability that she will survive until age 65 is 0.90. Ashley is 45, and the probability that she will survive until age 65 is 0.95.

a. Find the probability that both Mandy and Ashley will survive until age 65.

b. Find the probability that only Mandy will survive until age 65.

c. Find the probability that neither Mandy nor Ashley will survive until age 65.

d. What assumption about the lives of Mandy and Ashley did you make in answering the above questions?

19. An insurance company is considering insuring two large oil tankers against spills. The limit of the liability on the coverage is $10,000,000. The company believes that the probability of an oil spill requiring the maximum liability coverage during the policy period is 0.001 per tanker.

a. What is the probability that neither tanker would have a spill requiring the maximum liability coverage during the policy period?

b. What is the probability that only one tanker would have a spill requiring the maximum liability coverage during the policy period?

c. What is the probability that both tankers would have spills requiring the maximum liability coverage during the policy period?

20. Coin flipping can be used to model other real life phenomena and aid in certain probability calculations. An example of this would be to compute the probability that the World Series ends in some specified number of games. The World Series is a best of seven game series played at the end of the regular baseball season between the champion of the American League and the champion of the National League. The first team to win four games is declared the champion of baseball for that year. If we assume the probability of either team winning a game is approximately 0.5 and the games are independent events, the probability that the series ends in either 4, 5, 6, or 7 games can be computed.

a. What is the probability that the series ends in exactly 4 games? Write the sample space consisting of 16 equally likely outcomes similar to the sample space resulting from tossing a coin four times.

b. What is the probability that the series ends in exactly 5 games?

c. Assume the probability that the series ends in exactly 6 games is $\frac{5}{16}$. Use this information together with your answers to the first two parts of this problem to compute the probability that the series ends in exactly 7 games.

21. Drug usage in the workplace costs employers incredible amounts of money each year. Drug testing potential employees has become so prevalent that drug users are finding it extremely hard to find jobs. Drug tests, however, are not completely reliable. The most common test used to detect drugs is approximately 98% accurate. To decrease the likelihood of making an error, all potential employees are screened through two tests, which are independent, and each has about 98% accuracy.

a. If a person were drug-free, what is the probability they would fail both tests?

b. If a person were a drug user, what is the probability they would pass both tests?

22. Suppose you draw two cards out of a standard deck without replacement. What is the probability that you draw the ace of spades and then another spade?

23. Manchester United F.C. has 42 players on their first team roster, classified by position and age group as follows.

Manchester United F.C. First Team Roster		
	18-22 year olds	**23-30 year olds**
Defender	9	13
Midfielder	4	5
Forward	4	2
Goalkeeper	2	3

Their head coach must choose two players at random for a press conference.

a. What is the probability that a midfielder age 18-22 and a defender age 23-30 are chosen?

b. What is the probability that two players age 18-22 are chosen?

6.4 Combinations and Permutations

To compute certain probabilities, such as the probability of having winning numbers in the state lottery, requires the ability to count the number of possible outcomes for a given experiment or a sequence of experiments.

However, often it is impractical to list out all the possibilities. Therefore, we will develop some techniques to facilitate counting.

The Fundamental Counting Principle

E_1 is an event with n_1 possible outcomes and E_2 is an event with n_2 possible outcomes. The number of ways the events can occur in sequence is $n_1 \cdot n_2$. This principle can be applied for any number of events occurring in sequence.

PROCEDURE

Example 6.4.1

Using the Fundamental Counting Principle to Count the Number of Tomato Plots Needed for an Experiment

An agricultural research center is designing an experiment to determine the effects of soil type, fertilizer, and plant spacing on tomato yield. The plan calls for four soil types, three types of fertilizer, and four different spacings between plants. The plan also includes replication of each combination of fertilizer, soil type, and plant spacing five times. How many different plots of tomatoes will be needed to conduct the experiment?

Solution

$$\underset{\text{(soil types)}}{4} \cdot \underset{\text{(fertilizers)}}{3} \cdot \underset{\text{(plant spacings)}}{4} \cdot \underset{\text{(replications)}}{5} = 240$$

6.4 Exercises

Basic Concepts

1. What is the Fundamental Counting Principle?
2. What is a factorial and how is it calculated?
3. Describe the difference between permutations and combinations.
4. Give an example of a situation in which you would need to determine the number of distinguishable permutations.

Exercises

5. The blue plate lunch at a local cafeteria consists of an entrée, a side item, and a dessert. If there are 6 choices for an entrée, 5 choices for a side item, and 4 choices for a dessert, how many different lunches are available?

6. You are interested in buying a home in a new subdivision. The builder offers 3 basic floor plans, each with 4 possible arrangements for the garage, and siding in 6 different colors. How many different homes can be built?

7. Compute each of the following.

   a. 1!
   b. 3!
   c. 5!
   d. 7!

8. A wedding planner needs to select 6 songs from a playlist containing 12 songs to play at the reception. How many different sequences are possible?

9. In how many ways can 11 kids be picked for the 9 positions on a baseball team?

10. How many distinguishable permutations can be made from the word STATISTICS?

11. How many distinguishable permutations can be made from the word SASSAFRAS?

12. A person tosses a coin 11 times. In how many ways can he get 9 heads?

13. How many 5 card hands can be dealt from a deck of 52 cards?

14. There are eight people hosting a party. Three people are needed to decorate for the party. How many ways can the decorating crew be chosen?

15. In how many ways can a graduate student fulfill their degree requirements in statistics if 10 classes are needed from a choice of 15 classes?

16. The Johnson family is planning their vacation. Each of the five family members is allowed to nominate three places they would like to visit. If they want to visit four different places during the trip, in how many ways can they plan their trip, assuming that no family members choose the same place?

17. Kara was born on 11/21/1992. She would like to make an eight-digit password using all of the digits in her birth date. How many different eight-digit passwords could she create?

18. Employees at a local software company need a unique seven-digit code to access the building. The manager wants to make each person's code from the company's phone number, 555-8212.

a. If there are 509 employees who need codes, will the manager have enough unique codes using only the digits in the phone number?

b. Would there be enough ten-digit codes if he used the area code, 516, as well?

19. Which of the following words would produce the greatest number of different five-letter arrangements? (**Hint:** Think before you calculate!)

a. TEARS

b. STOPS

c. TESTS

d. ROOST

20. Receipts often show the last four digits of your credit card. Assume American Express offers 15-digit cards starting with 34 or 37. If a thief has the last four digits of your American Express credit card, what is the probability of them correctly guessing the first 11 digits? Express your answer as a fraction.

6.5 Bayes' Theorem

We discussed conditional probability and independence in Section 6.3. **Bayes' Theorem** (also referred to as **Bayes' Rule** or **Bayes' Law**) is a clever way of obtaining a conditional probability given new information. The additional information is obtained for a subsequent event and is used to revise the initial probability. We begin with an example.

Example 6.5.1

Using Bayes' Theorem to Determine a Probability

Suppose that 85% of all passengers in an airport fly on a major airline, while the remaining 15% fly on a small airline. Of those passengers traveling on a major airline, suppose we know that 65% are traveling for business. Of those passengers traveling on a small airline, 25% are traveling for business. (Notice that even though we only talk about business passengers, there are also implied non-business passengers as well.) Now a business passenger is selected at random. What is the probability that the business passenger traveled on a major airline?

Solution

Let's first define the events associated with this problem.

$M =$ Major Airline
$S =$ Small Airline
$B =$ Business Passenger

Here are the probabilities given to us in the problem.

$$P(M) = 0.85$$
$$P(S) = 0.15$$
$$P(B \mid M) = 0.65$$
$$P(B \mid S) = 0.25$$

To determine the probability that the selected business passenger traveled on a major airline, we need to find the probability $P(M \mid B)$. Using the four probabilities above and Bayes' Theorem, we proceed as follows.

$$P(M \mid B) = \frac{P(M \cap B)}{P(B)}$$

By the definition of a conditional probability.

$$= \frac{P(M \cap B)}{P(B \cap M) + P(B \cap S)}$$

The denominator says that all business passengers travel either on a major airline or on a small airline; those are the only two alternatives and they are mutually exclusive. Thus, $P(B)$ is equivalent to the denominator.

$$= \frac{P(M) \cdot P(B \mid M)}{P(M) \cdot P(B \mid M) + P(S) \cdot P(B \mid S)}$$

The numerator and denominator result from rearranging the conditional probability formula and solving for the probability of the intersection of two events. Note that $P(M \cap B)$ is equal to both $P(M) \cdot P(B \mid M)$ and $P(B) \cdot P(M \mid B)$ by the definition of conditional probability.

$$= \frac{0.85 \cdot 0.65}{0.85 \cdot 0.65 + 0.15 \cdot 0.25}$$

Substitute the probability values given in the problem.

$$= \frac{0.5525}{0.59} = 0.936 \approx 94\%$$

Therefore, we know that if the passenger was traveling for business, there is about a 94% chance that he or she will be traveling on a major airline.

Notice how the conditional probability of 94% is not intuitive. It is called the **posterior probability**. The **prior probability** of a passenger traveling on a major airline of 85% has been increased to 94%, given the information that the passenger was traveling for business purposes.

The following is a formal statement of Bayes' Theorem.

Bayes' Theorem

Let A be an event and B_1, B_2, ..., B_N be N mutually exclusive and collectively exhaustive events. Then Bayes' Theorem states,

$$\begin{aligned} P(B_i \mid A) &= \frac{P(B_i \cap A)}{P(A)} \\ &= \frac{P(B_i \cap A)}{P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_N)} \\ &= \frac{P(B_i) \cdot P(A \mid B_i)}{P(B_1) \cdot P(A \mid B_1) + P(B_2) \cdot P(A \mid B_2) + \cdots + P(B_N) \cdot P(A \mid B_N)} \\ &= \frac{P(B_i) \cdot P(A \mid B_i)}{\sum_{i=1}^{N} P(B_i) \cdot P(A \mid B_i)}. \end{aligned}$$

THEOREM

Let's look at another example using Bayes' Theorem.

Example 6.5.2

Using Bayes' Theorem to Determine the Probability of Having a Disease

Let D be the event that a person has a rare disease. Suppose that the rare disease has an incidence rate of 1% in the population, $P(D) = 0.01$. $\overline{D}$ is the event that a person does not have the rare disease (i.e., the complement of D). Suppose a machine is used to diagnose the disease. Let C be the event that the disease is confirmed as the diagnosis. Suppose that the probability of the machine falsely confirming the disease when one doesn't have it is $P\left(C \mid \overline{D}\right) = 0.15$, called a *false positive*; while $P(C \mid D) = 0.95$, which says that the machine correctly confirms the disease with an accuracy of 95%. Now, suppose that the machine confirms that a person has the disease. What is the probability that the person actually has the disease? In other words, what is $P(D \mid C)$?

Solution

Here are the probabilities given to us in the problem.

$$\begin{aligned} P(D) &= 0.01 \\ P(\overline{D}) &= 0.99 \\ P(C \mid D) &= 0.95 \\ P(C \mid \overline{D}) &= 0.15 \end{aligned}$$

To find the probability that a person with a positive diagnostic result actually has the disease we proceed as follows.

$$
\begin{aligned}
P(D \mid C) &= \frac{P(D \cap C)}{P(C)} \\
&= \frac{P(D \cap C)}{P(C \cap D) + P(C \cap \overline{D})} \\
&= \frac{P(D) \cdot P(C \mid D)}{P(D) \cdot P(C \mid D) + P(\overline{D}) \cdot P(C \mid \overline{D})} \\
&= \frac{0.01 \cdot 0.95}{0.01 \cdot 0.95 + 0.99 \cdot 0.15} \\
&= \frac{0.0095}{0.0095 + 0.1485} = \frac{0.0095}{0.1580} \approx 0.06
\end{aligned}
$$

This is somewhat of an assuring result in that you have only a 6% chance of having the disease even though the machine yielded a positive diagnostic.

6.5 Exercises

Basic Concepts

1. Briefly explain the relationship between conditional probability and Bayes' Theorem.
2. What is the difference between prior and posterior probabilities?
3. What is Bayes' Theorem?
4. How is Bayes' Theorem used to "revise" a probability based on additional information?

Excercises

5. In a production line, 8% of all items produced are defective. 75% of all defective items are fully inspected, while 10% of all non-defective items go through a complete inspection. Given that an item is completely inspected, what is the probability that it is defective?

6. The issue of Corporate Tax Reform has been cause for much debate in the United States, especially in the House Ways and Means Committee as well as the Senate Finance Committee. Among those in the legislature, 45% are Republicans and 55% are Democrats. It is reported that 30% of the Republicans and 70% of the Democrats favor some type of Corporate Tax Reform to prevent American companies from operating in foreign countries. Suppose a member of Congress is randomly selected and they are found to favor some type of corporate tax reform. What is the probability that this person is a Democrat?

Randomness and Statistics: Part 1

Henri Poincaré was a French mathematician, theoretical physicist, and philosopher of science and is considered one of the greatest mathematicians of all time. Concerning the game of roulette, he remarked that the game appears random because we are ignorant of what causes the various results (numbers). That is, if we had a complete mathematical model of the physics of the roulette wheel and the initial conditions (speed of the ball, speed of the wheel, and position on the wheel when the ball was thrown), we could predict with certainty what the outcome would be. However, since we don't have such a model, the whole process appears random because we are ignorant of the initial conditions and the physical model.

One of the reasons we study probability distributions is that they provide a means of describing our state of ignorance about the outcome of some random phenomena. One of the reasons we study statistics is that it can teach us how to develop statistical models that can reduce our ignorance of seemingly random phenomena. Statistical prediction models have been developed for climate modeling, weather forecasting, ecological modeling, sports modeling, crime prediction, stock market prediction, economic forecasting, credit scoring, disease prediction, epidemiology, horse racing, natural language processing, and autonomous systems, among others.

Probability Distribution	
x	$P(X = x)$
1	$\frac{1}{30}$
2	$\frac{4}{30}$
3	$\frac{9}{30}$
4	$\frac{16}{30}$
Total	$\sum P(X = x_i) = \frac{30}{30} = 1$

Note that the distribution possesses the essential properties of all probability distributions; that is, the probabilities sum to one, and all the probabilities are between 0 and 1.

Continuous Probability Distributions

Continuous random variables also have probability distributions. We often describe their probability distributions using an equation or graph rather than a table since it is impossible to list all of the values in the range of a continuous random variable. They will be the subject of the next chapter.

7.1 Exercises

Basic Concepts

1. What is a random variable?
2. What is a probability distribution?
3. Do all random variables have probability distributions?
4. Explain the concept of an "empirical law".
5. What are the two types of random variables discussed in this section? What distinguishes the two types?
6. What is the difference between randomness and uncertainty?
7. What is the difference between a probability distribution function and a probability model?
8. Discrete probability distributions always have two characteristics; what are they?
9. What is the value of describing a random variable with a probability distribution?
10. Identify three different ways to express possible values of a random variable along with their associated probabilities.
11. How is a probability distribution created?
12. What is a probability distribution function?

13. Describe what methods Fat the Butch could have used to develop a probability distribution for the game he was playing. What about Pascal and Fermat?

14. How are subjective probabilities usually expressed? Give an example.

15. Sabine Hossenfelder, the particle physicist and YouTube content creator with 850,000+ subscribers, mentioned in one of her videos that with respect to quantum mechanics the only way you can see the "wave function" of a particle is through its probability distribution. How do you see a probability distribution?

Exercises

16. Classify the following as either a discrete random variable or a continuous random variable.

- **a.** The number of pages in a standard math textbook.
- **b.** The amount of electricity used daily in a home.
- **c.** The number of customers entering a restaurant in one day.
- **d.** The time spent daily on the phone after supper by a teenager.
- **e.** Campers at a state park over Labor Day weekend.

17. Classify the following as either a discrete random variable or a continuous random variable.

- **a.** The speed of a train.
- **b.** The possible scores on the SAT exam.
- **c.** The number of pizzas eaten on a college campus each day.
- **d.** The daily takeoffs at Chicago's O'Hare Airport.
- **e.** The highest temperatures in Maine and Florida tomorrow.

18. Classify the following as either a discrete random variable or a continuous random variable.

- **a.** The number of emergency phone calls received per day by a local fire department.
- **b.** The speed of pitches of major league baseball pitchers.
- **c.** The weight of a lobster caught in Maine.
- **d.** The number of defective circuits on a computer chip.
- **e.** The time it takes for a 5-year battery to die.

19. Classify the following as either a discrete random variable or a continuous random variable.

- **a.** The total points scored per football game for a local high school team.
- **b.** The daily price of a stock.
- **c.** The interest rate charged by local banks for 30-year mortgages.
- **d.** The number of times a backup of the computer network is performed in a month.
- **e.** The amount of sugar imported by the U.S. in a day.

20. Determine whether or not the following distribution is a probability distribution. If the distribution is not a probability distribution, give the characteristic which is not satisfied by the distribution.

x	1	2	3
$P(X = x)$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3}$

21. Determine whether or not the following distribution is a probability distribution. If the distribution is not a probability distribution, give the characteristic which is not satisfied by the distribution.

x	-2	2	3
$P(X = x)$	0.25	0.50	0.25

22. Tell whether or not the following distribution is a probability distribution. If the distribution is not a probability distribution, give the characteristic which is not satisfied by the distribution.

x	2	3	4	5
$P(X = x)$	0.30	-0.50	0.50	0.70

23. Determine whether or not the following distribution is a probability distribution. If the distribution is not a probability distribution, give the characteristic which is not satisfied by the distribution.

x	5	10	15
$P(X = x)$	0.46	0.25	0.25

24. Determine whether or not the following distribution is a probability distribution. If the distribution is not a probability distribution, give the characteristic which is not satisfied by the distribution.

x	-10	-5	3	8
$P(X = x)$	0.18	0.39	0.08	0.35

25. Determine whether or not the following distribution is a probability distribution. If the distribution is not a probability distribution, give the characteristic which is not satisfied by the distribution.

x	100	200	300
$P(X = x)$	-0.10	0.50	0.50

26. Determine whether or not the following distribution is a probability distribution. If the distribution is not a probability distribution, give the characteristic which is not satisfied by the distribution.

$$P(X = x) = \frac{x}{16}, \text{ for } x = 1, 2, 3, 4, 5$$

7.2 Exercises

Basic Concepts

1. Why is the notion of expected value one of the most important concepts in the analysis of random phenomena?
2. True or False: The expected value of a discrete random variable is usually one of the possible outcomes of the random variable.
3. Suppose the expected value of a random variable was known to be 6.3. Interpret the meaning of the expected value.
4. Give an example of a situation in which expected value would be useful to compare alternatives, other than the one used in the section.
5. How is the variability of a random variable related to risk?
6. True or False: Two random variables that have exactly the same probability distributions will have the same expected value.
7. True or False: Two random variables with the same expected value will have the same probability distributions.

Exercises

8. Determine the expected value, the variance, and the standard deviation for a random variable with the following probability distribution.

x	−5	−2	0	2	5
$p(x)$	0.06	0.15	0.58	0.18	0.03

9. Determine the expected value, the variance, and the standard deviation for a random variable with the following probability distribution.

x	400	420	440	460	480	500
$p(x)$	0	0.1	0.1	0.2	0.2	0.4

10. A regional hospital is considering the purchase of a helicopter to transport critical patients. The relative frequency of X, the number of times the helicopter is used to transport critical patients each month, is derived for a similarly sized hospital and is given in the following probability distribution.

x	0	1	2	3	4	5	6
$p(x)$	0.15	0.20	0.34	0.19	0.06	0.05	0.01

   a. Determine the average number of times the helicopter is used to transport critical patients each month.
   b. Determine the variance of the number of times the helicopter is used to transport critical patients.
   c. Determine the standard deviation of the number of times the helicopter is used to transport critical patients.

d. Determine the probability that the helicopter will not be used at all during a month to transport critical patients.

e. Determine the probability that the helicopter will be used at least once to transport critical patients.

f. Determine the probability that the helicopter will be used at most twice to transport critical patients.

g. Determine the probability that the helicopter will be used more than three times to transport critical patients.

11. Based on past experience, an architect has determined a probability distribution for X, the number of times a drawing must be examined by a client before it is accepted.

x	1	2	3	4	5
$p(x)$	0.1	0.2	0.3	0.2	0.2

a. Find the average number of times a drawing must be examined by a client before it is accepted.

b. Find the variance of the number of times a drawing must be examined by a client before it is accepted.

c. Find the standard deviation of the number of times a drawing must be examined by a client before it is accepted.

d. What is the probability that a drawing must be examined five times before being accepted by the client?

e. Find the probability that the drawing must be examined at least twice before being accepted by the client.

f. Find the probability that a drawing must be examined at most three times before being accepted by the client.

g. Find the probability that a drawing must be examined less than twice before being accepted by the client.

12. The manager of a retail clothing store has determined the following probability distribution for X, the number of customers who will enter the store on Saturday.

x	10	20	30	40	50	60
$p(x)$	0.10	0.20	0.30	0.20	0.10	0.10

a. Find the expected number of customers who will enter the store on Saturday.

b. Find the standard deviation of the number of customers who will enter the store on Saturday.

c. Find the variance of the number of customers who will enter the store on Saturday.

d. Find the probability that more than 30 customers will enter the store on Saturday.

e. Find the probability that at most 20 customers will enter the store on Saturday.

f. Find the probability that at least 40 customers will enter the store on Saturday.

g. What is the probability that exactly 10 customers will enter the store on Saturday?

13. An entrepreneur is considering investing in a new venture. If the venture is successful, he will make $50,000. However, if the venture is not successful, he will lose his investment of $10,000. Based on past experience, he believes that there is a 40% chance that the venture will be successful.

a. Use the information in the problem to determine the probability distribution of the amount of money to be made (or lost) on the venture.

b. Determine the expected amount of money to be made on the venture.

c. Determine the standard deviation of the amount of money to be made on the venture.

14. An investor is considering two alternative investment options with the following payoff distributions.

	Option 1			Option 2		
Payoff	−$100,000	$30,000	$100,000	−$20,000	$0	$20,000
***P*(Payoff)**	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0.25	0.50	0.25

a. Determine the expected payoff for each of the investment options.

b. Determine the standard deviation of the payoff for each of the investment options.

c. Which investment option would you choose? Explain.

15. A cereal manufacturer has two new brands of cereal which it would like to produce. Because resources are limited, the cereal manufacturer can only afford to produce one of the new brands. A marketing study produced the following probability distributions for the amount of sales for each of the new brands of cereal.

Cereal A	
Sales	***P*(Sales)**
$150,000	0.2
$200,000	0.3
$300,000	0.3
$400,000	0.2

Cereal B	
Sales	***P*(Sales)**
$10,000	0.40
$300,000	0.40
$600,000	0.10
$1,000,000	0.10

a. What are the expected sales of each of the new brands of cereal?

b. What is the standard deviation of the sales for each of the brands of cereal?

c. If both of the brands of cereal cost the same amount to produce, which brand of cereal do you think the cereal manufacturer should produce? Explain.

7.3 Exercises

Basic Concepts

1. What is the most significant property of the discrete uniform distribution?
2. Under what circumstances is the uniform distribution often used as an initial alternative?
3. Why is the discrete uniform probability model selected to estimate the number of German tanks being produced per month during WWII?
4. What parameter in the discrete uniform probability model is being estimated in the German tank problem?
5. Can a 60-sided die be used to simulate the probability distribution of a roulette wheel with 38 unique numbers? If yes, how?

Exercises

6. Tristen walks into class and states that he has an extra ticket to a concert on Friday night. He asks anyone interested in attending the concert to put their name on a piece of paper and put it in a basket. He plans to draw from the basket to choose the person who will attend the concert with him. If you and 16 other people in the class put your names in the basket what is the chance of being chosen to attend the concert with him?

7. Sharlene has just put a down payment on a lot in a small subdivision. There are 10 lots in the subdivision, and all are approximately 0.25 acres in size. Five builders have been contracted by the subdivision manager to each build two homes in order to finish the subdivision in 6 months. Jonathan is one of the builders contracted by the subdivision manager. What is the probability that Jonathan will be the builder that builds her house?

8. You are going to the casino tonight and plan on playing roulette.
   a. Find the expected value and variance of betting \$1 on red.
   b. What would be your betting strategy to maximize winnings?

9. An experiment consists of tossing a coin and rolling a six-sided die simultaneously.
   a. List the sample space for the experiment.
   b. What is the probability of getting a head on the coin and the number 3 on the die?
   c. What is the probability of getting a tail on the coin and at least a 4 on the die?

10. Given the following discrete uniform probability distribution, find the expected value and standard deviation of the random variable.

x	0	1	2	3	4
$P(X = x)$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

Randomness and Statistics: Part 2

The roulette wheel features 36 red and black numbers from 1 to 36, along with two green slots labeled "0" and "00." When the croupier spins the wheel and the ball eventually lands in one of the 38 slots, this outcome is typically viewed as a random event. The uniform distribution serves as a descriptive model of this randomness, reflecting our state of ignorance about where the ball will land. In this model, each number is equally likely to occur—with a 1/38 probability—encapsulating our uncertainty about the outcome.

Claude Shannon and Edward Thorp (both at MIT) developed a system to model this seemingly random process. They used sensors and a timing mechanism to gather data on the ball's speed and the wheel's rotation, and they input this data into a model they implemented in a wearable computer—in their shoe! The computer then indicated the most likely winning number or section of numbers in which the ball would land. By using their model and data, they could place more accurate bets and increase their chances of winning. Essentially, they changed their level of ignorance about the random phenomena, which is the essence of effective modeling.

Variance and Standard Deviation of a Binomial Random Variable

To find the **variance** of a binomial random variable, use the expression

$$\sigma^2 = V(X) = np(1 - p).$$

Therefore, the **standard deviation** of a binomial random variable is given by

$$\sigma = \sqrt{V(X)} = \sqrt{np(1-p)}.$$

FORMULA

Example 7.4.6

Computing the Expected Value and Variance of a Binomial Random Variable

Compute the expected value and the variance of the number of profitable leases in Example 7.4.3.

Solution

Since the random variable is binomial, we can use the formulas above. Since $n = 12$ and $p = 0.1$, the expected value is given by the following expression.

$$\begin{aligned}\mu &= E(X) = np \\ &= 12(0.1) \\ &= 1.2\end{aligned}$$

The variance is

$$\begin{aligned}\sigma^2 &= V(X) = np(1-p) \\ &= 12(0.1)(0.9) \\ &= 1.08,\end{aligned}$$

which implies that the standard deviation is $\sqrt{1.08} \approx 1.039$.

Thus, if groups of 12 oil leases were purchased with the same probability of success (0.1 probability of a profitable lease), then the average number of profitable leases per group of 12 would be 1.2 and the standard deviation would be 1.039 leases.

7.4 Exercises

Basic Concepts

1. Describe the characteristics of a binomial experiment.
2. What are the parameters of a binomial probability model?
3. Give an example of a binomial experiment, other than the one used in the section.
4. What is the formula for the binomial probability distribution function?
5. What influences the shape of the binomial probability distribution?
6. How do you determine the expected value of a binomial random variable? The variance? The standard deviation?

Exercises

7. Compute ${}_nC_x$ for each of the following combinations of x and n.

a. $n = 5, x = 4$

b. $n = 10, x = 8$

c. $n = 15, x = 1$

d. $n = 20, x = 0$

8. Compute ${}_nC_x$ for each of the following combinations of x and n.

a. $n = 4, x = 2$

b. $n = 12, x = 8$

c. $n = 18, x = 15$

d. $n = 23, x = 20$

9. The random variable X is a binomial random variable with $n = 9$ and $p = 0.1$.

a. Find the expected value of X.

b. Find the standard deviation of X.

c. Find the probability that X equals 2. (Use the formula for $P(X = x)$.)

d. Find the probability that X is at most 3.

e. Find the probability that X is at least 2.

f. Find the probability that X is less than 5.

10. The random variable X is a binomial random variable with $n = 12$ and $p = 0.8$.

a. Find the expected value of X.

b. Find the standard deviation of X.

c. Find the probability that X equals 7. (Use the formula for $P(X = x)$.)

d. Find the probability that X is at most 4.

e. Find the probability that X is at least 1.

f. Find the probability that X is more than 10.

11. A real estate agent has ten properties that she shows. She feels that there is a ten percent chance of selling any one property during a week. The chance of selling any one property is independent of selling another property.

a. What probability model would be appropriate for describing the number of properties sold each week?

b. Compute the expected number of properties to be sold in a week.

c. Compute the standard deviation of the number of properties sold each week.

d. Compute the probability of selling one property in one week.

e. Compute the probability of selling five properties in one week.

f. Compute the probability of selling at least three properties in one week.

12. A small commuter airline is concerned about reservation no-shows and, correspondingly, how much they should overbook flights to compensate. Assume their commuter planes will hold 15 people. Industry research indicates that 20% of the people making a reservation will not show up for a flight. Whether or not one person takes the flight is considered to be independent of other persons holding reservations.

a. What probability model would be appropriate for the number of passengers that actually take the flight?

b. If the airline decides to book 18 people for each flight, how often will there be at least one person who will not get a seat?

c. If they book 17 people, how often will there be at least one person who will not get a seat?

d. If they book 16 people, how often will there be at least one person who will not get a seat?

e. If they book 18 people for each flight, how often will there be one or more empty seats?

f. If they book 17 people, how often will there be one or more empty seats?

g. If they book 16 people, how often will there be one or more empty seats?

h. Based on the results from parts **b.** to **g.** above, which booking policy do you prefer? Explain your answer.

13. Seven plants are operated by a garment manufacturer. They feel there is a ten percent chance for a power outage to occur in a month at any one plant and the risk of a power outage at one plant is independent of the risk of a power outage at another plant. Let X = the number of plants of the garment manufacturer that have a power outage in the next month.

a. Determine the probability distribution for X.

b. Interpret the results for $P(X=0)$, $P(X=4)$, and $P(X=7)$.

c. Compute the expected value of X.

d. Compute the standard deviation for X. Is this value large in relation to the expected value? In what units is the standard deviation expressed?

14. A company that makes traffic signal lights buys switches from a supplier. Out of each shipment of 1000 switches, the company will take a random sample of 10 switches. Let X equal the number of defective switches in the sample.

a. The company has a policy of rejecting a lot if they find any defective switches in the sample. What is the probability that the shipment will be accepted if, in fact, 2% of the switches are actually defective?

b. What is the probability that the shipment will be accepted if the percent of defective switches is actually 5%?

c. The company decides to change their policy and will accept the lot if they find no more than one defective switch. Repeat parts a. and b. for this new policy.

15. The probability of getting into medical school if one or both of your parents is a doctor is 0.7. If a doctor has four children answer the following questions.

a. Determine the probability of getting exactly two children into medical school.

b. What is the probability of getting at least two children into medical school?

16. A certain aspirin is advertised as being preferred by 4 out of 5 doctors. If the advertisement is assumed to be true, answer the following questions.

a. What is the probability that at least half of ten doctors chosen at random will prefer this brand of aspirin?

b. What is the probability that 9 out of 10 of the doctors will prefer this brand?

17. In manufacturing integrated circuits, the yield of the manufacturing process is the percentage of good chips produced by the process. The probability that an integrated circuit manufactured by the Ace Electronics Company will be defective is $p = 0.05$. If a random sample of 15 circuits is selected for testing, answer the following questions.

a. What is the probability that no more than one integrated circuit will be defective in the sample?

b. What is the expected number of defective integrated circuits in the sample?

18. A local employment service procures temporary office personnel for local businesses. They have found that 90% of the invoices for their services are paid within 10 working days. If a random sample of 12 invoices is checked, answer the following questions.

a. What is the probability that all of the invoices will be paid within 10 working days?

b. What is the probability that six or more of the invoices will be paid within 10 working days?

19. An experiment consists of rolling a pair of typical six-sided dice 10 times. On each roll the sum of the dots on the two dice is noted.

a. Find the probability that on any roll of the two dice the sum of the dots is either 7 or 11.

b. Find the probability that in the 10 rolls of the pair of dice, a 7 or 11 occurs 5 times.

c. Find the probability that in the 10 rolls of the pair of dice, a 7 or 11 does not occur at all.

d. Find the mean and variance of the number of times we see a 7 or 11 in the 10 rolls of the dice.

20. "Would you say you eat to live or live to eat?" was asked to each person in a sample of 1001 adults in a recent survey. Seventy-four percent of the respondents answered eat to live, 23% answered live to eat, and 3% had no opinion. Assuming these percentages are accurate, find the probability, in 12 randomly chosen adults, that the number who would answer "eat to live" is:

a. exactly 7.

b. no more than 10.

c. at most 11.

d. at least 3.

Poisson Random Variables for Length or Space

Instead of counting the number of successes in a time interval, there are a number of applications of the Poisson that measure the number of successes in some area or length. The average number of successes in the area or length will define the parameter of the Poisson random variable.

Example 7.5.2

Determining a Probability Using the Poisson Distribution

The telephone company is considering purchasing optical cable from Optica, Inc. The company wishes to replace approximately 100,000 feet of conventional cable with optical fiber. Since optical fiber is very difficult to repair, it is important that the number of optical cable defects are minimized. Optica claims that on average there is one defect per 200,000 feet of cable. What is the probability that the replaced cable will contain no defects?

Solution

Let $X =$ the number of defects in 100,000 feet of optical cable.

Based on previous experience, we assume that the number of defects is approximated by a Poisson distribution with Poisson parameter λ computed as follows.

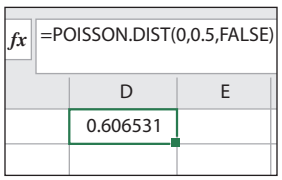
$$\lambda = \frac{100{,}000}{200{,}000} = \frac{1}{2} \quad \text{(the average number of defects per 100,000 feet of cable)}$$

Using the tables provided in Appendix A, Table F,

$$P(X = 0) = 0.6065.$$

Technology

A Poisson probability can also be found using the POISSON.DIST function in Excel. For instructions please visit stat.hawkeslearning.com and navigate to **Discovering Statistics and Data, Fourth Edition > Technology Instructions > Poisson Distribution > Poisson Probability (pdf)**.



7.5 Exercises

Basic Concepts

1. How is the Poisson distribution similar to the binomial distribution?
2. What are the two conditions that an experiment must meet in order to be considered a Poisson random variable?
3. What are some uses of the Poisson probability model in the real world?
4. What is the Poisson probability distribution function?
5. What is the parameter of the Poisson probability model?
6. What is the expected value of a Poisson random variable? The variance? The standard deviation?

Exercises

7. In the last six months, on average 5 shoppers downtown had their automobiles broken into each month while they shopped. What is the probability that exactly 2 shoppers will have their automobiles broken into next month?

8. The number of calls received by an office on Monday morning between 8:00 AM and 9:00 AM has a Poisson distribution with λ equal to 4.0.

a. Determine the probability of getting no calls between eight and nine in the morning.

b. Determine the probability of getting exactly five calls between eight and nine in the morning.

c. What will be the expected number of calls received by the office during this time period? What is the variance?

d. Graph the probability distribution of the number of calls using values from Appendix A, Table F.

9. The director of a local hospital is studying the occurrence of medication errors. Medication errors are deemed to occur when a patient is given the wrong amount of medication, or the wrong medication is given to a patient. Based on past experience, the director believes that medication errors follow a Poisson process with an average rate of 2 per week. (For the following problems, assume that 1 month = 4 weeks.)

a. What is the probability that there are no medication errors in one week?

b. What is the probability that there are no medication errors in one month?

c. Find the average number of medication errors in one week.

d. Find the average number of medication errors in one month.

e. Find the standard deviation of the number of medication errors in one month.

f. How likely is it that at least 4 medication errors will be observed in one month?

10. The number of weaving errors in a 20 ft by 10 ft roll of carpet has a Poisson distribution with $\lambda = 0.1$.

a. Using Appendix A, Table F, construct the probability distribution for the carpet.

b. What is the probability of observing less than 2 errors in the carpet?

c. What is the probability of observing more than 5 errors in the carpet?

11. A bank is evaluating their staffing policy to assure they have sufficient staff for their drive-up window during the lunch hour. If the number of people who arrive at the window in a 15-minute period has a Poisson distribution with $\lambda = 5$, answer the following questions.

a. How many people are expected to arrive during the lunch hour?

b. What is the probability that no one will show up during the lunch hour of 12:00 PM to 1:00 PM?

c. What is the probability that more than 6 people will show up in any 15-minute period?

12. An aluminum foil manufacturer wants to improve the quality of the product and is trying to develop a probability model for the flaws that occur in a sheet of foil. Assume that X, the number of flaws per square foot, has a Poisson distribution. If flaws occur randomly at an average of one flaw per 50 square feet, what is the probability that a box containing a 200 square foot roll will contain one flaw? More than one flaw?

13. A manufacturing company is concerned about the high rate of accidents that occurred on the production line last week. There were 6 accidents in the last week and this may require a report to be sent to the government agency for safety. Determine the probability of 6 accidents occurring in a week when the average number of accidents per week has been 3.5. Assume that the number of accidents per week follows a Poisson distribution.

7.6 The Hypergeometric Distribution

The binomial and the hypergeometric random variables are very similar. Both random variables have only two outcomes on each trial of the experiment. They both count the number of successes in n trials of an experiment. The hypergeometric distribution differs from the binomial distribution in the lack of independence between trials, which also implies that the probability of success will vary between trials. In addition, hypergeometric distributions have finite populations in which the total number of successes and failures are known. Hypergeometric distributions are widely used to model probability in various fields of biology: molecular biology, evolutionary biology, bioinformatics, cancer research, genomics, and malaria research.

Hypergeometric Probability Distribution

The **hypergeometric probability distribution** can be used when sampling from a population of finite size N without replacement and it is known that there are r successes in the population (therefore, $N - r$ failures). The hypergeometric distribution is used to find the probability of x successes in a sample of size n.

DEFINITION

Because the binomial and hypergeometric are closely related, a small change in an experiment can switch the distribution of the random variable. A binomial experiment, such as counting the number of red cards drawn in 8 draws from a deck with replacement, can easily be modified to a hypergeometric by not replacing the cards. Since there are 26 red cards (successes) and 26 black cards (failures), the probability of drawing a red card on the first draw is $\frac{26}{52}$ or $\frac{1}{2}$. If a red card is drawn on the first draw and not replaced, the probability of drawing a red card on the next draw is slightly less $\frac{25}{51}$, since there is one less red card in the deck. If the next card drawn is also red, then the probability of a red card on the third draw will be diminished to $\frac{24}{50}$.

Variance of a Hypergeometric Random Variable

The **variance** of a hypergeometric random variable is given by the expression

$$\sigma^2 = V(X) = n\left(\frac{r}{N}\right)\left(1-\frac{r}{N}\right)\frac{(N-n)}{(N-1)}.$$

FORMULA

Example 7.6.2

Determining the Expected Value and Variance of a Hypergeometric Random Variable

Compute the expected value and variance for the random variable defined in Example 7.6.1.

Solution

$$E(X) = 16\left(\frac{2}{30}\right) \approx 1.067$$

$$V(X) = 16\left(\frac{2}{30}\right)\left(1-\frac{2}{30}\right)\frac{(30-16)}{(30-1)} \approx 0.481$$

Thus, if the experiment were repeated many times, the average number of defective chips per board would be slightly greater than 1.

7.6 Exercises

Basic Concepts

1. How does the hypergeometric model differ from the binomial model?
2. What is the hypergeometric probability distribution function?
3. What are the parameters of the hypergeometric model?
4. How do you determine the expected value of a hypergeometric random variable? The variance?

Exercises

5. Suppose a batch of 50 light bulbs contains 3 light bulbs that are defective. Let $X =$ the number of defective light bulbs in a random sample of 10 light bulbs (where the sample is taken without replacement).
   - **a.** What probability model would be appropriate for describing the number of defective light bulbs in the sample?
   - **b.** Find the expected number of defective bulbs.
   - **c.** Find the standard deviation of the number of defective bulbs.
   - **d.** Find the probability that at least 1 of the bulbs sampled will be defective.
   - **e.** Find the probability that at most 2 of the bulbs sampled will be defective.
   - **f.** Find the probability that more than 3 of the bulbs sampled will be defective.

6. A small electronics firm has 60 employees. Ten of the employees are older than 55. An attorney is investigating a client's claim regarding age discrimination. The attorney randomly selects 15 employees without replacement and records the number of employees over age 55.

- **a.** What probability model would be appropriate for describing the number of employees over age 55 in a sample of 15 selected without replacement?
- **b.** Find the average number of employees over age 55 in the sample.
- **c.** Find the standard deviation of the number of employees over age 55 in the sample.
- **d.** Find the probability that at least 2 of the employees selected will be over age 55.
- **e.** Find the probability that less than 2 of the employees selected will be over age 55.
- **f.** Find the probability that at most 4 of the employees will be over age 55.

7. A bank has to repossess 100 homes. Fifty of the repossessed homes have market values that are less than the outstanding balance of the mortgage. An auditor randomly selects 10 of the repossessed homes (without replacement) and records the number of homes that have market values less than the outstanding balance of the mortgage.

- **a.** Find the expected number of homes the auditor will find with market values less than the outstanding balance of the mortgage. Find the standard deviation of the number of homes the auditor will find with market values less than the outstanding balance of the mortgage.
- **b.** What is the probability that all of the audited homes will have market values in excess of the outstanding balance of the mortgage?
- **c.** What is the probability that none of the audited homes will have market values in excess of the outstanding balance of the mortgage?

8. A small liberal arts college in the Northeast has 200 freshmen. Eighty of the freshmen are engineering majors. Suppose thirty freshmen are randomly selected (without replacement).

- **a.** Find the expected number of engineering majors in the sample.
- **b.** Find the standard deviation of the number of engineering majors in the sample.
- **c.** Find the probability that none of the selected students will be an engineering major.
- **d.** Find the probability that all of the selected students will be an engineering major.

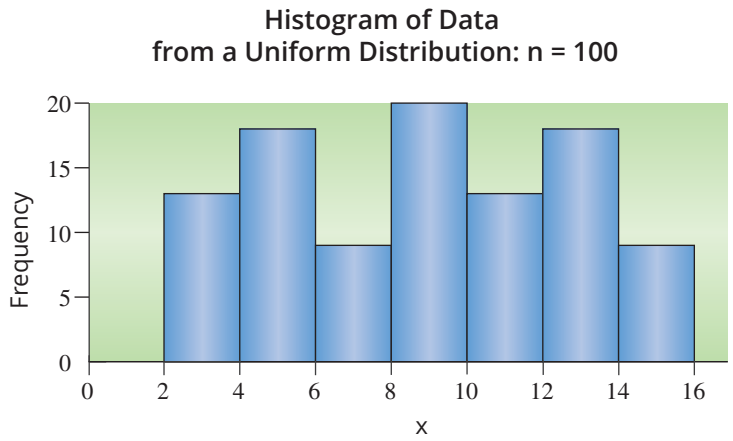


Figure 8.1.2

However, if we were to generate 1000 observations, the distribution would likely begin to level (see Figure 8.1.3).

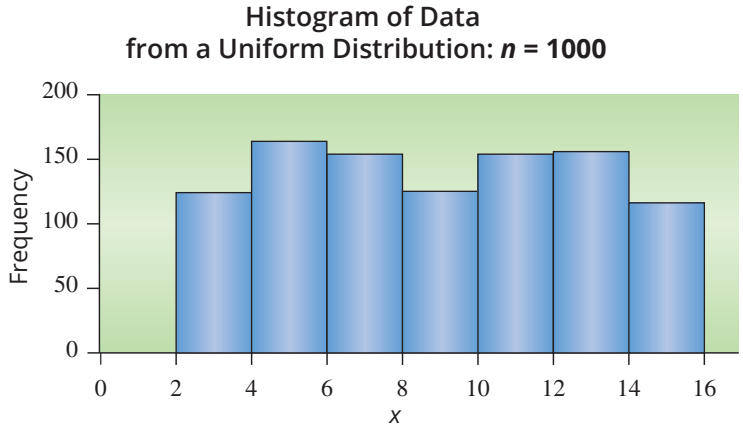


Figure 8.1.3

This idea is similar to the law of large numbers we discussed in Chapter 6 which states that a relative probability approaches the classical probability when enough trials are done. In the case of a random variable's probability distribution, a larger sample size should produce a distributional shape closer to the expected shape.

8.1 Exercises

Basic Concepts

1. Probability is defined differently for discrete and continuous random variables. Describe this difference.
2. How is the continuous uniform distribution different from the discrete uniform distribution?
3. What is the uniform probability density function?
4. Describe the shape of the density function for a uniform distribution.

Exercises

5. Suppose a continuous random variable is uniformly distributed between 10 and 70.

- **a.** What is the mean of the distribution?
- **b.** What is the standard deviation of the distribution?
- **c.** What is the probability that a randomly selected value will be above 45?
- **d.** What is the probability that a randomly selected value will be less than 30?
- **e.** What is the probability that a randomly selected value will be between 25 and 50?
- **f.** Find the probability that a randomly selected value will exactly equal 35.

6. A 14-volt lawn mower battery actually has a voltage that is uniformly distributed between 13.3 and 14.5 volts.

- **a.** What is the mean of the distribution?
- **b.** What is the standard deviation of the distribution?
- **c.** What is the probability that a particular battery has a voltage above 14.0 volts?
- **d.** What is the probability that a particular battery has a voltage less than 13.6 volts?
- **e.** What is the probability that a particular battery has a voltage between 14.0 and 14.3 volts?
- **f.** Find the probability that a particular battery has a voltage that is exactly 14.2 volts.

7. Polar Bear Frozen Foods manufactures frozen French fries for sale to grocery store chains. The final package weight is thought to be a uniformly distributed random variable. Assume X, the weight of French fries, has a uniform distribution between 57 ounces and 63 ounces.

- **a.** What is the mean weight for a package?
- **b.** What is the standard deviation for the weight of a package?
- **c.** What is the probability that a store will receive a package weighing less than 59 ounces?
- **d.** What is the probability that a package will contain between 60 and 63 ounces?
- **e.** What is the probability that a package will contain more than 62 ounces?
- **f.** Find the probability that a package will contain exactly 60 ounces.

8. The annual growth in height of cedar trees is believed to be distributed uniformly between 6 and 11 inches.

- **a.** Draw a picture of the distribution of growth in height of cedar trees.
- **b.** What is the mean growth per year?
- **c.** What is the standard deviation of the growth per year?
- **d.** What is the probability that a randomly selected cedar tree will grow between 9 and 10 inches in a given year?
- **e.** Find the probability that a randomly selected cedar tree will grow less than 8 inches in a given year.

f. Find the probability that a randomly selected cedar tree will grow more than 9 inches in a given year.

g. Find the probability that a randomly selected cedar tree will grow exactly 7 inches in a given year.

9. A particular employee arrives to work sometime between 8:00 AM and 8:30 AM. Based on past experience the company has determined that the employee is equally likely to arrive at any time between 8:00 AM and 8:30 AM.

a. On average, what time does the employee arrive?

b. What is the standard deviation of the time at which the employee arrives?

c. If a call comes in for the employee at 8:10 AM find the probability that the employee will be there to take the call.

d. Find the probability that the employee will arrive between 8:20 AM and 8:25 AM.

e. Find the probability that the employee will arrive after 8:15 AM.

f. Find the probability that the employee will arrive at exactly 8:10 AM.

8.2 The Normal Distribution

Now what if the values of a random variable are not expected to be distributed evenly across a sample space? Specifically, what if the majority of values fall somewhere around the middle of the data range, with fewer values falling towards the ends of the range? The normal distribution is one example of this type of distribution.

To be sure, the normal distribution is the preeminent distribution used in the statistical theory we will examine. Many statistical inference procedures either directly or indirectly have roots in normal theory. These procedures usually assume that the population from which a random sample is drawn is normally distributed.

The **normal distribution**, also called the Gaussian distribution, was named after Carl Gauss who published a work in 1823 describing the mathematical definition of the distribution.[1] Gauss developed this distribution to describe the error in predicting the orbits of planets.

Normal distributions are all bell-shaped, but the bells come in various shapes and sizes. Since all normal distributions are symmetric, the mean, mode, and median are all equal.

Although normally distributed random variables can range in value from minus infinity to positive infinity, values that are a great distance from the mean rarely occur.

The Origins of the Normal Distribution: Abraham de Moivre (1667 – 1754)

De Moivre was born in France but lived most of his life in England. In 1733, he published a paper containing the equation that describes the normal curve. He allegedly was doing calculations using the binomial distribution for gamblers and was looking for a shortcut in the very arduous calculations. He discovered the normal distribution as the limit of the binomial distribution. De Moivre was a highly respected mathematician and friend of Issac Newton. He lived to a quite old age for the time and died in obscure poverty.

De Moivre's discovery received little attention until Laplace began writing on probability in the 1770s.

There are two other mathematicians who discovered the equation of the normal curve, Adrain in 1808 and Gauss in 1809. Even though de Moivre published the equation for the normal distribution more than 75 years earlier than Gauss, the normal curve was called the Gaussian distribution for many years. Even now, you will hear the normal curve referred to as the Gaussian distribution.

Using the probability density to determine the probability of some interval would be complicated. Fortunately, there is an easier way. A special normal distribution, called the **standard normal**, can be used to determine probabilities for any normal random variable. The standard normal distribution will be discussed in Section 8.3.

8.2 Exercises

Basic Concepts

1. How was the normal distribution developed?
2. Are the normal and uniform distributions probability models?
3. List the properties of the normal distribution.
4. What is the shape of the normal distribution?
5. What are the parameters of the normal distribution?
6. If the variance of a normal distribution is constant, what effect will changes in the mean have on the distribution?
7. If the mean of a normal distribution is constant, what effect will changes in the standard deviation have on the distribution?

Exercises

8. Sketch a normal curve and mark each of the following on the x-axis.
   a. μ
   b. $\mu + \sigma$
   c. $\mu - \sigma$

9. Sketch a normal curve and use labels to illustrate the empirical rule.

10. Sketch three normal curves on a single axis that have the same standard deviation but different means.

11. Sketch three normal curves on a single axis that have the same mean, but different standard deviations.

12. Using the Human Development Trends data set and technology, create a histogram for the following variables in the data set using 10 classes (or bins in Excel). Does the distribution of each variable appear to be normally distributed? Why or why not?
    a. Human Development Index (HDI)
    b. Life expectancy

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Human Development Trends

Technology

For instructions on how to create a histogram, visit stat.hawkeslearning.com and navigate to **Discovering Statistics and Data, Fourth Edition > Technology Instructions > Graphs > Histogram**.

13. Using the following data on the length in feet of 40 great white sharks, create a stem-and-leaf plot of the data using a two-digit stem. Do the data appear to be normally distributed?

11.9	16.6	14.0	10.8	17.7	13.7	16.1	17.3
15.8	18.7	13.1	14.9	17.0	13.3	15.5	19.7
12.9	16.2	17.3	18.4	14.8	14.8	10.7	14.0
19.4	14.8	14.9	13.4	19.7	15.1	14.0	12.2
16.4	11.5	15.6	19.5	19.0	16.1	17.9	15.4

8.3 The Standard Normal Distribution

Given that the normal distribution is a function of two continuous parameters, μ and σ, there are an infinite number of combinations for μ and σ, and thus an infinite number of normal distributions. The **standard normal distribution** in Figure 8.3.1 is a special version of the normal distribution. This distribution provides a basis for computing probabilities for all normal distributions.

Standard Normal Distribution

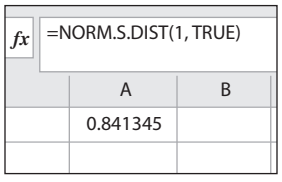
The **standard normal distribution** is a normal distribution with a mean of zero and a standard deviation of one.

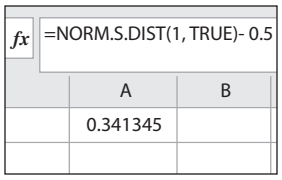
$$\mu = 0 \quad \text{and} \quad \sigma = 1$$

DEFINITION

Technology

A normal probability can also be easily found using technology. For full instructions on computing normal probabilities using technology, visit stat.hawkeslearning.com and navigate to **Discovering Statistics and Data, Fourth Edition > Technology Instructions > Normal Distribution > Normal Probability (cdf)**.





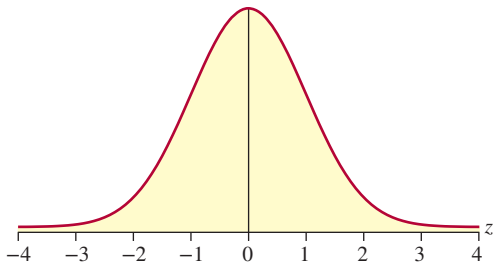


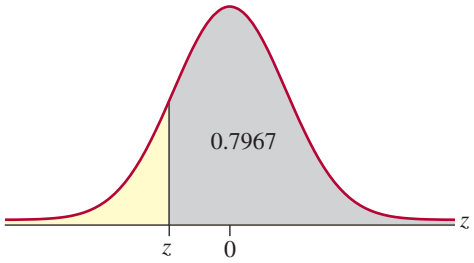
Figure 8.3.1

The technique used to translate any normal random variable into a standard normal random variable is called a ***z*-transformation** (or "standardizing" the random variable) and was discussed earlier in Chapter 4. Because the z-transformation gives the standard normal unique status among normals, the standard normal is also referred to as the ***z*-distribution**.

Tables A, B, and C in Appendix A contain probability calculations for various areas under the standard normal curve. Specifically, Tables A and B provide the probability that a standard normal random variable will be less than a specified value. Table C provides the probability that a standard normal random variable will be between 0 and a specified value. For example, to compute the probability that a standard normal random variable will be less than 1 (see Figure 8.3.2), look up the value 1.00 in Table B. The table value of 0.8413 is the area under the curve between negative infinity and 1, which is also the probability that the random variable will assume a value in that interval.

Please note that the value of z is to the left of 0. Thus, the value of z is going to be negative. Note that the area to the left of z represents the cumulative probability (in the above figure). So, to find the value of z, we only need to find 0.1469 in the body of Appendix A, Table A. The value of z with the area 0.1469 to the left of it is -1.05. That is, $P(z < -1.05) = 0.1469$.

d. Once again, a picture can be very helpful.



Note that from the picture, we have the area to the right of z. However, we know that the total area under the curve is 1. Thus, if the area to the right of z is 0.7967, then the area to the left of z is $1 - 0.7967 = 0.2033$. From the picture, it is clear that if we find 0.2033 in the body of Appendix A, Table A, the corresponding value of z is the value we are interested in. This value of z is -0.83. Therefore, the value of z with the area 0.7967 to the right is -0.83.

8.3 Exercises

Basic Concepts

1. When we say a random variable has a distribution, what does that mean?
2. Why is the standard normal distribution called the standard normal?
3. Would it be unusual for a standard normal distribution to have an observation greater than 6?
4. Would it be unusual for a standard normal distribution to have an observation less than -4?
5. What is the standard normal distribution? What are the parameters of the distribution?
6. Why is the standard normal distribution important?

Exercises

7. What proportion of the area under the standard normal curve falls between the following z-values?
   - **a.** 0 and 0.67
   - **b.** 0 and 1.645
   - **c.** 0 and 1.96
   - **d.** 0 and 2.575

8. What proportion of the area under the standard normal curve falls between the following z-values?

a. -0.67 and 0

b. -1.645 and 0

c. -1.96 and 0

d. -2.575 and 0

9. What proportion of the area under the standard normal curve falls between the following z-values?

a. -0.85 and 0.85

b. -0.55 and 0.55

c. -1.56 and 1.98

d. -2.23 and 2.96

10. What proportion of the area under the standard normal curve falls between the following z-values?

a. -0.97 and 0.97

b. -0.54 and 1.82

c. -1.95 and 2.28

d. -2.89 and 1.59

11. Using the standard normal tables in Appendix A, determine the following probabilities and draw the corresponding diagram..

a. $z \le 0$

b. $z \ge 0$

c. $z \le -1$

d. $z \le 1$

e. $z \ge -1$

f. $z \ge 1$

12. Using the standard normal tables in Appendix A, determine the following probabilities and draw the corresponding diagram.

a. $z \le -0.44$

b. $z \ge 0.44$

c. $-0.44 \le z \le 0.44$

d. $z \le -0.67$

e. $z \ge 0.67$

f. $-0.67 \le z \le 0.67$

13. Using the standard normal tables in Appendix A, determine the following probabilities and draw the corresponding diagram.

a. $z \le -1.28$

b. $z \ge 1.28$

c. $-1.28 \le z \le 1.28$

d. $z \le -1.96$

e. $z \ge 1.96$

f. $-1.96 \le z \le 1.96$

14. Using the standard normal tables in Appendix A, determine the following probabilities and draw the corresponding diagram.

a. $P(0 \le z \le 0.79)$

b. $P(-1.57 \le z \le 2.33)$

c. $P(z \ge 1.89)$

d. $P(z \le -2.77)$

15. Using the standard normal tables in Appendix A, determine the following probabilities and draw the corresponding diagram.

a. $P(0 \le z \le 1.24)$

b. $P(-2.64 \le z \le 3.32)$

c. $P(z \ge 3.22)$

d. $P(z \le -3.39)$

Note

We recommend drawing the corresponding areas on a normal curve when working on these exercises.

Note

We recommend drawing the corresponding areas on a normal curve when working on these exercises.

16. Find the value of z such that 0.05 of the area under the curve lies to the right of z and draw the corresponding diagram.

17. Find the value of z such that 0.01 of the area under the curve lies to the right of z and draw the corresponding diagram.

18. Find the value of z such that 0.10 of the area under the curve lies to the right of z.

19. Find the value of z such that 0.05 of the area under the curve lies to the left of z and draw the corresponding diagram.

20. Find the value of z such that 0.01 of the area under the curve lies to the left of z.

21. Find the value of z such that 0.10 of the area under the curve lies to the left of z.

22. Find the value of z such that 0.7458 of the area under the curve lies between $-z$ and z and draw the corresponding diagram.

23. Find the value of z such that 0.9505 of the area under the curve lies between $-z$ and z.

24. Find the value of z such that 0.90 of the area under the curve lies between $-z$ and z.

8.4 Applications of the Normal Distribution

Most normal distributions of real data do not have a mean of zero and standard deviation of one. However, we can perform a transformation to standardize any normal random variable into a standard normal distribution.

Standardizing a Normal Random Variable

The following formula can transform any normal random variable into a standard normal random variable, z:

$$z = \frac{x - \mu}{\sigma}$$

where x is a normal random variable with mean μ and standard deviation σ.

FORMULA

If we look at the individual pieces, exactly how the transformation works is not very mysterious. First, the numerator $x - \mu$ centers the z-distribution around zero. By subtracting the mean of the random variable from each data value, the mean of the resulting random variable will be zero. A short example illustrates the point.

8.4 Exercises

Basic Concepts

1. Describe the connection between the z-transformation and the standard normal random variable.

Exercises

2. Determine the z-value given $\mu = 15$, $\sigma = 2$, and $x = 19$. Indicate where the z-value would be on the standard normal distribution.

3. Determine the z-value given $\mu = 0.023$, $\sigma = 0.001$, and $x = 0.020$. Indicate where the z-value would be on the standard normal distribution.

4. The lengths of full-term pregnancies have a normal distribution with a mean of 266 days and a standard deviation of 16 days.

- **a.** Find the probability that the length of a pregnancy is between 250 and 282 days.
- **b.** Find the probability that the length of a pregnancy is greater than 275 days.
- **c.** Find the probability that the length of a pregnancy is less than 250 days.

5. If too much glucose binds to hemoglobin in the blood, this is an indicator of diabetes. A test called A1c measures the amount of glucose that binds to hemoglobin. A1c level in a nondiabetic person has a normal distribution with a mean of 5% and a standard deviation of 0.5%.

- **a.** Find the probability that the A1c level is between 5.5% and 5.7%.
- **b.** Find the probability that the A1c level is greater than 5.7%.
- **c.** Find the probability that the A1c level is less than 5.2%.

6. According to the Bureau of Labor Statistics, the average data scientist salary is \$100,560.[3] If the standard deviation of data scientist incomes is \$25,000 and we assume that salaries are normally distributed, what percentage of data scientists earn more than \$150,000?

7. A certain component for the newly developed electronic diesel engine is considered to be defective if its diameter is less than 8.0 mm or greater than 10.5 mm. The distribution of the diameters of these parts is known to be normal with a mean of 9.0 mm and a standard deviation of 1.5 mm. If a component is randomly selected, what is the probability that it will be defective?

8. A television manufacturer is studying television remote control unit usage. One of the criteria they are measuring is the distance at which people attempt to activate the television set with the remote unit. They have discovered that activation distances are normally distributed with an average activation distance of six feet with a standard deviation of three feet. If a remote unit's maximum range is ten feet, what proportion of the time will users attempt to operate the remote outside of the operating limit?

9. According to the Bureau of Labor Statistics, the mean weekly earnings for people working in a sales related profession in 2022 was \$938.[4] Assume that the weekly earnings are approximately normally distributed with a standard deviation of \$100.

a. What are the mean weekly earnings for people working in a sales related profession in 2022?

b. If a salesperson was randomly selected, find the probability that their weekly earnings exceed \$1000.

c. If a salesperson was randomly selected, find the probability that their weekly earnings are at most \$800.

d. If a salesperson was randomly selected, find the probability that their weekly earnings are between \$800 and \$950.

e. Do you feel that it is reasonable to assume that the weekly earnings have a normal distribution? Why or why not?

10. The repair time for air conditioning units is believed to have a normal distribution with a mean of 38 minutes.

a. What is the standard deviation of repair time if 40% of the units are repaired between 33 and 43 minutes?

b. Using the value of the standard deviation that you computed in **a.**, what is the probability that a repair will be longer than an hour?

c. Using the value of the standard deviation that you computed in **a.**, what is the probability that the repair time for an air conditioning unit will be less than 25 minutes?

11. LED monitors manufactured by TSI Electronics have life spans which have a normal distribution with an average life span of 100,000 hours and a standard deviation of 15,000 hours. If an LED monitor is selected at random, find the following probabilities.

a. The probability that the life span of the monitor will be less than 90,000 hours.

b. The probability that the life span of the monitor will be more than 120,000 hours.

c. The probability that the life span of the monitor will be between 100,000 hours and 120,000 hours.

12. A beer distributor believes the amount of beer in a 12-ounce can of beer has a normal distribution with a mean of 12 ounces and a standard deviation of 1 ounce. If a 12-ounce beer can is randomly selected, find the following probabilities.

a. The probability that the 12-ounce can of beer will actually contain less than 11 ounces of beer.

b. The probability that the 12-ounce can of beer will actually contain more than 12.5 ounces of beer.

c. The probability that the 12-ounce can of beer will actually contain between 10.5 and 11.5 ounces of beer.

13. A statistics teacher believes that the final exam grades for her business statistics class have a normal distribution with a mean of 82 and a standard deviation of 8.

a. Find the score which separates the top 10% of the scores from the lowest 90% of the scores.

b. The teacher plans to give all students who score in the top 10% of scores an A. Will a student who scored a 90 on the exam receive an A? Explain.

c. Find the score which separates the lowest 20% of the scores from the highest 80% of the scores.

d. The teacher plans to give all students who score in the lowest 10% of scores an F. Will a student who scored a 65 on the exam receive an F? Explain.

14. In order for you to become a member of Mensa, a worldwide organization with approximately 145,000 members, your IQ score must be in the top 2%. The word *mensa* is Latin for "table," and was chosen to denote a group or round table of people with equal ability. In 1996, Mensa, which was founded by two British barristers, celebrated its 50^{th} birthday. American Mensa Ltd., which was founded in 1960 has more than 57,000 members. Assuming that IQ scores have an approximately normal distribution with a mean and standard deviation of 100 and 15, respectively, answer the following questions.

a. What IQ must one have in order to become a member of Mensa?

b. What percent of all Americans have an IQ of at least 145?

c. What percent of all members of Mensa have an IQ of at least 145?

d. If Mensa decided to become more exclusive, and accepted only the top 1% instead of the top 2% as members, what IQ would one need in order to become a member of Mensa?

15. A farmer believes that the yields of his tomato plants have a normal distribution with an average yield of 10 lb. and a standard deviation of 2 lb. The farmer would like to identify the plants which yield the highest 5% and save them for breeding purposes.

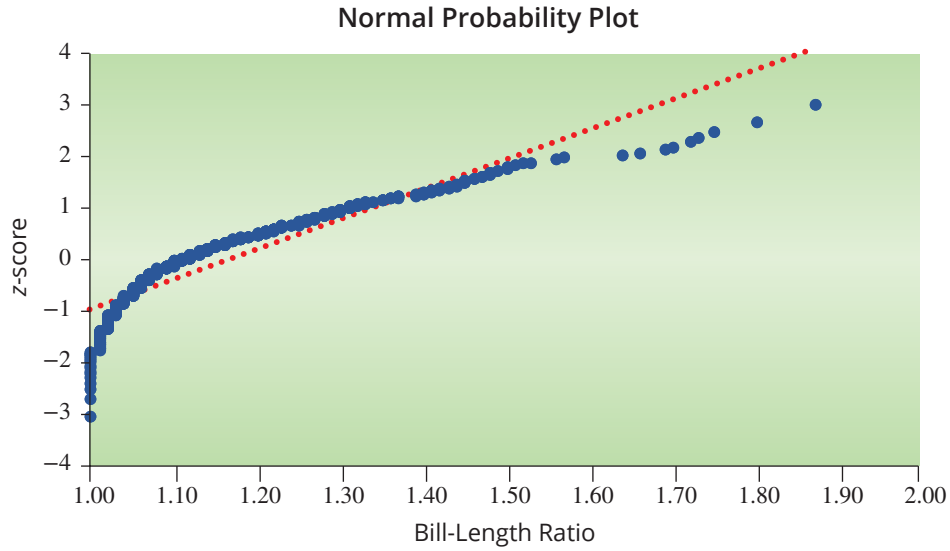
a. Compute the yield which separates the highest 5% of yields from the lowest 95% of yields.

b. If a tomato plant yielded 14 lb. would it be kept for breeding purposes? Explain.

c. If a tomato plant yielded 13 lb. would it be kept for breeding purposes? Explain.

8.5 Assessing Normality

Many of the statistical tests that are discussed in this book require that the data be a simple random sample from a population that has a *normal* distribution, or is at least approximately normal. If a histogram of the data is symmetric and bell-shaped, we can assume normality. However, the shape of a histogram can be hard to determine with a small sample of data. Therefore, we need additional ways to assess normality. One of these alternative methods is called a **normal probability plot** (or **normal quantile plot**).



From the graph, we can see that the plotted points do not follow a linear trend. This particular pattern indicates that the data are skewed right and not normally distributed, confirming what we observed in the histogram.

Evaluating the normality of a data set is crucial in parametric statistics. We'll revisit this subject when we delve into different statistical tests and models in upcoming chapters.

8.5 Exercises

Basic Concepts

1. List two ways to graphically assess the normality of a data set. Under what conditions are each appropriate?
2. Describe the general procedure for creating a normal probability plot.
3. How should a normal probability plot look to indicate normality?

Exercises

Please use technology for all the exercises in this section.

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Moneyball

4. Construct a histogram using the "BA" (batting average) column of the Moneyball data set. Can we assume batting averages have a normal distribution?

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Housefly Wing Lengths

5. Create a normal probability plot of the housefly data from Example 8.5.2. What do you observe? Does the plot lead you to the same conclusion as the histogram?

6. A pharmaceutical company wants to test whether a new cold medication will perform better than an existing medication. Laboratory technicians observe a sample of 25 patients and record the number of hours it takes for each patient to feel symptom relief after taking the medicine. Before the company performs a

test of the new medication against the current one, they need to know if the data are normally distributed. Use a normal probability plot to determine if the data appear to come from a population that is normally distributed.

3.00	1.50	0.20	1.62	1.06
3.01	2.45	0.66	1.94	0.21
1.51	3.08	5.37	6.96	1.32
0.79	7.20	1.36	4.45	3.29
1.74	3.87	1.90	3.50	3.09

7. Data on the total annual rainfall (in inches) in Asheville, North Carolina were gathered by a weather station in from 2000-2022.[7] Use a normal probability plot to determine if the data appears to come from a population that is normally distributed.

Total Annual Rainfall in Asheville, NC					
Year	Total Precipitation (in inches)	Year	Total Precipitation (in inches)	Year	Total Precipitation (in inches)
2000	35.59	2008	35.63	2016	33.40
2001	34.49	2009	62.13	2017	54.10
2002	44.47	2010	44.26	2018	79.48
2003	59.46	2011	46.04	2019	57.25
2004	52.36	2012	44.66	2020	64.71
2005	47.26	2013	75.22	2021	54.51
2006	48.29	2014	46.91	2022	45.43
2007	34.39	2015	54.35		

8. A professor is interested in examining the distribution of the grades his students received on the midterm exam. There are 18 students in the class, and no time limit was given for the exam. Use a normal probability plot to determine if the students' grades are normally distributed.

80.8	81.7	81.7	81.7	81.7	82.5
83.3	83.3	84.2	84.2	85.0	86.7
86.7	87.5	87.5	90.3	90.4	90.8

9. A group of students and professors are studying conifers in the Pacific Northwest United States. They take a sample of 25 Douglas fir trees and record several metrics, including the circumference of the trunks (in meters).[8] Use a normal probability plot to determine if the trunk circumference values are normally distributed.

4.97	0.45	0.40	0.15	2.84
6.65	0.62	0.39	0.86	1.24
4.93	0.64	0.62	2.22	2.23
0.29	0.18	0.27	1.97	2.45
0.19	0.55	0.41	2.85	9.09

10. A group of friends decide to run a marathon together. There are 16 runners in the group, and they are all in relatively good shape. Use a normal probability plot to determine if their marathon times are normally distributed.

4:07:58	4:18:34	4:21:15	4:24:23
4:08:07	4:18:40	4:22:17	4:25:12
4:16:28	4:19:39	4:23:52	4:25:14
4:17:30	4:19:45	4:23:55	4:26:34

8.6 Approximation to the Binomial Distribution

To approximate other distributions, the normal distribution can be very useful. Although it is a continuous distribution, it is used to approximate discrete distributions, specifically the binomial.

The Binomial Distribution

Calculating binomial probabilities can be quite time consuming if n is large. For example, suppose that you intend to sample 2000 subjects for a marketing research survey. If 50 percent of the population believes your product is superior to the competition's, what is the probability of obtaining 600 or fewer subjects who believe your company's product is superior?

$$P(X \le 600) = P(X = 0) + P(X = 1) + P(X = 2) + \cdots + P(X = 599) + P(X = 600)$$

Determining the appropriate probability using the binomial distribution would require the calculation of 601 individual probabilities, many of which would have extremely large combinations such as the following.

$${}_{2000}C_{400}\, 0.5^{400}(1 - 0.5)^{1600}$$

Computing this and the other 600 similar calculations would be a formidable task. The normal distribution is useful in approximating binomial probabilities. The larger the binomial parameter, n, the more accurate the approximation. Determining the probability described above using the normal approximation is trivial in comparison to calculating the exact probability using the binomial.

Recall that the normal distribution is a function of two parameters, the mean and the standard deviation. Thus, if the normal distribution is used to approximate the binomial distribution, it seems reasonable that the mean and standard deviation of the normal should be the same as the mean and standard deviation of the binomial that is being approximated. Specifically, let

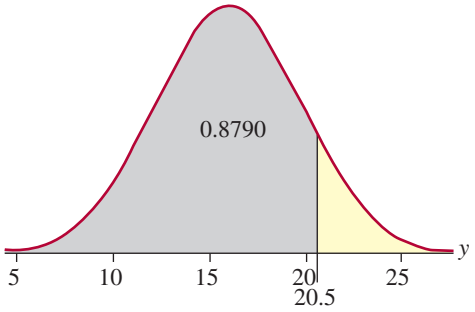
$$\mu = E(X) = np, \text{ and}$$

$$\sigma = \sqrt{V(X)} = \sqrt{np(1-p)}.$$

Using continuity correction,

$$P(Y \le 20.5) = P\left(z \le \frac{20.5 - 16}{3.8367}\right) \approx P(z \le 1.17) = 0.8790.$$

Normal Approximation to the Binomial, $n = 200$, $p = 0.08$



Thus, using the normal approximation and continuity correction, the probability that the restaurant will have no more than 20 no-shows is 0.8790. Notice that the continuity correction has a significant impact on the accuracy of the approximation. Using the binomial distribution, the exact probability is 0.8775.

Technology

For directions on computing a binomial probability using technology, please visit stat.hawkeslearning.com and navigate to **Discovering Statistics and Data, Fourth Edition > Technology Instructions > Binomial Distribution > Binomial Probability (cdf)**.

8.6 Exercises

Basic Concepts

1. Why would you want to use the normal distribution to approximate a binomial distribution?
2. What are the parameters of a normal distribution used to approximate a binomial distribution?
3. What is continuity correction? How does it improve the normal approximation to the binomial?

Exercises

4. The pandemic has closed the generational technology divide as more older adults have adopted technology. A recent AARP survey revealed that 44% of Americans 50 years old and older enjoy playing video games at least once a month.[9] Consider the probability that fewer than 15 out of the 123 seniors surveyed at a local shopping mall play video games Assume that the probability of a given senior playing video games is 44%. Verify that a normal distribution can be used to approximate the binomial probability, or show how the conditions have not been met.

5. Only 4% of Americans are truly vegetarian, including vegans.[10] For a survey of what vegetarian foods to include in a new product line at Whole Foods grocery stores, would a sample of 100 customers be sufficient to use a normal distribution to approximate probabilities for this survey? Verify that a normal distribution can be used to approximate a binomial probability for the survey or show how the conditions have not been met.

6. Management at a small engineering company is considering the addition of a company cafeteria area. A random sample of 50 persons out of the total number of persons employed by the firm will be surveyed to see if they are in favor of the addition. Assume that the true percentage of persons that favor the addition is 90%.

a. Find the expected number of employees in the sample who will favor the addition of the cafeteria area.

b. Find the standard deviation of the number of employees in the sample who will favor the addition of the cafeteria area.

c. What is the probability that between 35 and 37 employees (inclusive) in the sample will favor the cafeteria?

d. What is the probability that more than 40 of the employees in the sample will favor the cafeteria?

e. What is the probability that at most 38 of the employees in the sample will favor the cafeteria?

7. The accounting department of a large corporation checks the addition of expense reports submitted by executives before paying them. Historically, they have found that 15% of the reports contain addition errors. An auditor randomly selects 60 expense reports and audits them for addition errors.

a. Find the expected number of reports in the sample that will have addition errors.

b. Find the standard deviation of the number of reports sampled that will have addition errors.

c. Find the probability that fewer than 10 of the sampled expense reports will have addition errors.

d. Find the probability that at least 30 of the sampled expense reports will have addition errors.

e. Find the probability that between 5 and 15 (inclusive) of the sampled expense reports will have addition errors.

8. A local electronics store purchased a market research study which suggests that 40 percent of all homes have a video doorbell. A sample of 200 homes is selected to confirm the study's findings. If the marketing study is correct, answer the following questions.

a. Find the expected number of homes sampled which will have video doorbells.

b. Find the standard deviation of the number of homes in the sample which will have video doorbells.

c. What is the probability that at most 80 of the sampled homes will have video doorbells?

d. What is the probability that between 100 and 120 (inclusive) homes sampled will have video doorbells?

e. What is the probability that at least 130 of the sampled homes will have video doorbells?

9. Suppose a virus is believed to infect two percent of the population. If a sample of 3000 randomly selected subjects are tested, answer the following questions.

a. Find the expected number of subjects sampled that will be infected.

b. Find the standard deviation of the number of subjects sampled that will be infected.

c. What is the probability that fewer than 30 of the subjects in the sample will be infected?

d. What is the probability that between 40 and 80 (inclusive) of the subjects in the sample will be infected?

e. Find the probability that at least 70 of the subjects in the sample will be infected.

10. Based on a recent survey, approximately 71% of Americans who shop at Walmart purchase the store brands. A random sample of 200 shoppers at Walmart was conducted. If the survey is correct, answer the following questions.

a. Find the expected number of people that will purchase a Walmart store brand from the random sample of 200 Walmart shoppers.

b. Find the standard deviation of the number of people that will purchase a Walmart store brand from the sample of Walmart shoppers.

c. Find the probability that more than 150 Walmart shoppers will purchase a store brand.

d. Find the probability that less than 100 Walmart shoppers will purchase a store brand.

e. Find the probability that between 120 and 150 shoppers (inclusive) will purchase a store brand.

CR Chapter Review

Key Terms and Ideas

- Continuous Random Variables
- Continuous Uniform Distribution
- Probability Density Function
- Uniform Probability Density Function
- Normal Distribution
- Normal Probability Density Function
- Standard Normal Distribution
- z-Distribution
- z-Score
- Standard Normal Random Variable
- Normal Probability Plot
- Normal Approximation to the Binomial Distribution
- Continuity Correction

Key Formulas

Section	
8.1	**Uniform Probability Density Function** $f(x)=\begin{cases}\frac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{otherwise}\end{cases}$ **Expected Value for a Continuous Uniform Random Variable** $\mu = E(x) = \frac{a+b}{2}$ **Standard Deviation for a Continuous Uniform Random Variable** $\sigma = \frac{b-a}{\sqrt{12}}$

Point Estimator

A **point estimator** is a single-valued estimate calculated from the sample data, which is intended to be close to the true population value.

DEFINITION

Can you be sure that the sample mean will always be close to the population mean? When dealing with random variables, nothing is certain, but there are methods of reducing the probable error. To understand how this is achieved, we must examine how the sample mean varies.

9.1 Exercises

Basic Concepts

1. Is sampling part of a deductive or an inductive process?
2. Why is the quality of sample data so important?
3. Why is randomness useful in sampling?
4. What is the problem with a voluntary sample?
5. What makes a sample biased?
6. What is a sampling frame and why is it important?
7. Discuss how you would draw a simple random sample of the people in your town.
8. True or False: Determining a well-defined population is easy when developing a sampling frame.
9. What is a pseudo-random number generator? How is it not completely?
10. Is the sample mean always close to the population mean?
11. What is the sampling distribution of the sample mean?
12. Why is the sample mean a random variable?
13. What is a point estimator?
14. Explain the meaning of the "distribution of a variable" in the context of statistics.
15. Explain how a statistic can have a distribution.

Exercises

Data

stat.hawkeslearning.com
Data Sets > Beers and Breweries

16. Obtain a random sample of 15 beers from the Beers and Breweries data set. Describe how you selected the sample.

17. Obtain a random sample of 12 houses (list the IDs) from the Dunes West subdivision in the Mount Pleasant Real Estate data set. Describe how you selected the sample.

Data

stat.hawkeslearning.com

Data Sets > Mount Pleasant Real Estate Data

18. A magazine reported the results of a survey in which readers were asked to send in their responses to several questions regarding good eating. Consider the reported results to the question, *How often do you eat chocolate?*

Survey Responses	
Category	**% of Responses**
Frequently	13
Occasionally	45
Seldom	37
Never	5

a. Were the responses to this survey obtained using voluntary sampling techniques? Explain your answer.

b. What types of biases may be present in the responses?

c. Is 13% a reasonable estimate of the proportion of all Americans who eat chocolate frequently? Explain.

19. A magazine reported the results of a survey in which readers were asked to send in their responses to several questions regarding anger. Consider the reported results to the question, *How long do you usually stay angry?*

Survey Responses	
Category	**% of Responses**
A few hours or less	48
A day	12
Several days	9
A month	1
I hold a grudge indefinitely	22
It depends on the situation	8

a. Were the responses to this survey obtained using voluntary sampling techniques? Explain your answer.

b. What types of biases may be present in the responses?

c. Is 22% a reasonable estimate of the proportion of all Americans who hold a grudge indefinitely? Explain.

20. Students in a marketing class have been asked to conduct a survey to determine whether or not there is a demand for an insurance program at a local college. The students decide to randomly select students from the local college and mail them a questionnaire regarding the insurance program. Of the 150 surveys that were mailed, 50 students responded to the following survey item: *Pick the category which best describes your interest in an insurance program.*

Survey Responses	
Category	**% of Responses**
Very Interested	50
Somewhat Interested	15
Interested	10
Not Very Interested	5
Not At All Interested	20

a. What types of biases may be present in the responses?

b. Is 50% a reasonable estimate of the proportion of all students who would be very interested in an insurance program at the local college? Explain.

c. Is 50% a reasonable estimate of the proportion of all business majors who would be very interested in an insurance program at the local college? Explain.

d. What strategies do you think the marketing students could have used to get a less biased response to their survey?

e. Suppose the program was created and only a few people registered. How could the survey question have been reworded to better predict actual enrollment?

21. Shortly after acquiring Twitter, Elon Musk created a poll asking: *Should I step down as head of Twitter? I will abide by the results of this poll.* The results of the particular survey were 57.5% yes and 42.5% no.

a. How would you describe this sampling methodology?

b. What biases may exist in this sampling method?

c. Is it reasonable to believe that the results of the survey reflect the attitudes of the Twitter users on this issue?

9.2 Exercises

Basic Concepts

1. What is an estimator? Give an example.
2. What three questions should be asked when considering a random variable?
3. Explain the difference between a biased estimator and an unbiased estimator.
4. Give two examples of estimators that are unbiased.
5. Is an unbiased estimator always closer to the parameter being estimated than a biased estimator? Explain.
6. What is the standard error of the mean and what does it indicate?
7. What are two desirable characteristics of the sample mean?
8. Explain the Central Limit Theorem.
9. What effect does increasing the sample size have on the accuracy of an estimate?

Exercises

10. Suppose that the length of babies born in a local hospital has a mean of 20 inches and a standard deviation of 1 inch. Calculate the mean and the standard deviation of the sample mean for each of the following sample sizes. (Assume the population is infinite.)

a. $n = 35$

b. $n = 50$

c. $n = 75$

d. What happens to the size of the standard deviation of the sample mean as the sample size increases?

11. Suppose that the average ages of employees at the credit union where you bank has a mean of 50 years and a standard deviation of 10 years. Calculate the mean and standard error for each of the following sample sizes (assume the population is infinite).

a. $n = 40$

b. $n = 55$

c. $n = 100$

d. What happens to the size of the standard error as the sample size increases?

12. Suppose that the average time teenagers spend on social media per day is a normally distributed population with a mean of 8 hours (480 minutes) and a standard deviation of 2 hours (120 minutes). If $\bar{x}$ is the average (in minutes) of a sample of 50, find the following probabilities.

a. $P(\bar{x} \le 500)$

b. $P(\bar{x} \ge 450)$

c. $P(423 \le \bar{x} \le 514)$

d. $P(400 \le \bar{x} \le 496)$

13. A company fills bags with fertilizer for retail sale. The weights of the bags of fertilizer have a normal distribution with a mean weight of 15 lb and standard deviation of 1.70 lb.

a. What is the probability that a randomly selected bag of fertilizer will weigh between 14 and 16 pounds?

b. If 35 bags of fertilizer are randomly selected, find the probability that the average weight of the 35 bags will be between 14 and 16 pounds.

14. A travel agency conducted a survey of the prices charged by ocean cruise ship lines and determined they were approximately normally distributed with a mean of $210 per day (base ticket price and onboard spending) and a standard deviation of $20 per day.

a. If an ocean cruise ship line is chosen at random, find the probability that it will charge less than $175 per day.

b. What is the probability that the average charge for a randomly selected sample of 35 ocean cruise ship lines will be less than $175 per day?

15. The turkeys found in a particular county have an average weight of 15.6 pounds with a standard deviation of 4.00 pounds. Forty-five turkeys are randomly selected for a county fair.

a. Find the probability that the average weight of the turkeys will be less than 14.5 pounds.

b. What is the probability that the average weight of the turkeys will be more than 17 pounds?

c. Find the probability that the average weight of the turkeys will be between 13 and 18 pounds.

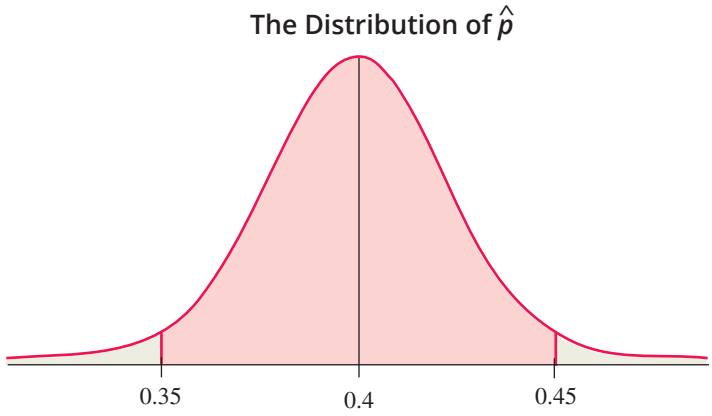
16. The average score for a water safety instructor (WSI) exam is 75 with a standard deviation of 12. Fifty scores for the WSI exam are randomly selected.

a. Find the probability that the average of the fifty scores is at least 80.

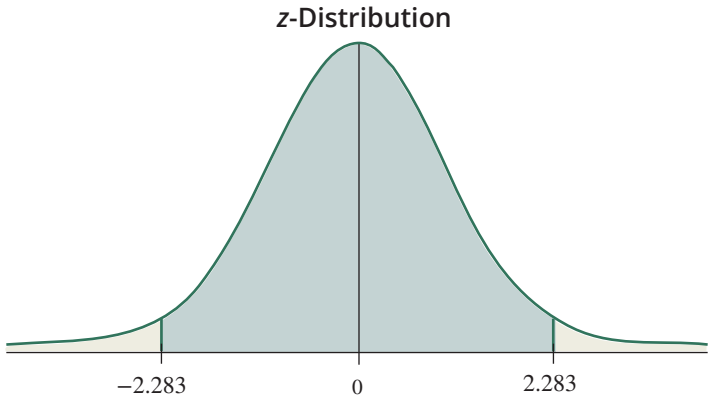
b. Find the probability that the average of the fifty scores is at most 70.

c. Find the probability that the average of the fifty scores is between 72 and 78.

17. A college food service buys frozen fish in boxes labeled 10 pounds. The true average weight of the boxes is 8 pounds with a standard deviation of 2 pounds. The food service director suspects that the boxes do not contain as much fish as advertised. He decides to inspect 40 boxes from the next shipment. If the average weight is less than 10 pounds he will reject the entire shipment. Find the probability that the food service director will not reject the shipment.



Using the z-transformation,



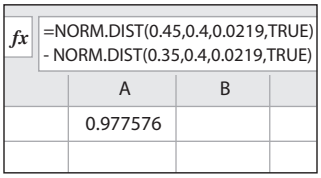
$$P\left(\frac{(0.35-0.4)}{0.0219} < z < \frac{(0.45-0.4)}{0.0219}\right)$$

$$= P(-2.283 < z < 2.283)$$
$$= P(z < 2.283) - P(z < -2.283)$$
$$= 0.9888 - 0.0112$$
$$= 0.9776$$

For a sample of 500, it is very probable (0.9776) that the error of estimation will be less than 0.05.

Technology

For directions on calculating the probability, please visit stat.hawkeslearning.com and navigate to **Discovering Statistics and Data, Fourth Edition > Technology Instructions > Normal Distribution > Normal Probability (cdf)**.



9.3 Exercises

Basic Concepts

1. What does the symbol $\hat{p}$ represent?
2. What is the connection between $\hat{p}$ and p?
3. Is $\hat{p}$ an unbiased estimator? If so, of what?

4. What are the conditions that make the sample size n "sufficiently large" for a sample proportion?

5. Describe the sampling distribution of $\hat{p}$ if n is sufficiently large.

6. Explain what it means for a sample proportion to have a normal distribution.

Exercises

7. Suppose that the true proportion of registered voters who favor a mayoral candidate is 0.45. Find the mean and standard deviation of the sample proportion for samples of the following sizes.

a. $n = 30$

b. $n = 45$

c. $n = 65$

d. What happens to the size of the standard deviation of the sample proportion as the sample size increases?

8. Suppose that the true proportion of Americans over 25 years old that have a 4-year college degree is 0.35. Find the mean and the standard deviation of the distribution of the sample proportion for the following sample sizes.

a. $n = 38$

b. $n = 52$

c. $n = 75$

d. What happens to the size of the standard deviation of the sample proportion as the sample size increases?

9. Suppose that the true population proportion of Americans over 60 years old with high blood pressure is $p = 0.50$. In a random sample of twenty Americans over the age of 60, what is the probability that the proportion with high blood pressure will be greater than 0.60?

10. Suppose the true population proportion of people with blood type A+ is $p = 0.30$. What is the probability that the sample proportion for a sample of size 60 ($n = 60$) will be less than 0.25?

11. Suppose that the true proportion of Americans who save at least 10% of their income is 0.15. If $\hat{p}$ is the sample proportion of Americans surveyed who save at least 10% of their income from a sample of size 68, find the following probabilities.

a. $P\left(\hat{p} > 0.25\right)$

b. $P\left(\hat{p} < 0.09\right)$

c. $P\left(0.10 < \hat{p} < 0.20\right)$

d. $P\left(0.18 < \hat{p} < 0.25\right)$

12. Suppose that the true proportion of companies in the United States who requested financial assistance during the Coronavirus pandemic through the Paycheck Protection Program (PPP) was 62%. If $\hat{p}$ is the sample proportion of U.S. companies surveyed who requested financial assistance using PPP from a sample of size 100, find the following probabilities.

a. $P\left(\hat{p} > 0.75\right)$

b. $P\left(\hat{p} < 0.50\right)$

c. $P\left(0.40 < \hat{p} < 0.80\right)$

d. $P\left(0.70 < \hat{p} < 0.90\right)$

13. The director of a radio station in a large metropolitan area believes that the proportion of young professionals (his target market) in the area who prefer country music has increased from 25% to 35%. The director randomly decides to select 50 young professionals and ask them if they prefer country to any other type of music. If the sample proportion is greater than 0.35, he will switch to a new format emphasizing country.

a. If the true proportion of young professionals who prefer country has not changed, find the probability that the radio director will switch to the new format.

b. If the true proportion of young professionals who prefer country has changed as the director suspects, find the probability that the radio director will switch to the new format.

14. The owner of a large office building plans on building a dedicated smoking area outside, but will not do so if less than 30% of his employees smoke. He decides to randomly select 50 of the workers in the building and ask them whether or not they smoke. If the sample proportion of workers who smoke is less than 0.30, the owner will not create the smoking area.

a. Find the probability that the owner will not create the smoking area when the true proportion of smokers is 0.5.

b. Find the probability that the owner will create the smoking area when the true proportion of smokers is 0.2.

15. Eighty percent of the flights arriving in Atlanta for a large US airline are on time. If the FAA randomly selects 50 of the airline's flights, find the probability that:

a. at least 85% of the sampled flights will be on time.

b. at most 70% of the sampled flights will be on time.

c. between 75% and 85% of the sampled flights will be on time.

16. Approximately 7% of the nation's public-school children in grades 2 through 5 take medication for attention deficit hyperactivity disorder (ADHD), a developmental disorder characterized by impulsiveness or difficulty concentrating or sitting still. The main treatment prescribed for ADHD is Ritalin, a relatively safe drug with few side effects. A sample of 286 students is taken.

a. Find the probability that at least 4% of the school children in the sample take medication for ADHD.

b. Find the probability that between 5% and 8% of the school children in the sample take medication for ADHD.

9.4 Other Forms of Sampling

Random sampling is an effective means of obtaining a sample that is representative of the population. As we discussed previously, acquiring an exact sampling frame for the population under study is a requirement for simple random sampling, a requirement which can be time-consuming and expensive. There are other sampling strategies that

9.4 Exercises

Basic Concepts

1. What are advantages and disadvantages of non-probability samples?
2. What is a judgment sample? Give an example not in the text of when a judgment sample would be appropriate.
3. What is a convenience sample? Are these samples usually representative of the population?
4. What are the worst forms of non-probability samples?
5. Explain the idea of systematic sampling. What are the advantages and disadvantages of this sampling procedure?
6. Explain the idea of cluster sampling. What are the advantages and disadvantages of this sampling procedure?
7. Explain the idea of stratified sampling. What are the advantages and disadvantages of this sampling procedure?
8. Explain why a systematic sample is not a random sample.

Exercises

9. An employee-owned company has 6000 hourly employees and 2000 salaried employees. The human resources department decides to develop a survey on several different benefit plans, including child care and retirement benefits, that they may offer employees in the future. The results of the survey are to be presented to the board of directors for consideration. Because the human resources department wants to be sure of equal representation of the employees, it has decided to randomly select 500 hourly employees and 500 salaried employees. What kind of sampling method is the human resources department using? If the sample is used to make inferences regarding the desirability of various benefits packages for all employees, discuss any deficiencies in the sampling procedure.

10. A social researcher in Florida wants to determine the average number of children per family in the state.
    - **a.** What is the population of interest?
    - **b.** What variable will be measured?
    - **c.** What level of measurement is the variable of interest?
    - **d.** Discuss the steps that would be necessary for each of the following sampling methods.
        - **i.** Simple random sampling
        - **ii.** Cluster sampling
        - **iii.** Stratified sampling
    - **e.** What sampling method do you believe would be the most cost-effective? Justify your answer.

11. A stock analyst wants to estimate the average yearly earnings of stocks on the New York Stock Exchange.

a. What is the population of interest?

b. Discuss the steps necessary to apply each of the following sampling methods.

i. Simple random sampling

ii. Cluster sampling

iii. Stratified sampling

12. A news reporter in Orlando, Florida wants to conduct a survey to determine how local residents feel about the institution of a state income tax. Since there will be a lot of people from which to choose, he goes to Disney World and randomly selects individuals entering the complex. He asks the selected people whether or not they favor a state income tax in Florida. The responses to the survey are as follows.

Survey Responses	
Category	**% of Responses**
Favor a Florida State Income Tax	50
Do Not Favor a Florida State Income Tax	50

a. What sampling technique was used for this survey?

b. What biases may be present in the responses?

c. Is 50% a reasonable point estimate of the proportion of Orlando residents who favor the state income tax? Explain.

CR Chapter Review

Key Terms and Ideas

- Sampling
- Sampling Distribution
- Voluntary Sampling
- Selection Bias
- Biased Sample
- Sampling Frame
- Simple Random Sample
- Random Number Table
- Random Number Generator
- Sampling Distribution of the Sample Mean
- Point Estimator
- Unbiased Estimator
- Standard Error of the Mean
- Central Limit Theorem
- Population Proportion
- Sample Proportion
- Sampling Distribution of the Sample Proportion
- Error of Estimation
- Probability Samples
- Non-Probability Samples
- Judgment Sample
- Convenience Sample
- Systematic Sampling
- Cluster Sampling
- Strata (stratum)
- Stratified Sampling

To estimate variation, the sample variance s^2 is the best point estimator of the population variance σ^2. The sample variance is an unbiased estimator of the population variance which means that the values of the sample variance tend to center around the value of the population variance. The sample standard deviation, s, tends to underestimate the population standard deviation, σ, although the bias becomes smaller as the size of the sample increases.

Example 10.1.3

Determining the Best Estimator of the Population Standard Deviation of Apple iPhone Costs

Using the data from Example 10.1.1, determine the best point estimate of the population standard deviation of iPhone costs.

Solution

The best estimator of the population standard deviation is the sample standard deviation. The sample standard deviation can be determined using technology or computed manually using the formula $s = \sqrt{\frac{\Sigma\left(x_i - \bar{x}\right)^2}{n-1}}$ (see Section 4.2). The sample standard deviation using either method is 43.004. Therefore, the best estimate of the population standard deviation of iPhone costs is approximately \$43.

Technology

For instructions on how to compute the sample standard deviation using technology, please visit stat.hawkeslearning.com and navigate to **Discovering Statistics and Data, Fourth Edition > Technology Instructions > Descriptive Statistics > One Variable.**

10.1 Exercises

Basic Concepts

1. What is statistical inference?
2. What is an estimator?
3. Explain, in your own words, the difference between the terms *estimator* and *estimate*.
4. Give three examples of point estimators. Identify the parameters being estimated by these estimators.
5. What are two important questions to consider when estimating a population mean?
6. What is mean squared error?
7. What is an unbiased estimator? Give an example.
8. Why is the sample mean considered the best point estimate of the population mean?
9. Are all estimators unbiased? Explain.
10. What are two characteristics of the best available estimate for a parameter?
11. Are biased estimators useful? Give an example of one.

Exercises

12. The mean monthly water bill for 79 randomly selected residents of the local apartment complex is \$138. What is the best point estimate for the mean monthly water bill for all residents of the local apartment complex?

13. At the local college, a study found that students earned an average of 8.8 credit hours per semester with a sample standard deviation of 2.8 credit hours. The study randomly selected a sample of 108 students. What is the best point estimate for the average number of credit hours per semester for all students at the local college?

14. Suppose a sample of 2500 new car buyers is drawn. Of those sampled, 882 preferred foreign cars over domestic cars. Determine a point estimate for the population proportion of new car buyers who prefer foreign cars over domestic.

15. An environmentalist draws a sample of 250 oil tankers. Thirty-two of the oil tankers had spills last year. Using this information, find a point estimate for the population proportion of oil tankers that had spills last year.

16. In a sample of 126 ChatGPT responses, the average response processing time was found to be 2.9 seconds with a variance of 1.2 seconds. Give a point estimate for the population standard deviation of ChatGPT response processing times.

17. Consider a sample of tires. Their diameters are measured and found to have a variance of 2.4 centimeters^2. Give a point estimate for the population variance of tire diameters.

18. A sample of 6 bottles of sesame oil are randomly selected from a shipment of 3000 bottles. Each bottle is designed to contain 240 ml of sesame oil. The contents of bottles are measured, and the results are given below. Give a point estimate for the population standard deviation of the shipment.

240, 241, 235, 233, 234, 229

10.2 Estimating the Population Mean, σ Known

Rarely will a point estimate of the population mean result in a value which exactly matches the population mean, μ. If an estimate is used for decision making, it is desirable that there be some indication of its potential error. One of the significant limitations of simply reporting a point estimate is the lack of information concerning the estimator's accuracy.

Interval estimates, however, are constructed to provide additional information about the precision of the estimate. An **interval estimator** is made by developing an upper and a lower boundary for an interval that will hopefully contain the population parameter. It would be easy to construct an interval estimator that would always contain the population parameter: for example, the interval from negative infinity to positive infinity. But this particular estimator would not contain any useful information about the location of the population parameter. In interval estimation, the smaller the interval for a given level of confidence, the better the estimator.

Interval Estimate

An **interval estimate** defines an upper and lower boundary for an interval that will hopefully contain the population parameter with a given confidence level.

DEFINITION

Note

It is important to note that confidence and probability are conceptually related, but they are not the same thing.

10.2 Exercises

Basic Concepts

1. **a.** What is an interval estimator?

b. What is the difference between a point estimate and an interval estimate?

2. What is the distinction between probability and confidence?

3. What is the role of the z-value in the confidence interval expression?

4. Describe in words the ideas behind the construction of a confidence interval.

5. Suppose a 95% confidence interval for an estimate of a mean was 111 to 189. Explain what is wrong with the following expression: $P(111 < \mu < 189) = 0.95$.

6. What are the conditions required in order to construct a $100(1-\alpha)\%$ confidence interval using the expression $\bar{x} \pm z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$?

7. Describe the effect on the width of a confidence interval as each of the following increases: n, $1-\alpha, \alpha$, $\bar{x}$.

8. What expression indicates the margin of error? Is this the same as the maximum error of estimation?

9. What is $\bar{x} \pm z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ an estimate of?

Exercises

10. Find $z_{\alpha/2}$ for the following levels of α.

a. $\alpha = 0.05$

b. $\alpha = 0.01$

c. $\alpha = 0.10$

d. $\alpha = 0.04$

e. $\alpha = 0.02$

f. $\alpha = 0.08$

11. Find $z_{\alpha/2}$ for the following confidence levels.

a. 98%

b. 94%

c. 92%

d. 96%

e. 88%

f. 85%

12. Construct a 90% confidence interval for the true mean of a normal population if a random sample of size 40 from the population yields a sample mean of 75 and the population has a standard deviation of 5.

13. A paint manufacturer is developing a new type of paint. Thirty panels were exposed to various corrosive conditions to measure the protective ability of the paint. The mean life for the samples was 168 hours before corrosive failure. The life of paint samples is assumed to be normally distributed with a population standard deviation of 30 hours. Find the 95% confidence interval for the mean life of the paint.

14. The chief purchaser for the State Education Commission is reviewing test data for a metal link chain which will be used on children's swing sets in elementary school playgrounds. The average tensile strength for a sample of 50 pieces of chain is

5000 psi. Based on past experience, the tensile strength of metal chains is known to be normally distributed with a standard deviation of 100 pounds. Estimate the actual mean tensile strength of the metal link chain with 99% confidence.

15. A research scholar wants to know how many times per week a strain of E. coli reproduces. From a sample of 476 organisms there was an average of 3 reproductions per hour. The population standard deviation is known to be 0.3 reproductions per hour. Construct a 95% confidence interval for the true population mean number of reproductions per hour for this bacteria.

16. An educational psychologist wishes to know the mean number of words that a third grader can read per minute. She wants to make an estimate at an 80% level of confidence. For a sample of 196 third graders, the mean words per minute was 27.1. Assume a population standard deviation of 3.2. Construct the confidence interval for the mean number of words that a third grader can read per minute and interpret the results.

10.3 Estimating the Population Mean, σ Unknown

In the last section we assumed that the population standard deviation is known. In practice this assumption is not very realistic, since the standard deviation describes variability about the mean. If the population standard deviation is known, the mean is usually also known, and there is no need to create an interval estimate for it. Why estimate something we already know? The methodology for creating the confidence interval when σ is unknown is quite similar to the methodology when σ is known but much more useful.

Interval Estimation of the Population Mean for a Normal Population with σ Unknown

If σ is not known and either the population is normally distributed or $n > 30$, the derivation of the confidence interval must be changed slightly.

Student's *t*-Distribution

Provided the population from which the sample is drawn is normally distributed or $n > 30$, the distribution of the quantity

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

where s is the standard deviation of the sample, has a **Student's *t*-distribution**.

FORMULA

The *t*-distribution is very much like the normal distribution (see Figure 10.3.1). It is a symmetrical, bell-shaped distribution with slightly thicker tails than a normal distribution. The shape of the *t*-distribution approaches the normal distribution as the **degrees of freedom**, the one parameter of the *t*-distribution, becomes larger.

William Sealy Gossett: The Student

Upon graduating from New College, Oxford with a strong understanding of mathematics, W.S. Gossett began working at the Guinness brewery in Dublin, Ireland. While working at Guinness, Gossett applied his statistical knowledge to find the best yielding varieties of barley, and in 1908, he developed the *t*-distribution. Few other statisticians at the time saw the merit in developing small-sample methods since most of their work required large data sets; however, Gossett was convinced of the importance of his work. Unfortunately, Guinness had prohibited its employees from publishing papers to protect trade secrets, and thus did not originally allow Gossett to publish his findings. After convincing the brewery that his statistical methods would be of no use to competing breweries, Guinness allowed Gossett to publish his conclusions on the *t*-distribution, but only under the pseudonym, "Student", to avoid issues with other staff members and company policies. To this day, Gossett's most noteworthy achievement is known simply as the "Student's *t*-distribution."

Solution

Using the results from the initial sample,

$$n = \left(\frac{1.96 \cdot 4.3}{0.25} \right)^2 \approx 1136.50$$

$n \approx 1137$. (Always round up to assure the required confidence.)

Notice that while the sample data values are being collected they can be used to improve the estimate of the population standard deviation. For example, suppose the sample standard deviation after sampling the first 100 observations was 4.1. Using this estimate of s instead of 4.3 results in a sample size of 1034 compared to the original specification of 1137. The notion of modifying the sample size estimate as additional data is observed can be applied at regular intervals during the sampling process until the estimate of the standard deviation stabilizes.

10.3 Exercises

Basic Concepts

1. Why is the assumption that the population standard deviation is known when estimating the population mean not very realistic?
2. What effect does knowing the standard deviation of the population have on the construction of the confidence interval?
3. What is the Student's t-distribution?
4. What is the parameter of the t-distribution? How is it calculated?
5. What is the value of having a confidence interval with a small width?
6. Can a confidence interval be constructed with a width of your choice? Explain.
7. What is the margin of error? What is the connection between the expression for the margin of error and the equation to determine the sample size?
8. What is the difference between the method of determining the sample size when σ is known versus when σ is unknown?

Exercises

9. Find the t-value such that 0.025 of the area under the curve is to the right of the t-value. Assume the degrees of freedom equal 13.

10. Find the t-value such that 0.01 of the area under the curve is to the right of the t-value. Assume the degrees of freedom equal 21.

11. Find $t_{\alpha/2,df}$ for the following combinations of α and n.
    - **a.** $\alpha = 0.05, n = 15$
    - **b.** $\alpha = 0.01, n = 20$
    - **c.** $\alpha = 0.10, n = 8$
    - **d.** $\alpha = 0.05, n = 12$
    - **e.** $\alpha = 0.01, n = 18$
    - **f.** $\alpha = 0.10, n = 22$

12. A random sample, consisting of the values listed below, was taken from a normally distributed population. Assuming the standard deviation of the population is unknown, construct a 99% confidence interval for the population mean.

27.4	26.5	25.7	31.4
28.2	21.9	16.3	22.7
18.8	34.4	29.2	20.5

13. Construct an 80% confidence interval for the mean of a normal population assuming that the values listed below comprise a random sample taken from the population. The population standard deviation is unknown.

83.9	87.4	65.2	86.0	73.1
80.3	92.7	87.5	69.3	77.5
91.9	71.1	79.1	72.4	88.2

14. An FDA representative randomly selects 8 packages of ground chuck from a grocery store and measures the fat content (as a percent) of each package. The resulting measurements are given below.

13%	12%	14%	17%
15%	16%	18%	15%

a. Calculate the sample mean and the sample standard deviation of the fat contents.

b. Construct a 90% confidence interval for the true mean fat content of all the packages of ground beef.

c. What assumption did you make about the fat content in constructing your interval?

15. A hospital would like to determine the mean length of stay for its patients having abdominal surgery. A sample of 15 patients revealed a sample mean of 6.4 days and a sample standard deviation of 1.4 days.

a. Find a 95% confidence interval for the mean length of stay for patients with abdominal surgery.

b. Interpret this interval and state any assumptions that were made in the construction of the interval.

16. An independent group of food service personnel conducted a survey on tipping practices in a large metropolitan area. They collected information on the percentage of the bill left as a tip for 25 randomly selected bills. The average tip was 18.3% of the bill with a standard deviation of 2.7%.

a. Construct an interval to estimate the true average tip (as a percent of the bill) with 99% confidence.

b. Interpret the interval, and state any assumptions that were made in the construction of the interval.

17. A travel agent is interested in the average price of a hotel room during the summer in a resort community. The agent randomly selects 15 hotels from the community and determines the price of a regular room with a king size bed. The average price of the room for the sample was \$160 with a standard deviation of \$30.

a. Construct an interval to estimate the true average price of a regular room with a king size bed in the resort community with 90% confidence.

b. Interpret the interval, and state any assumptions that were made in the construction of the interval.

18. A technician working for the Chase-National Food Additive Company would like to estimate the preserving ability of a new additive. This additive will be used for Auntie's brand preserves. Based on past tests, it is believed that the time to spoilage for this additive has a standard deviation of 6 days. To be 90% confident of the true mean time to spoilage, what sample size will be needed to estimate the mean time to spoilage with an accuracy of one day?

19. A computer software company would like to estimate how long it will take a beginner to become proficient at creating a graph using their new spreadsheet package. Past experience has indicated that the time required for a beginner to become proficient with a particular function of the new software product has an approximately normal distribution with a standard deviation of 15 minutes. Find the sample size necessary to estimate the true average time required for a beginner to become proficient at creating a graph with the new spreadsheet package to within 5 minutes with 95% confidence.

20. A food truck vendor is evaluating a downtown location by counting the number of people who walk past the prospective location on a particular day during lunch time (i.e. 11:00 AM to 2:00 PM). A preliminary study has indicated a standard deviation of about 30 people per lunch period. How many lunch periods will be needed to estimate the average number of people who walk past the prospective location during the lunch period to within 9 people with 90% confidence?

10.4 Estimating the Population Proportion

Recall that we discussed point estimation of proportions in Section 10.1. In this section we will develop a confidence interval for a population proportion based on the sampling distribution of the point estimate (see Section 9.3 for review).

Note

Proportions have various equivalent formats, including fractions, decimals, and percentages.

$$\frac{3}{5} = 0.6 = 60\%$$

Interval Estimation of a Population Attribute

The use of confidence intervals to apprise a decision maker of the reliability of estimates of a population mean can also be applied to estimating proportions. The random variable, $\hat{p}$, has a binomial distribution that can be approximated with a normal random variable.

Thus, the sample proportion, $\hat{p}$, is distributed normally with mean, p, and variance,

Predictive Analytics: Direct Mail Simulation

One of the illusions magicians try to create is the ability to accurately forecast complex future events. The goal of predictive analytics is similar and rather magical, but it is not an illusion. As the name implies, predictive analytics is a broad set of statistical, machine learning, and data mining techniques used to forecast complex future events with a significant degree of accuracy.

In the appendix of the Hawkes Learning courseware, there is a role playing game called Direct Mail. The purpose of the game is for the player to develop a predictive strategy for direct mail marketing.

The direct mail game is a simulation game in which you will play the role of a junior marketing manager in charge of direct mail marketing for a product you have been given to help manage. You are a new hire and are anxious to impress the marketing manager with some statistical predictive analytics.

In your company the marketing department will develop a brochure to send out to members of various mailing lists. The only thing you will know about the mailing lists are that the marketing manager has given you a set of lists which she believes may be potentially profitable using the developed brochure. You have been asked to develop a strategy to evaluate the risks and opportunities associated with each list for any potential profitability. If the marketing manager believes your initial assessments are good, you will be allowed to do the remainder of the lists on your own. Is this just a guessing game, or is there some statistical science that can be applied to aid you? As you likely guessed, the latter is the correct answer.

In direct mail marketing one of the big problems is getting the recipient to open your mail. How personal the mail looks will influence the probability it will be opened. In addition to choosing whether to mail a list or not, you can define your own mailing tactic that may impact whether your mail is opened. There are two mailing choices. First, you must decide whether to use first class or bulk mail. First class costs $0.68 per piece and bulk mail costs $0.49 for each piece mailed. First class looks more personal, but bulk is cheaper; which will be best for your particular list? Another personalizing feature you can invest in is whether you want to use mailing labels or have the envelopes typed with the name and address.

The marketing manager is anxiously awaiting your predictive analytic model and mailing strategy.

10.4 Exercises

Basic Concepts

1. Describe, in layman's terms, how a confidence interval is constructed for a population proportion.
2. It seems that estimating proportions produces estimates which are much more precise than those for means. Explain why this is the case or is not the case.
3. When determining the sample size required to estimate a population proportion within some specified error level with a specific level of confidence, what is the guideline to follow when there is no estimate available for the population proportion? Why is this done?

Exercises

4. Acid rain accumulations in lakes and streams in the northeastern part of the United States are a major environmental concern. A researcher wants to know what fraction of lakes contain hazardous pollution levels. From 200 randomly selected lakes, it is determined that 45 of them have an unsafe concentration of acid rain pollution.

a. Calculate the best point estimate of the population proportion of lakes that have unsafe concentrations of acid rain pollution.

b. Determine a 95% confidence interval for the population proportion.

c. If a local politician states that only 20% of the lakes are contaminated, does the study provide overwhelming evidence at the 95% level to contradict his views?

5. *The Richland Gazette*, a local newspaper, conducted a poll of 1000 randomly selected readers to determine their views concerning the city's handling of snow removal. The paper found that 650 people in the sample felt the city did a good job.

a. Compute the best point estimate for the percentage of readers who believe the city is doing a good job of snow removal.

b. Construct a 90% confidence interval for this percentage.

6. The clinical testing of drugs involves many factors. For example, patients that have been given placebos, which are harmless compounds that have no effect on the patient, often will still report that they feel better. Assume that in a study of 500 random subjects conducted by the Poppins Sucre Drug Company, the percentage of patients reporting improvement when given a placebo was 37%.

a. What would be a 95% confidence interval for the true proportion of patients who exhibit the placebo effect? Interpret this interval in terms of the problem.

b. What would the 99% confidence interval be?

c. To gain the additional 4% of confidence how much wider did the interval become?

7. The Peacock Electric Company thinks that 40% of their customers would be interested in bundling solar power purchasing with their electric bill. A random sample of 400 households reveals that 110 of the households are interested in this.

a. Construct a 99% confidence interval for the true proportion of households interested in bundling solar with their utilities.

b. Do you feel the company is accurate in its belief about the proportion of customers who have interest in bundling solar power with their utilities? Justify your answer.

8. Running continues to be a very popular sport in America. At a major race, there may be over 10,000 people entered to run. The race promoters for a road race in the Pacific Northwest took a random sample of 750 runners out of the 5000 runners entered to estimate the number of runners who will need hotel accommodations. Five hundred runners indicated they would need hotel accommodations.

a. Construct a 90% confidence interval for the true proportion of runners who will need hotel accommodations.

b. Is the confidence interval obtained sufficiently narrow to be of help in planning the number of hotel rooms which will be necessary to accommodate the runners? Justify your answer.

9. Over the past few years the national rate for homeownership has been stable near 64% likely due to low mortgage interest rates and an active real estate market. For various reasons, which could include short supply of affordable homes and demographical factors, the homeownership rate in West Palm Beach, Florida is suspected to be lower than the national rate. A sample of 80 households was randomly selected from the population of Palm Beach County in Florida. Suppose that 47 of the households sampled were owned by the residents of the homes.

a. Construct a 95% confidence interval for the proportion of households in the area sampled that are owned by the residents of the homes.

b. Is there evidence at the 95% level that the proportion of the households in the area sampled that are owned by residents is less than the national rate?

10. In the *Gallup Poll Monthly*, it was reported that 31% of the people surveyed in a recent poll claimed that vegetables were their least favorite food. Surprisingly, only 14% responded with liver, and 10% of those surveyed did not submit a response because they claimed that they liked everything. The poll was based upon a sample of 1001 people. Assume that a random sample of Americans was chosen, and construct a 90% confidence interval for the percentage of all Americans who say that vegetables are their least favorite food.

11. The Big Green Poster Company wants to estimate the percentage of poster sites controlled by their competition, Bird's Billboard Service. What sample size would be necessary to estimate this percentage to within 3% with 95% confidence? (They think Bird's controls about 33% of the boards.)

12. Researchers working in a remote area of Africa feel that 40% of families in the area are without adequate drinking water either through contamination or unavailability. What sample size will be necessary to estimate the percentage without adequate water to within 5% with 99% confidence?

13. Companies that provide environmental cleanup for hazardous waste and toxic chemicals are growing rapidly. W.R. Gross is thinking about entering this field with a subsidiary called Saf-t-Soil. They wish to estimate the true proportion of U.S. corporations that produce hazardous waste as a by-product of their manufacturing process to within 10% with 80% confidence. What sample size will be needed?

14. The public relations manager for a political candidate would like to determine if the registered voters in the candidate's district agree with the politician's view on a particular issue. Find the sample size necessary for the public relations manager to estimate the true proportion to within 5% with 85% confidence.

terms of the problem, to find a 95% confidence interval for the standard deviation, we take the square root of the endpoints of the confidence interval for the variance, yielding

$$0.0275 < \sigma < 0.0730.$$

The 95% confidence interval for the standard deviation of fill for the bottles is between 0.0275 ounce and 0.0730 ounce, indicating that the process meets the specifications of being less than 0.1 ounce.

10.5 Exercises

Basic Concepts

1. What is the sampling distribution for $\dfrac{(n-1)s^2}{\sigma^2}$?

2. What assumption must hold to use the chi-square distribution to make inferences about the population variance?

3. True or False: The chi-square distribution is skewed to the right.

4. What is the symbol for a critical value for the chi-square distribution? Describe the meaning of this critical value.

5. Give an example where we would want to calculate a confidence interval for σ^2.

Exercises

6. Determine the critical value(s) for each of the following tests for a population variance where the assumption of normality is satisfied.

a. Right-tailed test, $\alpha = 0.01$, $n = 20$

b. Right-tailed test, $\alpha = 0.005$, $n = 5$

7. Determine the critical value(s) for each of the following tests for a population variance where the assumption of normality is satisfied.

a. Right-tailed test, $\alpha = 0.025$, $n = 18$

b. Right-tailed test, $\alpha = 0.05$, $n = 41$

8. A bolt manufacturer is very concerned about the consistency with which his machines produce bolts that are $\dfrac{3}{4}$ inches in diameter. When the manufacturing process is working normally the standard deviation of the bolt diameter is 0.05 inches. A random sample of 30 bolts has an average diameter of 0.25 inches with a standard deviation of 0.07 inches.

a. Construct a 95% confidence interval for the standard deviation of the bolt diameter. Interpret the interval.

b. What assumption did you make about the diameter of the bolts in constructing the confidence interval in part **a.**?

9. A drug that is used for treating cancer has potentially dangerous side effects if it is taken in doses that are larger than the required dosage for the treatment. The pharmaceutical company that manufactures the drug must be certain that the standard deviation of the drug content in the tablet is not more than 0.1 mg. Twenty-five tablets are randomly selected and the amount of drug in each tablet is measured. The sample has a mean of 20 mg and a variance of 0.015 mg.

a. Construct a 99% confidence interval for the variance of the amount of drug in each tablet. Interpret the interval.

b. What assumption did you make about the amount of drug contained in the tablets in constructing the confidence interval in part **a.**?

10. A conservative investor would like to invest some money in a bond fund. The investor is concerned about the safety of her principal (the original money invested). Colonial Funds claims to have a bond fund which has maintained a consistent share price of $7. They claim that this share price has not varied by more than $0.25 on average since its inception. To test this claim, the investor randomly selects 25 days during the last year and determines the share price for the bond fund. The average share price of the sample is $7 with a standard deviation of $0.35.

a. Construct a 90% confidence interval for the standard deviation of the share price of the bond fund. Interpret the interval.

b. What assumption did you make about the share prices of the bond fund in constructing the confidence interval in part **a.**?

11. A manufacturer of automobile batteries is concerned about the life of the batteries that are produced. The manufacturer is comfortable with the average life of the batteries but more concerned about the standard deviation. Research has shown that the average life of the automobile batteries is 60 months. However, the manufacturer would like the standard deviation of the life of the automobile batteries to be relatively small, say, approximately six months. To determine a reliable range of the standard deviation of the batteries currently being produced, the manufacturer took a random sample of 15 batteries and found that the average life was 58 months with a standard deviation of seven months.

a. Construct a 98% confidence interval for the standard deviation of the life of their automobile batteries. Interpret this interval.

b. What assumptions did you make about the life of a battery being produced by the manufacturer?

12. Almost all smart devices (phones, tablets, and computers) are made with touch screens. A concern of many consumers is the shelf life of the "touch" component of the screens. A consumer advocacy group wanted to inform its members of a range that they can expect their touch screens to last. The group took a sample of 29 screens and measured the life of the "touch" function of the screens. That is, they used digital devices to simulate billions of touches to determine the life of the screens. Of the 29 screens sampled, the average "touch" life was 90 months with a standard deviation of six months. Construct an 80% confidence interval for the standard deviation of the life of the touch screens. Interpret this interval.

Note

Between **Step 2** and **3** of the hypothesis testing procedure, it is typical to collect sample data from the population and prepare it for analysis. However, for the problems presented in this textbook, the required data has already been provided, so this step is not necessary for our purposes.

In real-world scenarios, gathering sample data can be a challenging undertaking. In fact, according to data science literature, approximately 80% of the time required to perform a statistical analysis is devoted to collecting, extracting, preparing, cleaning, and organizing the sample data, with only 20% of the time dedicated to the actual statistical analysis.

Note

Note at **Step 4** there are two options; you can find the critical value of the test statistic or the *P*-value of the test statistic. Both methods will always produce equivalent results; meaning, the decision regarding the hypothesis test will always be the same with both methods. We will often cover both methods in an example to illustrate this. Even though we may show a critical value and a *P*-value, only one of these is required to make the decision to reject or fail to reject the null hypothesis. You or your instructor may have a preference of one method over another.

Steps in the Hypothesis Test

Step 1: Determine the population parameter to be used and develop the null and alternative hypotheses.

Step 2: Specify the significance level α.

**see sidebar*

Step 3: Validate the assumptions of the hypothesis test, identify the appropriate test statistic, and compute its value.

Step 4: Determine the critical value(s) or *P*-value.

Critical Value Method	*P*-Value Method
Find the critical value(s). (It may help to draw a graph displaying the critical value(s), the rejection region, and the test statistic.)	Find the *P*-value based on the value of the test statistic and the alternative hypothesis. (It may help to draw a graph displaying the test statistic and *P*-value.)

Step 5: Choose between the null and alternative hypotheses.

Critical Value Method	*P*-Value Method
·Reject H_0 if the test statistic is in the rejection region. ·Fail to reject H_0 if the test statistic is not in the rejection region.	·Reject H_0 if *P*-value is $< \alpha$. ·Fail to reject H_0 if *P*-value is $\geq \alpha$.

Step 6: State the conclusion in terms of the original question.

PROCEDURE

11.1 Exercises

Basic Concepts

1. What is a hypothesis?
2. What is the first step in the test of a hypothesis?
3. Describe the common elements present in all hypothesis tests.

4. Summarize the difference between the null and alternative hypotheses.
5. Define and give an example of a one-sided alternative. How does this differ from a two-sided alternative?
6. Is there a way to be absolutely certain your decision is correct when performing a hypothesis test? Explain.
7. What are the three important things you must be able to do in order to be successful at formulating hypothesis testing problems?
8. Describe a Type I error.
9. Describe a Type II error.
10. Explain how Type I and Type II errors influence the construction of a hypothesis.
11. Can both Type I and Type II errors be controlled in the hypothesis testing procedure? Explain.
12. What is the level of the test?
13. Why is a Type II error difficult to express numerically?
14. In the hypothesis testing process, in what way does the data guide our decision-making?
15. Explain the relationship between induction and the hypothesis testing process.

Exercises

16. The town mayor believes that more than 47% of the town residents favor annexation of a new community. How should she formulate the hypotheses to test her claim?

17. A chocolate chip manufacturer would like to know if its bag filling machine works correctly at the 450 gram setting. Assume the population is normally distributed. How should the manufacturer formulate the hypotheses to test if the bags are being overfilled?

18. A hospital director believes that 29% of the lab reports contain errors and feels an audit is required. A sample of 300 reports found 99 errors. Is there sufficient evidence at the 0.02 level to refute the hospital director's claim? State the null and alternative hypotheses for this test.

19. An engineer has designed a valve that will regulate water pressure on an automobile engine. The valve was tested on 140 engines and the mean pressure was 7.7 lbs/square inch. Assume the variance is known to be 0.64. If the valve was designed to produce a mean pressure of 7.9 lbs/square inch, is there sufficient evidence at the 0.10 level that the valve performs below the specifications? State the null and alternative hypotheses.

20. Using traditional methods, it takes 10.9 hours to receive a basic flying license. A new license training method using Computer Aided Instruction (CAI) has been proposed. Set up the hypotheses to test the claim at the 0.05 level that the new technique performs differently than the traditional method. State the null and alternative hypotheses.

21. Our environment is very sensitive to the amount of ozone in the upper atmosphere. The level of ozone normally found is 7.6 parts/million (ppm). A researcher believes that the current ozone level is higher than the normal level. Set up the hypotheses to test the researcher's claim.

22. An automobile manufacturer claims that their van has a 36 miles per gallon (MPG) rating. An independent testing firm has been contracted to test the MPG for this van. After testing 53 vans they found a mean MPG of 32.8 with a standard deviation of 0.3 MPG. Is there sufficient evidence at the 0.05 level that the vans underperform the manufacturer's MPG rating? State the null and alternative hypotheses for this test.

23. A restaurant owner believes that tardiness has become a problem with her staff. In past years around 5% of her employees showed up late for their shift. She believes that the current rate is much higher. How should she formulate the hypotheses to test her belief?

24. A sample of 800 computer chips revealed that 79% of the chips do not fail in the first 1000 hours of their use. The company's promotional literature claimed that more than 76% do not fail in the first 1000 hours of their use. Is there sufficient evidence at the 0.02 level to support the company's claim? State the null and alternative hypotheses for this test.

25. A lumber company is making boards that are 2784 mm tall. If the boards are too long they must be trimmed, and if they are too short they cannot be used. A sample of 31 boards is made, and it is found that they have a mean of 2779.6 mm with a standard deviation of 10 mm. Is there evidence at the 0.10 significance level that the boards are too short and unusable? State the null and alternative hypotheses for this test.

26. For the following situations, develop the appropriate H_0 and H_a and state what the consequences would be for Type I and Type II errors.

a. The Standard Tire Company has introduced a new tire in Europe that will be guaranteed to last at least 50,000 kilometers. Standard Tire has hired an independent agency to determine if there is overwhelming evidence that their tires will last through the warranty period.

b. A fisheries scientist claims that tilapia fed once a day with a specified formula will grow at least 200 grams in 110 days. Test whether there is overwhelming evidence to support this claim.

c. A horticulturist is experimenting with a new type of basil to grow and sell at the local farmer's market. Regular basil plants require at least 55 days to germinate at a temperature of 20° Celsius. This new type of basil is touted as being fast growing, so she expects that it will germinate in less time, resulting in larger plants.

27. For the following situations, develop the appropriate H_0 and H_a and state what the consequences would be for Type I and Type II errors.

a. A company that manufactures one-half inch bolts selects a random sample of bolts to determine if the diameter of the bolts differs significantly from the required one-half inch.

b. A company that manufactures safety flares randomly selects 100 flares to determine if the flares last at least three hours on average.

c. A consumer group claims that a new electric vehicle (EV) model gets significantly fewer miles on a single battery charge than advertised by the manufacturer. To confirm this claim, researchers randomly drive several of the EV vehicles of this particular model and measure the distance traveled after a single full battery charge.

In the previous two examples, we have considered a two-sided alternative and a one-sided "greater than" alternative. When considering a one-sided "less than" alternative, the procedure is very similar to that of a one-sided "greater than" alternative. The null hypothesis will be rejected if the calculated value of the test statistics, z, is less than the critical value, z_α, for the specified level of significance.

One-Sided "Less Than" Alternatives

For tests which are based on a test statistic which has a standard normal distribution and "less than" alternatives, find the value of z that *cuts off* α worth of probability in the left-hand tail of the distribution. The critical values for "less than" alternative hypotheses are given in Table 11.2.4 for typical values of α.

Table 11.2.4 – Critical Values of the *z*-Test Statistic for One-Sided (Less Than) Alternatives

Significance Level	Definition of Ordinary Variability	$-z_\alpha$
0.20	Upper 80% of the distribution	–0.84
0.10	Upper 90% of the distribution	–1.28
0.05	Upper 95% of the distribution	–1.645
0.01	Upper 99% of the distribution	–2.33

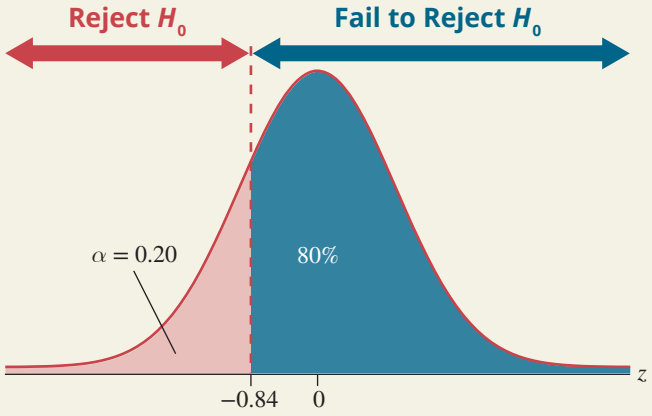


Figure 11.2.4

The figure above shows the rejection region for a test with a one-sided "less than" alternative hypothesis, and a significance level of 0.20. For this test, the null hypothesis will be rejected if the calculated value of the test statistic is less than –0.84.

The P-value is given by $P(z \leq z_0)$, where z_0 is the observed value of the test statistic.

PROCEDURE

11.2 Exercises

Basic Concepts

1. What is the rationale for the z-statistic?
2. Describe the distribution of the z-test statistic.

3. What are critical values? How do critical values influence the decision rule in the hypothesis testing procedure?
4. Suppose a null hypothesis was rejected at $\alpha = 0.05$. Would it be rejected at 0.10? Explain.
5. What is a *P*-value?

Exercises

6. Determine the critical value(s) of the test statistic for each of the following tests for the population mean with σ known.
   a. Left-tailed test, $\alpha = 0.01$
   b. Right-tailed test, $\alpha = 0.10$
   c. Two-tailed test, $\alpha = 0.05$

7. For each of the following combinations of the *P*-value and α, decide whether you would reject or fail to reject the null hypothesis.
   a. *P*-value = 0.0839, $\alpha = 0.05$
   b. *P*-value = 0.0174, $\alpha = 0.02$
   c. *P*-value = 0.0444, $\alpha = 0.10$
   d. *P*-value = 0.0374, $\alpha = 0.01$

8. Consider the following hypothesis tests for the population mean with σ known. Compute the *P*-value for each test and decide whether you would reject or fail to reject the null hypothesis at $\alpha = 0.05$.
   a. $H_0: \mu = 15,\ H_a: \mu > 15,\ z = 1.58$
   b. $H_0: \mu = 1.9,\ H_a: \mu < 1.9,\ z = -2.25$
   c. $H_0: \mu = 100,\ H_a: \mu \neq 100,\ z = 1.90$

9. Consider the following hypothesis tests for the population mean with σ known. Compute the *P*-value for each test and decide whether you would reject or fail to reject the null hypothesis at $\alpha = 0.01$.
   a. $H_0: \mu = 10,\ H_a: \mu > 10,\ z = 2.00$
   b. $H_0: \mu = 82,\ H_a: \mu < 82,\ z = -2.45$
   c. $H_0: \mu = 100, H_a: \mu \neq 100,\ z = 2.70$

10. The Tesla S electric vehicle (EV) can travel on average up to 405 miles on a single battery charge. Another manufacturer claims that its EV can travel further on a full battery charge. A random sample of 50 vehicles from the new manufacturer produces a sample mean of 408 miles. Test the hypothesis that the mean distance traveled by the manufacturer's EV is greater than 405 at $\alpha = 0.05$. Assume that the population standard deviation is 12 miles.

11. Hurricane Ian swept through southeastern Florida causing billions of dollars of damage. Because of the severity of the storm and the type of residential construction used in this semitropical area, there was some concern that the average claim size would be greater than the historical average hurricane claim.

Historically, the average claim size was \$24,000 with standard deviation \$2400.[3] Several insurance companies collaborated in a data gathering experiment. They randomly selected 84 homes and sent adjusters to settle the claims. In the sample of 84 homes, the average claim was \$26,000.

a. What is the population being studied?

b. What statistical measure should you use in your hypothesis?

c. State your hypotheses.

d. Test the hypothesis at the 0.01 level.

e. Is there overwhelming evidence (at the 0.01 level) that home damage is greater than the historical average? Write your conclusion in the context of the original problem.

12. For adults, a cholesterol value under 200 mg/dl is preferred. High cholesterol values over a period of time can result in a stroke or heart attack. A random sample of 10 cholesterol readings for a patient has an average of 204 mg/dl. The population standard deviation of the test is 6 mg/dl.

a. Is there overwhelming evidence to conclude that the patient's cholesterol level is above 200 mg/dl at a 0.05 significance level?

b. What is the lowest cholesterol average that would allow the patient to conclude that their cholesterol level is above 200 mg/dl?

13. A horticulturist working for a large plant nursery is conducting experiments on the growth rate of a new shrub. Based on previous research, the horticulturist feels the average daily growth rate of the new shrub is 0.2 cm per day with a standard deviation of 0.03 cm. A random sample of 45 shrubs has an average growth of 0.15 cm per day. Will a test of hypothesis at the 0.05 significance level support the claim that the growth rate is less than 0.2 cm per day?

14. Government regulations restrict the amount of pollutants that can be released to the atmosphere through industrial smokestacks. To demonstrate that their smokestacks are releasing pollutants below the mandated limit of 5 parts per billion pollutants, REM Industries collects a random sample of 300 readings. The mean pollutant level for the sample is 4.85 parts per billion. The population standard deviation is known to be 0.30 parts per billion.

a. Does the data support the claim that the average pollutants produced by REM Industries are below the mandated level at a 0.01 significance level?

b. What assumption did you make in performing the test in part **a.**?

15. In 2022, 3.96 billion people used social media, which is more than half the population of the world.[4] The average time spent on social media around the world is 147 minutes a day. Jennifer's daily average time spent on social media for the last ten days was 132 minutes.

a. Perform a hypothesis test to determine if Jennifer's time spent on social media is significantly less than the average. Use $\alpha = 0.01$. Assume the population standard deviation is 32.5 minutes.

b. What assumption did you make in performing the test in part **a.**?

- *Is my result practically significant?* With very large samples, it's easy to get a small *P*-value (large test statistic) even though the sample value may not be practically different than the hypothesized value.
- *Are the assumptions I made to perform the test justifiable?*
- *Can I reproduce my results?*

11.3 Exercises

Basic Concepts

1. What are the two key questions to be asked in the hypothesis testing procedure in order to determine which test statistic is appropriate?

2. If the variance for a population is not known, how is the test statistic affected?

3. Suppose a null hypothesis was rejected at $\alpha = 0.05$. Would it be rejected at 0.01? Explain.

4. Discuss how *P*-values are used in the test of a hypothesis.

Exercises

5. Researchers studying the effects of diet on growth would like to know if a vegetarian diet affects the height of a child. The researchers randomly select 12 vegetarian children that are six years old. The average height of the children is 42.5 inches with a standard deviation of 3.8 inches. The average height for all six year old children is 45.75 inches.

a. What is the population being studied?

b. Conduct an hypothesis test to determine whether there is overwhelming evidence at $\alpha = 0.05$ that six year old vegetarian children are not the same height as other six year old children?

c. What assumption did you make in performing the test in part **b.**?

6. Consider the following σ unknown hypothesis tests for the population mean. Compute the *P*-value for each test and decide whether you would reject or fail to reject the null hypothesis at $\alpha = 0.01$.

a. $H_0: \mu = 25, H_a: \mu > 25, t = 2.7, n = 15$

b. $H_0: \mu = 0.85, H_a: \mu < 0.85, t = -2.5, n = 7$

c. $H_0: \mu = 1000, H_a: \mu \neq 1000, t = 2.0, n = 15$

7. Consider the following σ unknown hypothesis tests for the population mean. Compute the *P*-value for each test and decide whether you would reject or fail to reject the null hypothesis at $\alpha = 0.05$.

a. $H_0: \mu = 120, H_a: \mu > 120, t = 1.5, n = 20$

b. $H_0: \mu = 0.2, H_a: \mu < 0.2, t = -2.75, n = 18$

c. $H_0: \mu = 50, H_a: \mu \neq 50, t = 2.4, n = 5$

8. The average number of points scored by a team during an NFL football game is known to be 19.55. Use the Super Bowl data set to test whether the number of points scored by a team during the Super Bowl is different than 19.55 at $\alpha = 0.05$.

Data
stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Super Bowl

9. The American IPA style of beer has on average 6.47% alcohol by volume (ABV). Use the Beers and Breweries data set , which is a sample of American canned beers brewed in the U.S., to determine if the American IPAs brewed in California have more ABV than average at $\alpha = 0.05$. Calculate the *P*-value for this hypothesis test.

Data
stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Beers and Breweries

10. According to Trulia[12], the average price per square foot for Mount Pleasant homes sold in 2017 was $210. Using the Mount Pleasant Real Estate data set, which is a sample of homes for sale in three neighborhoods on the north side of Mount Pleasant, perform a hypothesis test to test the claim that the average price per square foot is lower in the Park West neighborhood than the city's average at $\alpha = 0.10$. Calculate the *P*-value for this hypothesis test.

Data
stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Mount Pleasant Real Estate Data

11. Del Valley Foods requires that corn supplied for canning must weigh more than 5 ounces per ear. South Valley Farms claims that the corn they supply meets the required specifications. A sample of 200 ears of corn are selected at random from a delivery. The sample has a mean of 5.01 ounces and a standard deviation of 0.30 ounces. Will a test of hypothesis at $\alpha = 0.10$ support South Valley Farms' claim?

12. The director of the IRS has been flooded with complaints that people must wait more than 45 minutes before seeing an IRS representative. To determine the validity of these complaints, the IRS randomly selects 400 people entering IRS offices across the country and records the times that they must wait before seeing an IRS representative. The average waiting time for the sample is 55 minutes with a standard deviation of 15 minutes.

a. What is the population being studied?

b. Are the complaints substantiated by the data at $\alpha = 0.10$?

13. NarStor, a computer disk drive manufacturer, claims that the average time to failure for its hard drives is 14,400 hours. You work for a consumer group that has decided to examine this claim. Technicians ran 16 drives continuously for three years. Recently the last drive failed. The time to failure (in hours) are given below.

Time Until Failure (Hours)							
330	620	1870	2410	4620	6396	7822	8102
8309	12,882	14,419	16,092	18,384	20,916	23,812	25,814

a. What is the population being studied?

b. What is the variable being measured?

c. What level of measurement does the variable possess?

d. Conduct a hypothesis test to determine whether there is overwhelming evidence that the average time to failure is less than the manufacturer's claim. Use $\alpha = 0.01$.

e. What assumption did you make in performing the test in part **d.**?

14. Officials in charge of televising an international chess competition in South America want to determine if the average time per move for the top players has remained under five minutes over the last two years. Videos of matches which have been played over the two-year period are reviewed and a random sample of 50 moves are timed. The sample mean is 3.5 minutes with a standard deviation of 1.5 minutes.

a. What is the population under study?

b. Can the officials conclude at $\alpha = 0.05$ that the time per move is still under five minutes?

15. High power experimental engines are being developed by the Stevens Motor Company for use in their new sports coupe. The engineers have calculated the maximum horsepower for the engine to be 600 HP. Sixteen engines are randomly selected for horsepower testing. The sample has an average maximum HP of 620 with a standard deviation 50 HP.

a. Perform an hypothesis test to determine whether the data suggests that the average maximum HP for the experimental engine is significantly different than the maximum horsepower calculated by the engineers? Use a significance level of $\alpha = 0.10$.

b. What assumption did you make in performing the test in part **a.**?

16. The nutrition label for Oriental Spice Sauce states that one package of the sauce has 1190 milligrams of sodium. To determine if the label is accurate the FDA randomly selects two hundred packages of Oriental Spice Sauce and determines the sodium content. The sample has an average of 1167.34 milligrams of sodium per package with a sample standard deviation of 252.94 milligrams.

a. Find the *P*-value for the test of hypothesis that the sodium content is different than the nutrition label states.

b. Is there sufficient evidence to reject the null hypothesis at a significance level of 0.01?

11.4 Exercises

Basic Concepts

1. How can a confidence interval be used to test a hypothesis?
2. Can a confidence interval be used to test a one-sided hypothesis?
3. Describe the difference between statistical significance and practical significance.
4. Give an example of a situation in which results could be statistically significant but not practically significant.

Exercises

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Super Bowl

5. Historically, the average number of points scored by a team during an NFL football game is known to be 19.551. Use the Super Bowl data set and a confidence interval approach to test whether the number of points scored by a team during the Super Bowl is different than 19.551 at $\alpha = 0.05$.

6. AAA Controls makes a switch that is advertised to activate a warning light if the power supplied to a machine reaches 100 volts. A random sample of 250 switches is tested and the mean voltage at which the warning light occurs is 98 volts. The population standard deviation in known to be 3 volts. Using the confidence interval approach, test the hypothesis that the mean voltage activation is different from AAA Controls' claim at the 0.05 level.

7. Researchers studying the effects of diet on growth would like to know if a vegetarian diet affects the height of a child. The researchers randomly selected 12 vegetarian children that were six years old. The average height of the children is 42.5 inches. The average height for all six-year-old children is 45.75 inches with a standard deviation of 3.8 inches.
   a. Using confidence intervals, test to determine whether there is overwhelming evidence at $\alpha = 0.05$ that six-year-old vegetarian children are not the same height as other six-year-old children.
   b. What assumption did you make in performing the test?

8. High-power experimental engines are being developed by the Stevens Motor Company for use in its new sports coupe. The engineers have calculated the maximum horsepower for the engine to be 600 HP. Sixteen engines are randomly selected for horsepower testing. The sample has an average maximum HP of 620 with a standard deviation of 50 HP.
   a. Use the confidence interval approach to determine whether the data suggests that the average maximum HP for the experimental engine is significantly different than the maximum horsepower calculated by the engineers. Use a significance level of $\alpha = 0.01$.
   b. What assumption did you make in performing the test?

9. The nutrition label for Oriental Spice Sauce states that one package of sauce has 1190 milligrams of sodium. To determine if the label is accurate, the FDA randomly selects two hundred packages of Oriental Spice Sauce and determines the sodium content. The sample has an average of 1167.34 milligrams of sodium per package with a sample standard deviation of 252.94 milligrams.

a. Calculate a 99% confidence interval for the mean sodium content in Oriental Spice Sauce.

b. Using the confidence interval approach, is there evidence that the sodium content is different than the nutrition label states?

10. Officials in charge of televising an international chess competition in South America want to determine if the average time per move for the top players has remained under five minutes over the last two years. Data from matches which have been played over the two-year period were collected and a random sample of 50 moves was analyzed. The sample mean is 3.5 minutes with a standard deviation of 1.5 minutes. Using the confidence interval approach, test the hypothesis that the average time per move is different from 5 minutes at a 0.01 significance level.

11. For adults, a cholesterol value under 200 mg/dl is preferred. High cholesterol values over a period of time can result in a stroke or heart attack. A random sample of 10 cholesterol readings for a patient has an average of 204 mg/dl. The population standard deviation of the test is 6 mg/dl. Is there overwhelming evidence to conclude that the patient's cholesterol level is above 200 mg/dl at a 0.05 significance level? Discuss the statistical and practical significance for this problem.

12. A horticulturist working for a large plant nursery is conducting experiments on the growth rate of a new shrub. Based on previous research, the horticulturist feels the average weekly growth rate of the new shrub is 1 cm per week. A random sample of 45 shrubs has an average growth of 0.90 cm per week with a standard deviation of 0.30 cm. Will a test of hypothesis at the 0.05 significance level support the claim that the growth rate is less than 1 cm per week? Discuss the statistical and practical significance for this problem.

13. The director of the IRS has been flooded with complaints that people must wait more than 45 minutes before seeing an IRS representative. To determine the validity of these complaints, the IRS randomly selects 400 people entering IRS offices across the country and records the times which they must wait before seeing an IRS representative. The average waiting time for the sample is 55 minutes with a standard deviation of 15 minutes. Are the complaints substantiated by the data at $\alpha = 0.10$? Discuss the statistical and practical significance for this problem.

14. In 2022, 3.96 billion people used social media, which is more than half the population of the world.[13] The average time spent on social media around the world is 147 minutes. Jennifer's daily average time spent on social media for the last ten days was 132 minutes. Perform a hypothesis test to determine if Jennifer's time spent on social media is significantly less than the average. Use $\alpha = 0.01$. Assume the population standard deviation is 35 minutes. Discuss the statistical and practical significance for this problem.

$\alpha = 0.05$, the null hypothesis is rejected in favor of the alternative, since the test statistic has a *P*-value (0.0392) less than 0.05.

Table 11.5.2 – If *P*-Value = 0.0392

Level of the Test	Reject or Fail to Reject
0.10	Reject
0.05	Reject
0.03	Fail to Reject
0.01	Fail to Reject

11.5 Exercises

Basic Concepts

1. How does testing a hypothesis about a proportion differ from testing a hypothesis about a mean?
2. What is the appropriate test statistic to be used in hypothesis testing of a population proportion?
3. What conditions must be met in order to perform a hypothesis test about a population proportion?
4. How are *P*-values determined for a proportion?

Exercises

5. Determine the critical value(s) of the test statistic for each of the following large sample tests for the population proportion.
   - **a.** Left-tailed test, $\alpha = 0.05$
   - **b.** Right-tailed test, $\alpha = 0.01$
   - **c.** Two-tailed test, $\alpha = 0.10$

6. Determine the critical value(s) of the test statistic for each of the following large sample tests for the population proportion.
   - **a.** Left-tailed test, $\alpha = 0.07$
   - **b.** Right-tailed test, $\alpha = 0.04$
   - **c.** Two-tailed test, $\alpha = 0.09$

7. A commercial airline is concerned about the increase in usage of carry-on luggage. For years, the percentage of passengers with one or more pieces of carry-on luggage has been stable at approximately 68%. The airline recently selected 300 passengers at random and determined that 237 possessed carry-on luggage. Is there overwhelming evidence of an increase in carry-on luggage at a significance level of 0.01?

8. Ordinarily, when a company recruits a technical staff member, about 25% of the applicants are qualified. However, based on the information in 120 recently received resumes, 18 appear to be technically qualified.

a. Is there overwhelming evidence that the percentage of qualified applicants is less than 25%? Test at the 0.05 level.

b. What concerns might you have about the data in this problem?

c. What concerns might you have about creating a hypothesis after examining the data?

9. In a survey on social media usage, a survey was emailed to 7500 randomly selected people. There were 1413 surveys returned.

a. Test that the survey return rate is less than 20% at the 0.05 level.

b. What concerns might you have about the data in this problem?

10. Paper International, Inc. has a large staff of salespeople nationwide. Top officials of the company believe that 75% of their salespeople have met their monthly sales goals by the end of the third week of each month. To investigate this, they randomly select 250 salespeople and examine their sales records at the end of the third week of the current month. One-hundred seventy-five of the 250 salespeople surveyed had already met their monthly sales goals.

a. Does this sample support the belief of the top officials at the company at $\alpha = 0.10$?

b. What concerns might you have about the manner in which the data were collected?

11. Ships arriving in US ports are inspected by customs officials for contaminated cargo. Assume, for a certain port, that 20% of the ships arriving in the previous year contained cargo that was contaminated. A random selection of 50 ships in the current year included five that had contaminated cargo.

a. Does the data suggest that the proportion of ships arriving in the port with contaminated cargoes has decreased in the current year at $\alpha = 0.01$?

b. Do you have any concerns about the sample size? Explain.

c. Do you have any concerns about constructing a hypothesis after looking at the data?

12. Electronic circuit boards are randomly selected each day to determine if any of the boards are defective. A random sample of 300 boards from one day's production has twelve boards that are defective.

a. Based on the data, is there overwhelming evidence that more than 5% of the circuit boards are defective? Test at the $\alpha = 0.10$ level.

b. Do you have any concerns about the sample size? Explain.

13. Loch Ness Fish Farm breeds fish for commercial sale. The fish are kept in breeder tanks until at least 70% of the fish are five inches long at which time they are transferred to outdoor ponds. To determine if it is the appropriate time to transfer the fish, 50 fish are randomly selected and measured. If 33 of the fish are found to be over five inches long, does the sample data suggest that it is the appropriate time to transfer the fish at $\alpha = 0.05$?

14. Digger and Digger, a precious metals mining company, is considering the development of a new mining area. They have a lease on an area which they believe contains gypsum. The area will be profitable to mine if more than 15% of the samples contain more than trace amounts of the mineral. Eighty identical samples are randomly selected and the amount of gypsum is measured. Thirteen samples are observed to have more than trace amounts of the mineral. Based on the sample data, should Digger and Digger conclude that the area will be profitable to mine? Use $\alpha = 0.01$.

15. A socially conscious corporation wants to relocate their headquarters to another part of town. One concern expressed by workers is that their commuting distance will increase. The corporation has decided that if more than 50% of the employees will have to drive farther to the proposed new location, they will cancel the move. In a random sample of 398 employees, 201 indicated that their commuting distance to the new office will be longer. Based on the sample data, should the corporation cancel the move? Use a significance level of 0.01.

16. A production process will normally produce defective parts 0.2% of the time. In a random sample of 1400 parts, three defectives are observed.

- **a.** Is this overwhelming evidence at the 0.05 level to indicate that the defective rate of the process has increased?
- **b.** Compute the *P*-value for the test statistic.
- **c.** Based on the *P*-value, would the decision change at $\alpha = 0.01$?

17. Bombay Charlie's, a fast-food Indian restaurant, is thinking about adding a certain spice to their chicken curry dish to attract more customers. The restaurant manager has decided to add the spice if more than 80% of his customers prefer the taste of the chicken curry with the spice added. Sixty-five customers are randomly selected to participate in a blind taste test. Fifty-four of these customers prefer the chicken curry with the added spice.

- **a.** Find the *P*-value for the hypothesis test that the manager will perform to decide if more than 80% of the customers prefer the taste of the chicken curry with the added spice.
- **b.** Does the data suggest that more than 80% of the customers prefer the curry with the new spice at $\alpha = 0.05$?

18. The news program for KOPE, the local television station, claims to have 40% of the market. A random sample of 500 viewers conducted by an independent testing agency found 192 who claim to watch the KOPE news program on a regular basis.

- **a.** Find the *P*-value for testing the hypothesis that the news program for KOPE does not have at least 40% of the market as it claims.
- **b.** Is there sufficient evidence to reject the hypothesis that KOPE does not have at least 40% of the market at a significance level of 0.05?

19. The length of time that a storm window will last before beginning to leak is of interest to a window manufacturer who wishes to guarantee his windows. He believes that more than 50% of the windows will last at least four years. To research this, 931 windows, which were installed at least four years ago, are randomly selected and checked for leakage. Five hundred of the windows are found to still be leak-free.

a. Find the *P*-value for testing the hypothesis that more than 50% of the windows will be leak-free in four years.

b. Does the sample support the hypothesis that more than 50% of the windows will be leak-free in four years at $\alpha = 0.05$?

20. Some cities, like Berkeley, California, have introduced a "soda tax" to raise money for the city and to reduce the consumption of soda and other sugar-sweetened drinks which can lead to obesity and diabetes. Taxing at the rate of \$0.015 per ounce increases the cost of sugar-sweetened drinks by as much as 20%. Prior to the price increase the percentage of consumers that drank at least one sugar-sweetened drink per day was 48%. Suppose that two months after the price increase, a random sample of consumers was selected to determine changes in consumption of sugar-sweetened drinks. With $\alpha = 0.05$, can we conclude that the price increase was effective in decreasing the consumption of sugar-sweetened drinks if 78 of the 200 consumers drink sugar-sweetened drinks on a daily basis?

21. About 44% of the households in the United States had cable television in 2021.[14] Suppose that a sample of 200 households is selected in 2024 and it is determined that 78 of them have cable television.

a. With $\alpha = 0.05$, can it be concluded that a lower proportion of households in 2024 have cable television as compared with 2021?

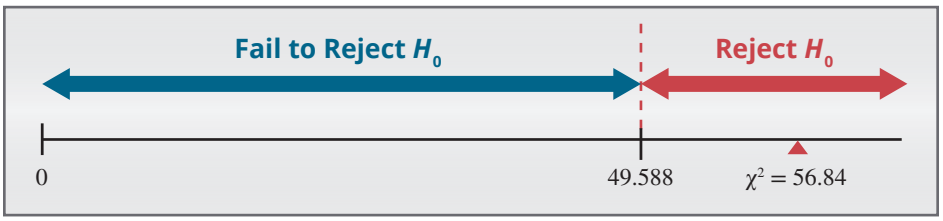
b. In the sample of 200, what is the most number of people who have cable television that would allow the conclusion that a lower proportion of households in 2024 have cable television as compared to 2021?

Data

stat.hawkeslearning.com

Discovering Statistics and Data, Fourth Edition > Data Sets > Super Bowl

22. The winner of the coin toss in a football game has their choice of one of three privileges: deciding which team receives the kickoff, deciding which goal his team will defend, or deferring and deciding in the second half whether to kick or receive the kickoff. Using the Super Bowl data set, determine if winning the coin toss seemed to impact winning the game by looking at the proportion of game winners who also won the coin toss. Test at the $\alpha = 0.05$ level.



Since the P-value of 0.0015 is less than $\alpha = 0.01$, we reach the same conclusion; reject the null hypothesis.

Step 6: State the conclusion in terms of the original question.

There is overwhelming evidence that the process variation exceeds the desired level. Therefore, the manager is willing to shut down the manufacturing process at a significance level of 0.01.

11.6 Exercises

Basic Concepts

1. How does testing a hypothesis about a variance differ from testing a hypothesis about a mean?
2. What is the sampling distribution for $\dfrac{(n-1)s^2}{\sigma^2}$?
3. What assumption must hold to use the chi-square distribution to make inferences about the population variance?
4. True or False: The chi-square distribution is skewed to the right.
5. What is the symbol for a critical value for the chi-square distribution? Describe the meaning of this critical value.
6. Give an example where we would want to calculate a confidence interval for σ^2.

Exercises

7. Determine the critical value(s) of the test statistic for each of the following tests for a population variance where the assumption of normality is satisfied.
   - **a.** Right-tailed test, $\alpha = 0.01$, $n = 20$
   - **b.** Right-tailed test, $\alpha = 0.05$, $n = 24$
   - **c.** Right-tailed test, $\alpha = 0.005$, $n = 5$

8. Determine the critical value(s) of the test statistic for each of the following tests for a population variance where the assumption of normality is satisfied.
   - **a.** Right-tailed test, $\alpha = 0.025$, $n = 18$
   - **b.** Right-tailed test, $\alpha = 0.10$, $n = 24$
   - **c.** Right-tailed test, $\alpha = 0.05$, $n = 41$

9. A bolt manufacturer is very concerned about the consistency with which the machines produce bolts that are $\frac{3}{4}$ inches in diameter. When the manufacturing process is working normally the standard deviation of the bolt diameter is 0.05 inches. A random sample of 30 bolts has an average diameter of 0.25 inches with a standard deviation of 0.07 inches.

a. Can the manufacturer conclude that the standard deviation of bolt diameters is greater than 0.05 inches at $\alpha = 0.05$?

b. What assumption did you make about the diameter of the bolts in performing the test in part **a.**?

10. A drug that is used for treating cancer has potentially dangerous side effects if it is taken in doses that are larger than the required dosage for the treatment. The pharmaceutical company that manufactures the drug must be certain that the standard deviation of the drug content in the tablet is not more than 0.1 mg. Twenty-five tablets are randomly selected and the amount of drug in each tablet is measured. The sample has a mean of 20 mg and a variance of 0.015 mg.

a. Does the data suggest at $\alpha = 0.01$ that the standard deviation of drug content in the tablets is greater than 0.1 mg?

b. What assumption did you make about the amount of drug contained in the tablets in performing the test in part **a.**?

11. The refrigeration coolers in a local grocery store must stay at the same daily temperature with little variance to ensure the quality of the items placed in it. Daily temperatures are measured in degrees Fahrenheit (ºF), and the store manager assumes the standard deviation in daily temperatures is 3.8ºF. The assistant manager claims that the standard deviation is more than 3.8ºF and decides to test the claim using a hypothesis test. For a random sample of 30 days, the assistant manager finds that the standard deviation in the daily temperatures for one cooler is 4.4ºF. At the 0.01 level of significance, does the evidence support the claim that the standard deviation in the daily temperatures for the cooler is more than 3.8ºF?

12.1 Exercises

Basic Concepts

1. What questions are we interested in answering when comparing two population means?
2. What is an independent experimental design?
3. How does the determination of the critical value(s) for a two-sample hypothesis test differ from a one-sample hypothesis test?
4. What conditions are necessary to use the normal distribution to perform a hypothesis test for the difference between two independent population means?
5. Describe how the data guides us to a conclusion in testing a hypothesis about the difference in population means?

Exercises

6. A researcher compares the effectiveness of two different instructional methods for teaching anatomy. A sample of 134 students using Method 1 produces a testing average of 54.6. A sample of 150 students using Method 2 produces a testing average of 53.4. Assume the population standard deviation is known to be 5.1 for Method 1 and 12.47 for Method 2. Determine the 90% confidence interval for the true difference between testing averages for students using Method 1 and students using Method 2.

7. A student researcher compares the ages of cars owned by students and cars owned by faculty at a local state college. A sample of 109 cars owned by students had an average age of 7.81 years. A sample of 126 cars owned by faculty had an average age of 5.93 years. Assume the standard deviation is known to be 2.70 years for age of cars owned by students and 2.18 years for age of cars owned by faculty. Determine the 95% confidence interval for the difference between the true mean ages for cars owned by students and faculty. Let Population 1 be cars owned by students and Population 2 be cars owned by faculty.

8. Red Auerbach, a Hall of Fame coach of the Boston Celtics, is quoted as saying, "You can't teach height."[1] Although the heights of NBA point guards have been increasing over time, this trend may not apply to all positions. To investigate whether there is a difference in the heights of NBA players over time, data from the rosters of the NBA teams during the 1991 and 2021 seasons were compared. The sample mean height of 28 players who played during the 1991 season was 79.65 inches and a sample mean height of 26 players from the 2021 season was 78.81 inches. Height is commonly assumed to follow a normal distribution with a population standard deviation of 3 inches. Let Population 1 be the 1991 season NBA players and Population 2 be the 2021 season NBA players.
   - **a.** Construct a 90% confidence interval for the difference between the true mean heights of NBA players from the 1991 season and the 2021 season.
   - **b.** Does the interval contain the value of zero? Explain what this means in the context of NBA player heights.
   - **c.** What are possible sources of uncertainty about the claim about the difference in heights of NBA players over the past several decades?

9. *Popular Science* (Vol. 242, No. 3) reported the results of a comparison of several popular minivans.[2] One of the features that they compared was the time required to accelerate from 0 to 60 miles per hour in seconds. The Dodge Grand Caravan ES was able to accelerate from 0 to 60 mph in 11.3 seconds, on average. The Volkswagen Eurovan took 16.5 seconds on average to accelerate from 0 to 60 mph. Suppose that 35 minivans of each type were tested and that the population standard deviation of the times required to accelerate from 0 to 60 for each type of minivan is expected to be 4 seconds, based on historical data. Let Population 1 be the acceleration times of the Dodge Grand Caravan ES and Population 2 be the acceleration times of the Volkswagen Eurovan.

a. Calculate a 95% confidence interval for the difference in average acceleration time between the two types of minivans. Interpret the interval.

b. Does the data suggest that there is a significant difference in the time required to accelerate from 0 to 60 between the two types of minivans at $\alpha = 0.05$?

c. What assumptions did you make about the time required to accelerate from 0 to 60 mph in calculating the confidence interval in part **a.** and for performing the test in part **b.**?

10. Determine the critical value(s) of the test statistic for each of the following tests for the comparison of two population means. Assume the population standard deviations are known and $n_1 = n_2 = 40$.

a. Left-tailed test, $\alpha = 0.05$

b. Right-tailed test, $\alpha = 0.10$

c. Two-tailed test, $\alpha = 0.01$

11. A researcher compares two compounds (A and B) used in the manufacture of car tires that are designed to reduce braking distances. The mean braking distance at a speed of 25 mph for tires made with compound A is 61 feet, with a population standard deviation of 8.5. The mean braking distance for tires made with compound B is 66 feet, with a population standard deviation of 14.7. Suppose that a sample of 65 braking tests are performed for each compound. Using these results, test the claim that the braking distance for tires using compound A is shorter than the braking distance when compound B is used. Let μ_1 be the true mean braking distance corresponding to compound A and μ_2 be the true mean braking distance corresponding to compound B. Use the 0.1 level of significance.

12. A medical researcher wants to compare the pulse rates of smokers and non-smokers. He believes that the pulse rate for smokers and non-smokers is different and wants to test this claim at the 0.05 level of significance. A sample of 70 smokers has a mean pulse rate of 68, and a sample of 82 non-smokers has a mean pulse rate of 71. The population standard deviation of the pulse rates is known to be 8 for smokers and 9 for non-smokers. Let μ_1 be the true mean pulse rate for smokers and μ_2 be the true mean pulse rate for non-smokers.

13. A certain test preparation course is designed to improve students' SAT Math scores. The students who took the prep course have a mean SAT Math score of 512, while the students who did not take the prep course have a mean SAT Math score of 504. Assume that the population standard deviation of the SAT Math scores for students who took the prep course is 41.9 and for students who did not take the prep course is 33.4. The SAT Math scores are taken for a sample of 74

students who took the prep course and a sample of 92 students who did not take the prep course. Conduct a hypothesis test of the claim that the SAT Math scores for students who took the prep course is higher than the SAT Math scores for students who did not take the prep course. Let μ_1 be the true mean SAT Math score for students who took the prep course and μ_2 be the true mean SAT Math score for students who did not take the prep course. Use a 0.10 level of significance.

14. The manager of a city bus system is trying to assess commuter use of a particular bus line. He suspects that, on a weekday morning at 8 a.m., more passengers ride that line on the Northbound route than the Southbound route. The manager asks his Northbound driver and his Southbound driver to count how many passengers are on their 8 a.m. weekday routes for two weeks. The resulting data is shown below. Assume that the population standard deviation for passengers on the Northbound route is 2.6 and that the population standard deviation for passengers on the Southbound route is 3.5. Is there sufficient evidence at the 0.05 level of significance to say that, on a weekday morning at 8 a.m., more passengers ride the bus line on the Northbound route than the Southbound route? Assume that both populations are approximately normally distributed. Let passengers on the Northbound route be Population 1 and let passengers on the Southbound route be Population 2.

Northbound	38	32	35	34	31	36	33	37	36	31
Southbound	34	35	30	34	33	32	28	28	25	35

12.2 Inference about Two Population Means: Independent Samples, σ_1 and σ_2 Unknown

In empirical research it is unlikely that population means or variances of interest will be known. It is still possible to make comparisons between two population means if the population means and standard deviations are unknown when the populations are (approximately) normally distributed or the samples are large. In this section we will examine two methods of comparing population means when the variances of the population are unknown. The first method will assume the variances of the populations are unknown but equal, and the second method will assume the variances are unknown and not equal.

Inferences About the Mean of Two Independent Populations, σ_1 and σ_2 Unknown and Assumed Equal

Inference about two means with equal but unknown variances is a common statistical method used to test whether there is a significant difference between the means of two independent populations. This technique is used when we have two independent samples

12.2 Exercises

Basic Concepts

1. **a.** What are the random variables in the test statistics in this section?

b. Why are they random variables?

c. Which sampling distribution do we use in the formulation of the test statistic when comparing two population means with unknown population variances or standard deviations?

d. What are the properties of the distributions referenced in part **c.**?

2. What assumptions are necessary to perform a hypothesis test for the difference between two independent population means when the population variances or standard deviations are unknown?

3. What is the test statistic for a hypothesis test about two population means when the population variances or standard deviations are unknown? How does this statistic differ from the test statistic used when the population standard deviation is known?

4. What is a pooled variance? Why is it used?

Exercises

5. Determine the critical value(s) of the test statistic for each of the following tests for the comparison of two population means. Assume the population standard deviations are unknown but equal, and $n_1 = n_2 = 40$.

a. Left-tailed test, $\alpha = 0.04$

b. Right-tailed test, $\alpha = 0.08$

c. Two-tailed test, $\alpha = 0.02$

6. A luxury car dealer is considering two possible locations for a new auto mall. The rent on the south side of town is cheaper. However, the dealer believes that the average household income is significantly higher on the north side of town. The dealer has decided that he will locate the new auto mall on the north side of town if the results of a study that he commissioned show that the average household income is significantly higher on the north side of town. Let Population 1 be the North Side of town and Population 2 be the South Side of town. The results of the study are as follows.

Income (Thousands of Dollars)			
	n	$\bar{x}$	s
North Side	35	72	10
South Side	40	68	5

a. Calculate a 90% confidence interval for the difference in average income between the north and south sides of town, assuming that the population standard deviations are equal. Interpret the interval.

b. Based on the study, will the auto dealer decide to locate the new auto mall on the north side of town? Use $\alpha = 0.05$.

7. An internal auditor for Tiger Enterprises has been asked to determine if there is a difference in the average amount charged for daily expenses by two top salespeople, Mrs. Ellis and Mr. Ford. The auditor randomly selects 45 days and determines the daily expenses for each of the salespeople. Let Population 1 be the daily expenses of Mrs. Ellis and Population 2 be the daily expenses of Mr. Ford.

Expenses (Dollars)			
	n	$\bar{x}$	s
Mrs. Ellis	45	\$55	\$8
Mr. Ford	45	\$60	\$3

a. Calculate a 95% confidence interval for the difference in the average amounts charged for daily expenses between Mrs. Ellis and Mr. Ford, assuming that the population standard deviations are not equal. Interpret the interval.

b. Based on the survey, can the auditor conclude that there is a difference in the average amounts charged for daily expenses by the two top salespeople? Use $\alpha = 0.05$.

c. Explain how the 95% confidence interval in part **a.** would lead you to make the same decision that was made in part **b.**

8. The military has two different programs for training aircraft personnel. A government regulatory agency has been commissioned to evaluate any differences that may exist between the two programs. The agency administers standardized tests to randomly selected groups of students from the two programs. The results of the tests for the students in each of the programs are as follows. Let Population 1 be the test results of students in training program A and Population 2 be the test results of students in training program B.

Military Training Programs			
	n	$\bar{x}$	s
Program A	50	85	10
Program B	55	87	9

a. Calculate a 99% confidence interval for the difference between the average scores of the two military programs, assuming that the population standard deviations are equal. Interpret the interval.

b. Can the agency conclude that there is a difference in the average test scores of students in the two programs? Use $\alpha = 0.01$.

9. Tom Sealack, an electrical engineer with the Navy, has been asked to determine if a new battery that has been offered to the Navy (at a reduced price) has a shorter average life than the battery they are currently using. He randomly selects batteries of each type and allows them to run continuously so that he can measure the time until failure for each battery. The results of the test are as follows. Assume that the population standard deviations are equal. Let Population 1 be the battery life of the new battery and Population 2 be the battery life of the old battery.

Battery Life (Hours)			
	n	$\bar{x}$	s
New Battery	35	700	30
Old Battery	35	710	35

a. Does the data suggest at $\alpha = 0.10$ that the time until failure for the new battery is significantly less than the time until failure for the old battery?

b. Calculate the *P*-value for the test in **a.**

c. Based on the *P*-value, would the decision change at $\alpha = 0.05$?

10. The City Bank believes that checking account balances are significantly larger for customers who are aged 40 to 49 than those who are aged 30 to 39. To investigate this belief, they randomly select customers from each age group and determine the average daily account balance for each customer for the current month. The results of the study are as follows. Assume that the population standard deviations are not equal. Let Population 1 be the 30 - 39 age group and Population 2 be the 40 - 49 age group.

Checking Account Balances (Dollars)			
Age Group	***n***	$\bar{x}$	***s***
30 – 39	200	\$2500	\$550
40 – 49	150	\$3500	\$950

a. Does the data suggest at $\alpha = 0.05$ that the average daily account balances are significantly higher for the 40 to 49 age group than the 30 to 39 age group?

b. Calculate the *P*-value for the test in **a.**

c. Based on the *P*-value, would the decision change at $\alpha = 0.10$?

11. Determine the critical value(s) of the test statistic for each of the following tests for the comparison of two population means where the assumptions of normality have been satisfied and the population standard deviations are unknown but equal.

a. Left-tailed test, $\alpha = 0.05$, $n_1 = 10$, $n_2 = 15$

b. Right-tailed test, $\alpha = 0.10$, $n_1 = 8$, $n_2 = 12$

c. Two-tailed test, $\alpha = 0.01$, $n_1 = 5$, $n_2 = 7$

12. Determine the critical value(s) of the test statistic for each of the following tests for the comparison of two population means where the assumptions of normality have been satisfied and the population standard deviations are unknown and unequal.

a. Left-tailed test, $\alpha = 0.025$, $n_1 = 13$, $n_2 = 25$

b. Right-tailed test, $\alpha = 0.005$, $n_1 = 7$, $n_2 = 18$

c. Two-tailed test, $\alpha = 0.10$, $n_1 = 15$, $n_2 = 15$

13. A cereal manufacturer has advertised that its product, Fiber Oat Flakes, has a lower fat content than its competitor, Bran Flakes Plus. Because of complaints from the manufacturers of Bran Flakes Plus, the FDA has decided to test the claim that Fiber Oat Flakes has a lower average fat content than Bran Flakes Plus. Several boxes of each cereal are selected and the fat content per serving is measured. The results of the study are as follows. Assume that the population variances are approximately equal and that the assumptions of normality have been satisfied. Let Population 1 be the fat content of Fiber Oat Flakes and Population 2 be the fat content of Bran Flakes Plus.

Fat Content (Grams)			
	n	$\bar{x}$	s
Fiber Oat Flakes	16	5	1
Bran Flakes Plus	15	6	2

a. Calculate a 90% confidence interval for the difference in average fat content between Fiber Oat Flakes and Bran Flakes Plus. Interpret the interval.

b. Does the study performed by the FDA substantiate the claim made by the manufacturer of Fiber Oat Flakes at $\alpha = 0.10$?

c. What assumptions must be made in order to calculate the confidence interval in part **a.** and perform the hypothesis test in part **b.**?

14. A large construction company would like to expand its operations into a new geographic area. The company has narrowed the choice of locations down to two cities. A major consideration in deciding between the two cities will be the average hourly wage they must pay for general laborers. The company randomly selects laborers from each city and determines their hourly wage with the following results. Assume that the population variances are approximately equal and that the assumptions of normality have been satisfied. Let Population 1 be City A hourly wages and Population 2 be City B hourly wages.

Hourly Wages (Dollars)			
	n	$\bar{x}$	s
City A	20	\$17	\$3
City B	20	\$14	\$2

a. Calculate a 99% confidence interval for the difference in average hourly wage between City A and City B. Interpret the interval.

b. Does the data indicate that there is a significant difference in hourly wages at $\alpha = 0.05$?

c. Calculate the P-value for the test performed in part **b.**

d. What assumptions must be made in order to calculate the confidence interval in part **a.** and perform the hypothesis test in part **b.**?

15. A Hollywood studio believes that a movie that is considered a drama will draw a larger crowd on average than a movie that is a comedy. To test this theory, the studio randomly selects several movies that are classified as dramas and several movies that are classified as comedies and determines the box office revenue for each movie. The results of the survey are as follows. Assume that the population variances are not equal and that the assumptions of normality have been satisfied. Let Population 1 be drama box office revenues and Population 2 be comedy box office revenues.

Box Office Revenues (Millions of Dollars)			
	n	$\bar{x}$	s
Drama	15	180	50
Comedy	13	150	30

a. Calculate a 95% confidence interval for the difference in average revenue at the box office for drama and comedy movies. Interpret the interval.

b. Does the data substantiate the studio's belief that dramas will draw a larger crowd on average than comedies at $\alpha = 0.01$?

c. Calculate the *P*-value for the test you conducted in part **b.**

d. What assumptions must be made in order to calculate the confidence interval in part **a.** and to perform the hypothesis test in part **b.**?

16. *Consumer Magazine* is reviewing the top of the line amplifiers produced by two major stereo manufacturers. One of the most important qualities of the amplifiers is the maximum power output. Brand A has redone their internal design and claims to have a higher maximum power level than Brand B. To test this claim, *Consumer Magazine* randomly selects amplifiers from each brand and determines the maximum power output. The results of the test are as follows. Assume that the population variances are approximately equal and that the assumptions of normality have been satisfied. Let Population 1 be Brand A amplifiers and Population 2 be Brand B amplifiers.

Amplifier Power Output (Watts)			
	n	$\bar{x}$	s
Brand A	12	800	25
Brand B	10	780	25

a. What assumptions must be made in order to perform the hypothesis test?

b. Does the data substantiate the claim that the Brand A amplifier has a higher average maximum power output than Brand B at $\alpha = 0.05$?

17. The State Environmental Board wants to compare pollution levels in two of its major cities. Sunshine City thrives on the tourist industry and Service City thrives on the service industry. The environmental board randomly selects several areas within the cities and measures the pollution levels in parts per million with the following results. Assume that the population variances are approximately equal and that the assumptions of normality have been satisfied. Let Population 1 be Sunshine City and Population 2 be Service City.

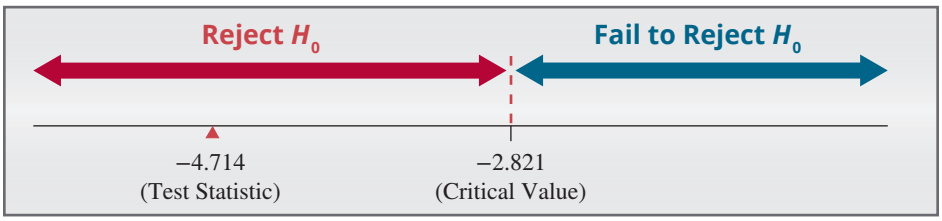
Pollution Levels (ppm)			
	n	$\bar{x}$	s
Sunshine City	15	8.5	0.57
Service City	10	7.9	0.50

a. What assumptions must be made in order to perform a hypothesis test for the difference between these two population means?

b. Will the State Environmental Board conclude at $\alpha = 0.01$ that Service City has a lower pollution level on average than Sunshine City?

c. Repeat part **b.**, assuming that the population variances are not equal.

d. Compare the results of part **b.** and part **c.**



Since the P-value of 0.00055 is much smaller than $\alpha = 0.01$, the null hypothesis is rejected.

Step 6: State the conclusion in terms of the original problem.

There is sufficient evidence for the researcher to conclude at $\alpha = 0.01$ that the average response time is significantly higher for those participants who have drunk one ounce of 100-proof alcohol than those who have not.

12.3 Exercises

Basic Concepts

1. Describe the differences between an independent experimental design and a paired design.
2. What are the assumptions for a paired difference experimental design?
3. What is the appropriate statistical measure to use when performing a hypothesis test about a paired difference experiment?
4. How does the hypothesis testing procedure for a paired difference experiment differ from that of a two-sample t-test?
5. What is the test statistic used in a paired difference hypothesis test?

Exercises

6. Determine the critical value(s) of the test statistic for each of the following paired difference tests (assume the differences have an approximately normal distribution).
   - **a.** Left-tailed test, $\alpha = 0.01$, $n = 15$
   - **b.** Right-tailed test, $\alpha = 0.10$, $n = 20$
   - **c.** Two-tailed test, $\alpha = 0.05$, $n = 8$

7. Determine the critical value(s) of the test statistic for each of the following paired difference tests (assume the differences have an approximately normal distribution).
   - **a.** Left-tailed test, $\alpha = 0.005$, $n = 12$
   - **b.** Right-tailed test, $\alpha = 0.025$, $n = 5$
   - **c.** Two-tailed test, $\alpha = 0.10$, $n = 25$

8. Given that most textbooks can now be purchased online, one wonders if students can save money by comparison shopping for textbooks at online retailers and at their local bookstores. To investigate, students at a university randomly sampled 25 textbooks on the shelves of their local bookstores. The students then found the "best" available price for the same textbooks via online retailers. The prices for the textbooks are listed in the following table. Let the difference d = bookstore price - online retailer price.

a. Is a paired design appropriate for the above study? Explain.

b. What assumption must be made in order to perform the test of hypothesis?

c. Does the data appear to satisfy the assumption described in part **b.**? Why or why not?

d. Based on the data, is it less expensive for the students to purchase textbooks from the online retailers than from local bookstores? Use $\alpha = 0.01$.

e. Calculate a 99% confidence interval for the mean difference in cost between the bookstores and the online retailers. Interpret the interval.

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Textbook Prices

Textbook Prices

	Price ($)	
Textbook	**Bookstore**	**Online Retailer**
1	70	60
2	38	36
3	88	89
4	165	149
5	80	136
6	103	95
7	42	50
8	98	111
9	89	65
10	97	86
11	140	130
12	40	30
13	175	150
14	85	75
15	100	85
16	68	62
17	67	69
18	140	142
19	49	40
20	149	127
21	126	130
22	92	93
23	144	129
24	98	84
25	40	52

9. The management for a large grocery store chain would like to determine if a new scanner will enable cashiers to process a larger number of items on average than the scanner they are currently using. Seven cashiers are randomly selected, and the number of grocery items they can process in three minutes is measured for both the old scanner and the new scanner. The results of the test are as follows. Let the difference d = number of items processed by the old scanner - number of items processed by the new scanner.

Number of Grocery Items Processed in Three Minutes

Cashier	1	2	3	4	5	6	7
Old Scanner	60	70	55	75	62	52	58
New Scanner	65	71	55	75	65	57	57

a. Is a paired design appropriate for the above experiment? Explain.

b. What assumption must be made in order to perform the test of hypothesis?

c. Does the data appear to satisfy the assumption described in part **b.**? Why or why not?

d. Calculate a 95% confidence interval for the mean difference between the number of items processed using the old scanner and the new scanner. Interpret this interval.

e. Can the management conclude that the new scanner will allow cashiers to process a significantly larger number of items on average than the old scanner at $\alpha = 0.05$?

10. An auto dealer is marketing two different models of a high-end sedan. Since customers are particularly interested in the safety features of the sedans, the dealer would like to determine if there is a difference in the braking distance (the number of feet required to go from 60 mph to 0 mph) of the two sedans. Six drivers are randomly selected and asked to participate in a test to measure the braking distance for both models. Each driver is asked to drive both models and brake once they have reached exactly 60 mph. The distance required to come to a complete halt is then measured in feet. The results of the test are as follows. Let the difference d = braking distance for Model A - braking distance for Model B.

Braking Distance of High-End Sedans (Feet)						
Driver	1	2	3	4	5	6
Model A	150	145	160	155	152	153
Model B	152	146	160	157	154	155

a. Is a paired design appropriate for the above experiment? Explain.

b. What assumption must be made in order to perform the test of hypothesis?

c. Does the data appear to satisfy the assumption described in part **b.**? Why or why not?

d. Calculate a 90% confidence interval for the average difference between braking distances for Model A and Model B. Interpret the interval.

e. Can the auto dealer conclude that there is a significant difference in the braking distances of the two models of high-end sedans? Use $\alpha = 0.10$.

11. A sleep disorder specialist wants to test the effectiveness of a new drug that is reported to increase the number of hours of sleep patients get during the night. To do so, the specialist randomly selects eight patients and records the number of hours of sleep each gets with and without the new drug. The results of the two-night study are listed in the table below.

Patient	1	2	3	4	5	6	7	8
Hours of sleep without the drug	5.7	6.2	5.1	5.6	4.8	6.8	5.4	5.9
Hours of sleep with the drug	6.0	5.8	5.4	6.1	5.3	6.5	5.7	6.2

Let d = (hours of sleep with the new drug) – (hours of sleep without the new drug). Assume that the hours of sleep are normally distributed for the population of patients both before and after taking the new drug. Using this data, determine the 95% confidence interval for the true difference in hours of sleep between the patients when using and when not using the new drug.

12. A psychology graduate student wants to test the claim that there is a significant IQ difference between husbands and wives. To test this claim, she measures the IQs of 8 married couples using a standard IQ test. The results of the IQ tests are listed in the following table.

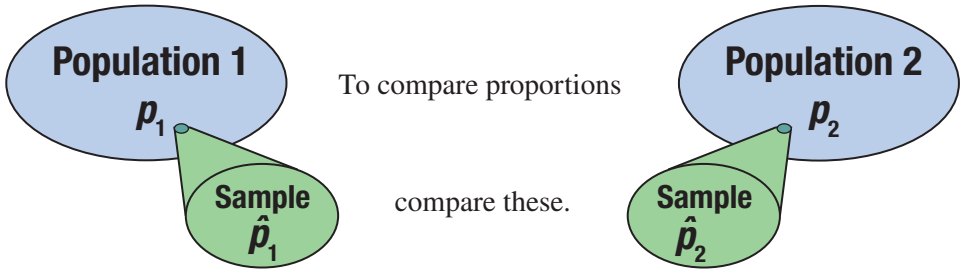
IQs of Married Couples								
Husband	109	112	102	130	119	106	121	116
Wife	105	110	109	124	123	111	115	120

Using a 0.10 level of significance, test the claim that there is a significant difference between the IQs of husbands and wives. Assume that the population distribution of the paired differences is approximately normal. Let the group "Husband" be Population 1 and let the group "Wife" be Population 2.

12.4 Inference about Two Population Proportions

Techniques are developed in this section for comparing two population proportions. A methodology for comparing two population proportions is particularly useful because proportions are among the few measures that can be used for summarizing categorical data. For a more extensive treatment of comparisons for categorical data, see Chapter 16.

There are many situations where comparing two population proportions may be of interest. For example, a sociologist may be interested in comparing the proportion of females who believe that it is okay to cry in public to the proportion of males who think it is okay to cry in public. A marketing manager may be interested in comparing the proportion of customers who favor Product A to the proportion of customers who favor Product B.



In order to perform a comparison of two population proportions, the assumptions outlined below must be met.

12.4 Exercises

Basic Concepts

1. Why is comparing two population proportions particularly useful?
2. Give two examples of situations in which someone would be interested in comparing population proportions.
3. What assumptions are necessary to perform a hypothesis test for the difference between two population proportions?
4. Which sampling distribution is used in a two-sample test of hypothesis about population proportions? What are the characteristics of this sampling distribution?
5. What is the test statistic that is used when comparing two population proportions?
6. True or False: In order to use the specified test statistic, the hypothesized difference in the null hypothesis between the two population proportions must be zero.

Exercises

7. Determine the critical value(s) of the test statistic for each of the following tests for the comparison of two population proportions. Assume that the samples are large enough that the normality assumption is met.
   a. Left-tailed test, $\alpha = 0.01$
   b. Right-tailed test, $\alpha = 0.05$
   c. Two-tailed test, $\alpha = 0.10$
8. Determine the critical value(s) of the test statistic for each of the following tests for the comparison of two population proportions. Assume that the samples are large enough that the normality assumption is met.
   a. Left-tailed test, $\alpha = 0.025$
   b. Right-tailed test, $\alpha = 0.02$
   c. Two-tailed test, $\alpha = 0.04$
9. A fundraiser believes that individuals over the age of 40 are more likely to say "Yes" when asked to donate to a worthy cause than individuals who are between the ages of 25 and 40. To test this theory, she randomly selects 200 individuals who are between the ages of 25 and 40 and 190 individuals over the age of 40 and asks for donations to the same cause. Let Population 1 be individuals between 25-40 years old and Population 2 be individuals over 40 years old. The results of the survey are as follows.

Fund-Raiser Survey		
Age	**Number Surveyed**	**# of "Yes" Responses**
25-40 years	200	16
Over 40 years	190	21

   a. Are the sample sizes large enough such that a hypothesis test for the difference between two population proportions may be performed? If so, does the data substantiate the fund-raiser's theory at $\alpha = 0.10$?

b. Calculate the *P*-value for the test and interpret its meaning.

c. Calculate a 95% confidence interval for the difference in the proportion of individuals who are between the ages of 25 and 40 and individuals who are over the age of 40 who would most likely donate to a worthy cause. Interpret the interval.

10. A poll is conducted to determine if US citizens think that there should be a national health care system in the U.S. The results of the poll were as follows: 69% of the 300 women surveyed and 64% of the 250 men surveyed think that there should be a national health care system in the U.S. Are the sample sizes large enough such that a hypothesis test for the difference between two population proportions may be performed? If so, is there sufficient evidence to conclude at $\alpha = 0.05$ that men and women feel differently about this issue? Let Population 1 be the women surveyed in the poll and Population 2 be the men surveyed in the poll.

11. A manufacturer is comparing shipments of machine parts from two suppliers. The parts from Supplier A are less expensive; however, the manufacturer is concerned that the parts may be of a lower quality than those from Supplier B. The manufacturer has decided that he will purchase his supplies from Supplier A unless he can show that the proportion of defective parts is significantly higher for Supplier A than for Supplier B. He randomly selects parts from each supplier and inspects them for defects. The results are as follows. Determine whether the sample sizes are large enough such that inferences about the difference between the population proportions can be made. If so, which supplier will the manufacturer choose at $\alpha = 0.05$? Explain. Let Population 1 be Supplier A parts and Population 2 be Supplier B parts.

Number of Defective Parts

	Number Surveyed	Number of Defective Parts
Supplier A	550	11
Supplier B	700	13

12. Suppose you have recently become interested in photography and are shopping on Amazon for a digital single-lens reflex (DSLR) camera. You've narrowed your choice down to two cameras and are leaning towards purchasing the Nikon, unless the Canon has a significantly higher proportion of 5-Star ratings. Given the Amazon rating distributions for the two cameras, which will you choose at $\alpha = 0.05$? Explain. Let the Nikon D3500 reviews be Population 1 and the Canon Rebel T7 reviews be Population 2.

Amazon Ratings

Stars	Nikon D3500 Number of Reviews	Canon Rebel T7 Number of Reviews
5-Star	354	266
4-Star	53	33
3-Star	12	10
2-Star	9	4
1-Star	14	9

Since the test statistic value of 2.7778 does not fall in the rejection region, we fail to reject the null hypothesis. Likewise, since the P-value of 0.0564 is greater than the significance level of 0.05, we fail to reject the null hypothesis.

Step 6: State the conclusion in terms of the original problem.

We conclude that there is insufficient evidence at a 0.05 significance level to indicate that the variances in revenue between dramas and comedies are significantly different.

Please note that the methods presented in this section work very poorly when the normality assumption is violated. It is very important to validate the assumption of normality before developing confidence intervals or performing hypothesis tests on the ratio of the variances.

12.5 Exercises

Basic Concepts

1. Give two examples of situations in which someone would be interested in comparing population variances (or standard deviations).
2. What assumptions are necessary to perform a hypothesis test for two population variances?
3. What is the test statistic that is used when comparing two population variances?
4. What are the parameters of the distribution of the test statistic in the previous question?

Exercises

5. Find a point on the F-distribution with 7 numerator degrees of freedom and 22 denominator degrees of freedom such that the following area lies to the right of this value.

   a. $\alpha = 0.100$ **c.** $\alpha = 0.025$

   b. $\alpha = 0.050$ **d.** $\alpha = 0.010$

6. Find a point on the F-distribution with 30 numerator degrees of freedom and 8 denominator degrees of freedom such that the following area lies to the right of this value.

   a. $\alpha = 0.100$ **c.** $\alpha = 0.025$

   b. $\alpha = 0.050$ **d.** $\alpha = 0.010$

7. Find $F_{0.025}$ for an F-distribution with the following parameters.

   a. 1 numerator degree of freedom, 25 denominator degrees of freedom

   b. 6 numerator degrees of freedom, 11 denominator degrees of freedom

   c. 8 numerator degrees of freedom, 40 denominator degrees of freedom

   d. 3 numerator degrees of freedom, 18 denominator degrees of freedom

8. Find $F_{0.010}$ for an F-distribution with the following parameters.

a. 15 numerator degrees of freedom, 19 denominator degrees of freedom

b. 10 numerator degrees of freedom, 29 denominator degrees of freedom

c. 60 numerator degrees of freedom, 24 denominator degrees of freedom

d. 12 numerator degrees of freedom, 21 denominator degrees of freedom

9. State the null and alternative hypotheses for each scenario.

a. A professor believes that the variance of SAT scores of honor students is less than that of all students who take the SAT. Let σ_1^2 represent the population variance for honor students.

b. A quality control inspector believes that the variance in the diameters of soda cans produced by Machine 1 is greater than the variance in the diameters of soda cans produced by Machine 2. Let σ_1^2 represent the population variance for Machine 1.

10. Calculate the test statistic for a hypothesis test for two population variances using the given information. Assume that both population distributions are approximately normal.

$$n_1 = 4, \quad s_1^2 = 0.961, \quad n_2 = 6, \quad s_2^2 = 0.899$$

11. State the critical value(s) of the test statistic, and determine the rejection region for the hypothesis test for the two population variances using the given information. Then give the appropriate conclusion for the hypothesis test. Assume that both population distributions are approximately normal.

a. $n_1 = 14, \quad s_1^2 = 3.152, \quad n_2 = 11, \quad s_2^2 = 9.300, \quad H_a: \sigma_1^2 < \sigma_2^2, \quad \alpha = 0.05$

b. $n_1 = 12, \quad s_1^2 = 1893, \quad n_2 = 26, \quad s_2^2 = 1066, \quad H_a: \sigma_1^2 > \sigma_2^2, \quad \alpha = 0.01$

c. $n_1 = 20, \quad s_1^2 = 27.08, \quad n_2 = 29, \quad s_2^2 = 11.77, \quad H_a: \sigma_1^2 \neq \sigma_2^2, \quad \alpha = 0.05$

For exercises 12-16, complete the following steps. Assume that both population distributions are approximately normal in each scenario.

a. State the null and alternative hypotheses.

b. Determine which distribution to use for the test statistic and state the level of significance.

c. Calculate the test statistic.

d. Draw a conclusion and interpret the decision.

12. A golf pro believes that the variances of his driving distances are different for different brands of golf balls. In particular, he believes that his driving distances, measured in yards, have a smaller variance when he uses Titleist golf balls than when he uses a generic store brand. He hits 10 Titleist golf balls and records a sample variance of 201.65. He hits 10 generic golf balls and records a sample variance of 364.57. Test the golf pro's claim using a 0.05 level of significance. Assume the samples are from populations that are approximately normally distributed. Does the evidence support the golf pro's claim? Let Population 1 be the Titleist golf balls and Population 2 be the generic golf balls.

13. A quality control inspector believes that the variance in the diameters of soda cans, measured in millimeters, is greater for soda cans produced by Machine A than for soda cans produced by Machine B. The sample variance of a random sample of 15 soda cans from Machine A is 2.788. The sample variance for a random sample of 17 soda cans from Machine B is 1.982. Test the inspector's claim using a 0.10 level of significance. Assume the samples are from populations that are approximately normally distributed. Does the evidence support the inspector's claim? Let Population 1 be the soda cans produced on Machine A and Population 2 be the soda cans produced on Machine B.

14. A medical researcher believes that the variance of total cholesterol levels in men is greater than the variance of total cholesterol levels in women. The sample variance for a random sample of 8 men's cholesterol levels, measured in mg/dL, is 277. The sample variance for a random sample of 7 women is 89. Test the researcher's claim using a 0.10 level of significance. Assume the samples are from populations that are approximately normally distributed. Does the evidence support the researcher's belief? Let Population 1 be men's cholesterol levels and Population 2 be women's cholesterol levels.

15. A basketball coach believes that the variance of the heights of adult male basketball players is different from the variance of heights for the general population of men. The sample variance of heights, measured in inches, for a random sample of 12 basketball players is 24.76. The sample variance for a random sample of 13 other men is 25.87. Test the coach's claim using a 0.01 level of significance. Assume the samples are from populations that are approximately normally distributed. Does the evidence support the coach's claim? Let Population 1 be male basketball player's heights and Population 2 be the heights of the other men.

16. One study claims that the variance in the resting heart rates of smokers is different than the variance in the resting heart rates of nonsmokers. A medical student decides to test this claim. The sample variance of resting heart rates, measured in beats per minute, for a random sample of 5 smokers is 545.1. The sample variance for a random sample of 5 nonsmokers is 103.7. Test the study's claim using a 0.01 level of significance. Assume the samples are from populations that are approximately normally distributed. Does the evidence support the study's claim? Let Population 1 be smoker's heart rates and Population 2 be nonsmokers heart rates.

Note

See page 242 for a quick review of how to obtain the sample estimates for the linear regression model. The estimates can also be obtained using technology.

In order to perform inference on the linear model, some assumptions about the nature of the error terms are required.

Assumptions about the Error Term in the Linear Model

1. The ε_i are presumed to be normally distributed with a mean of 0 and a variance of σ_e^2.
2. The ε_i are presumed to be independent of each other.

PROPERTIES

With the addition of the error term, the model's parameters are β_0, β_1, and σ_e^2. The estimation of these quantities was discussed in Chapter 5. The actual verification of these assumptions cannot be made prior to a regression analysis, but they can be validated by doing an analysis of the **residuals** (errors). A residual analysis is beyond the scope of this book, therefore, we will assume that the error terms satisfy the necessary assumptions in order to proceed with inference in the regression analysis.

In addition to the formal assumptions stated above, a linear model should be used to fit data that appears to be reasonably linearly related. Because of the wide availability of computer programs that calculate least squares estimates, you will not need to manually calculate estimates very often.

Technology

The instructions for calculating the coefficients for the simple linear regression model using various technologies can be found on stat.hawkeslearning.com under **Discovering Statistics and Data, Fourth Edition > Technology Instructions > Regression > Simple Linear Regression**.

13.1 Exercises

Basic Concepts

1. Why is an error term incorporated in the simple linear model?

2. What does the error term represent?

3. What assumptions are made about the error term in the simple linear model?

4. What are the parameters of the simple linear regression model? Identify their estimates from the sample.

Exercises

5. Consider the following data and regression output relating a student's grade to the number of absences from class.

a. Determine the independent and dependent variables and write the equation of the model we desire to estimate.

b. Determine the estimated model (regression equation) for predicting a student's grade based on the number of absences from class.

Class Absences and Grade

Number of Absences	Grade
3	3.9
5	3.8
6	2.9
6	2.7
6	2.4
7	2.3
8	1.9

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.917589
R Square	0.84197
Adjusted R Square	0.810364
Standard Error	0.329598
Observations	7

ANOVA		
	df	*SS*
Regression	1	2.89396978
Residual	5	0.543173077
Total	6	3.437142857

	Coefficients	*Standard Error*	*t Stat*
Intercept	5.427885	0.51610437	10.51703
Number of Absences	–0.44135	0.085509913	–5.16134

c. Determine the mean square error (MSE). (See Section 5.3.)

6. The following table of data gives the weeks of gestation and corresponding birth weights for a sample of ten babies.

Gestation Period and Birth Weight	
Weeks of Gestation	**Birth Weight**
34	5.9
34	5.7
35	6.2
36	6.6
36	6.8
37	7.0
38	7.2
38	7.5
39	8.0
40	8.2

a. Determine the independent and dependent variables and write the equation of the model we desire to estimate.

b. Determine the estimated model (regression equation) for predicting a baby's birth weight based on the number of weeks of gestation.

c. Determine the mean square error (MSE). (See Section 5.3.)

d. Determine the coefficient of determination and explain its meaning in terms of the problem.

Data

The data set can be found on stat.hawkeslearning.com under **Discovering Statistics and Data, Fourth Edition > Data Sets > Global Statistics by Country**.

7. Use the Global Statistics by Country data set to answer the following questions.

a. Determine the regression equation for predicting birth rate based on female literacy rate.

b. Determine the coefficient of determination and explain its meaning in terms of the problem.

c. Find the predicted birth rate for a female literacy rate of 70.

d. Find the predicted birth rate for a female literacy rate of 90.

e. As the female literacy rate rises what happens to the birth rate? Does this make sense? Explain your answer.

13.2 Inference Concerning β_1

Since β_1 specifies the rate of change between x and y, in most linear models the parameter of interest is β_1. Two inferential techniques are useful in evaluating the estimate of β_1. Confidence intervals, similar in structure to those used for means and proportions, will be developed. In addition, a hypothesis testing procedure will be presented to test whether β_1 is equal to some particular value.

The Confidence Interval for β_1

Developing a confidence interval for β_1 requires thinking about the estimate b_1 as a random variable. Each random sample from the population will produce different data and hence different least squares estimates of b_0 and b_1. The confidence interval will serve two purposes, to place bounds on the location of β_1 as well as to provide information about the quality of the point estimate b_1. The form of the confidence interval is familiar.

$$\text{Sample estimate of parameter} \pm \left(\begin{array}{c}\text{A certain number of standard deviations units depending on the desired confidence}\end{array}\right) \cdot \left(\begin{array}{c}\text{The standard deviation of the sample estimate}\end{array}\right)$$

The **sample estimate** of β_1 is b_1. The variance of b_1 is given by

$$\sigma_{b_1}^2 = \frac{\sigma_e^2}{\sum\left(x_i - \bar{x}\right)^2}$$

but like all population measurements, $\sigma_{b_1}^2$ usually has to be estimated from the data. The sample estimate of the variance of b_1 is given by

$$s_{b_1}^2 = \frac{s_e^2}{\sum\left(x_i - \bar{x}\right)^2}.$$

The only difference in the computation of $\sigma_{b_1}^2$ and $s_{b_1}^2$ is the replacement of the population variance of the error terms, σ_e^2, with the corresponding sample statistic, s_e^2. The **standard deviation (standard error) of the sample estimate** b_1 is

$$s_{b_1} = \sqrt{\frac{s_e^2}{\sum\left(x_i - \bar{x}\right)^2}}.$$

Step 6: State the conclusion in terms of the original problem.

There is overwhelming evidence at the 0.05 level that $H_a : \beta_1 \neq 0$. This implies that it is reasonable to believe (at the 0.05 level) that there is a linear relationship between the age and the price of a Jeep Cherokee. In fact, there appears to be a negative linear relationship between the age and the price of a Jeep Cherokee. However, our hypothesis test did not address the issue of a *negative* relationship, so we cannot make this conclusion.

If a data analyst feels that the assumptions of the simple linear model have been met and decides to make an inference about the model, the *P*-value of b_1 will be one of the first pieces of computer output that will be examined. The analyst will also look at the value of R^2 (R Square in the Excel output) to see what proportion of the variation in the data is explained by the regression model (see Section 5.3).

Thus far, the focus in this chapter has been inference on β_1. What about β_0? Since β_0 is merely a constant term, in most problems its value is not of great concern. However, if a confidence interval or test of hypothesis is needed, the methods used would be virtually identical to those presented for analyzing β_1.

13.2 Exercises

Basic Concepts

1. Identify two purposes that a confidence interval for β_1 serves.
2. What is the formula for the $100(1-\alpha)\%$ confidence interval for β_1?
3. What are the three pieces of information needed to calculate a confidence interval for β_1?
4. A 99% confidence interval for β_1 is found to be $(5.6,\ 10.2)$. Give two interpretations of this interval.
5. For the confidence interval given in the previous question, what is b_1, the sample estimate for β_1?
6. If there is no linear relationship between two variables, what is the value of β_1? Explain.
7. What is the test statistic for testing the hypothesis that $\beta_1 \neq 0$? Describe how this test statistic is similar to other test statistics used in hypothesis testing.
8. What are the degrees of freedom for the test statistic in the previous question?
9. Can we make inferences about β_0? Why are we more interested in inferences about β_1?
10. Explain why the *P*-value corresponding to b_1 is one of the first values examined by data analysts.

Exercises

11. Consider the random sample of data in the following table regarding the age of a particular model of car and the asking price for that car.

Car Data			
Age (Years)	**Asking Price ($)**	**Age (Years)**	**Asking Price ($)**
1	27,288	4	18,998
1	25,984	5	18,800
2	24,858	5	18,500
2	25,551	6	16,897
3	20,199	6	17,997

a. Draw a scatterplot of the data. Describe the relationship you observe in the scatterplot.

b. Using statistical software, estimate the simple linear model relating age to asking price .

c. What is the standard error of b_1?

d. Find a 99% confidence interval for β_1.

e. Interpret the confidence interval found in part **d.**

12. An economist is studying the relationship between income and IRA contributions. He has randomly selected eight subjects and obtained annual income and IRA contribution data from them. He wishes to predict the amount of money contributed to an IRA based on annual income.

Income and IRA Contributions							
Annual Income (Thousands of Dollars)	56	50	66	75	84	70	92
IRA Contribution (Thousands of Dollars)	0.3	0	1.2	1.8	3.3	2.2	5.2

a. Draw a scatterplot of the data. Describe the relationship that you observe between income and IRA contribution.

b. Estimate the parameters of the following model using statistical software.

$$IRA\ Contribution = \beta_0 + \beta_1 Income + \varepsilon_i$$

c. Calculate and interpret a 95% confidence interval for β_1.

d. What assumptions are being made about the error term in the construction of the confidence interval for β_1?

13. Consider the following summary output, which was generated from a random sample of 8 employees relating age to annual salary.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.732431223
R Square	0.536455496
Adjusted R Square	0.459198079
Standard Error	15.60374155
Observations	8

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	1690.639497	1690.639	6.943741
Residual	6	1460.860503	243.4768	
Total	7	3151.5		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-2.132440745	20.99597109	-0.10156	0.922412
Age	1.564320608	0.593648001	2.635098	0.038794

a. What is the estimated regression equation?

b. Is there evidence of a linear relationship between age and salary at the 0.05 level?

c. Does the decision in part **b.** change at the 0.01 level? Explain.

d. What proportion of the variation in annual salary is explained by the model? (See Section 5.3.)

14. The college placement office is developing a model to relate grade point average (GPA) to starting salary for liberal arts majors. Twenty recent graduates have been randomly selected, and their graduating GPAs and starting salaries were recorded.

GPA and Starting Salary										
GPA	2.2	3.5	2.1	2.8	1.9	3.2	2.5	2.4	2.9	3.1
Starting Salary (Thousands of Dollars)	55.1	65.2	56.3	59.3	54.3	61.4	57.6	54.8	45.7	63.2
GPA	3.7	2.0	3.3	2.7	3.5	2.6	3.4	3.9	3.0	2.8
Starting Salary (Thousands of Dollars)	59.5	47.8	62.5	55.9	72.4	57.6	62.3	72.5	60.3	58.0

a. Draw a scatterplot of the data. Describe the relationship you observe between GPA and starting salary.

b. Using statistical software, estimate the parameters of the model

$$Starting\ Salary = \beta_0 + \beta_1 GPA + \varepsilon_i.$$

c. Is there evidence of a linear relationship between GPA and starting salary? Test at the 0.05 level.

d. Predict the starting salary for a student with a GPA of 2.5.

e. Interpret the coefficient of GPA in the model.

f. What proportion of the variation in starting salaries is explained by GPA? (See Section 5.3.)

g. To perform statistical inference on the model, what assumptions are being made?

15. A statistics professor would like to build a model relating student scores on the first test to the scores on the second test. The test scores from a random sample of 21 students who have previously taken the course are given in the table.

Test Scores					
Student	First Test Grade	Second Test Grade	Student	First Test Grade	Second Test Grade
1	69	73	12	54	67
2	66	56	13	57	65
3	69	65	14	85	67
4	75	51	15	75	67
5	57	59	16	79	77
6	75	76	17	44	51
7	75	76	18	82	84
8	82	76	19	57	81
9	91	82	20	75	90
10	66	73	21	69	73
11	88	67			

a. Draw a scatterplot of the two test grades and describe the relationship you observe.

b. Using statistical software, estimate the parameters of the model

$$Second\ Test\ Grade = \beta_0 + \beta_1 First\ Test\ Grade + \varepsilon_i.$$

c. What proportion of the variation in the grades on the second test is explained by the grades on the first test?

d. Is there a linear relationship between the first test grades and the second test grades? Test at the 0.05 level.

e. Suppose you're enrolled in the professor's course this semester. If you scored a 75 on the first test, use the model to predict your second test score. Round your answer to the nearest whole number.

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Mount Pleasant Real Estate Data

16. Using the Mount Pleasant Real Estate data set, answer the following questions.

a. Using statistical software, estimate the simple linear regression model relating *List Price* (dependent variable) to *Square Footage* (independent variable).

b. Interpret the slope coefficient of the model.

c. Calculate and interpret a 95% confidence interval for β_1.

d. Is there evidence of a linear relationship between *List Price* and *Square Footage* at the 0.05 level?

e. What proportion of the variation in *List Price* is explained by *Square Footage*? (See Section 5.3.)

f. Using the estimated linear regression model in part **a.**, predict the price of a home in Mount Pleasant that has 3000 square feet.

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > US County Data

17. Using the US County Data data set, answer the following questions.

a. Using statistical software, estimate the simple linear regression model relating *Diabetes.percent* (dependent variable) to *Adult.obesity.percent* (independent variable).

b. Interpret the slope coefficient of the model.

c. Calculate and interpret a 95% confidence interval for β_1.

d. Is there evidence of a linear relationship between *Diabetes.percent* and *Adult. obesity.percent* at the 0.05 level?

e. What proportion of the variation in *Diabetes.percent* is explained by *Adult. obesity.percent*? (See Section 5.3.)

18. Using the Global Statistics by Country data set, answer the following questions regarding birth rate and female literacy rate.

a. Draw a scatterplot of the data. Describe the relationship you observe in the scatterplot.

b. Using statistical software, estimate the simple linear model relating female literacy rate to birth rate.

c. Find a 99% confidence interval for the slope.

d. Interpret the confidence interval found in part **c.** in terms of the problem.

Data

The full data set is available on stat.hawkeslearning.com under **Data Sets > Global Statistics by Country**.

13.3 Inference Concerning the Model's Prediction

The vast majority of regression models are developed for predictive purposes. For example, if you built the model relating the price of a Jeep Cherokee Limited to its age, it was probably because you want to use it to predict prices. While it is important to evaluate b_1, the estimate of the slope, the real concern of the model builder is the accuracy of a model's predictions. In the case of the Jeep Cherokee Limited model, how accurate are the prices that the model predicts? If the assumptions of the linear model (detailed in Section 13.1) have been met, then it is possible to make inferences as to the quality of a model's predictions.

The Regression Line as the Mean Value of *y* Given *x*

Examining the Jeep Cherokee data in Example 13.2.1 reveals two cars that are one-year old. For a given value of age (say one year) the prices of the one-year-old cars were \$44,998 and \$42,768. For anyone who has ever observed the car market, price variation is not unexpected. If you use the model,

$$Estimated\ Asking\ Price\ of\ Jeep\ Cherokee = \$47{,}030.83 - \$3846.09\,Age$$

for predictive purposes, then the predicted value of a one-year-old Jeep Cherokee will be

$$Asking\ Price = \$47{,}030.83 - \$3846.09(1) = \$43{,}184.74.$$

Using this model, all one-year-old Jeep Cherokees will have a predicted value of \$43,184.74. Since the prices of one-year-old Jeep Cherokees vary, how do you interpret the predicted price of \$43,184.74? The model's predicted value when *Age* is set to one is considered to be the average price of a one-year-old Jeep Cherokee. In other words, it is the mean value of *y* (price) when *x* (age) equals one. But wait a minute! The prices

The graph for the prediction interval shown in Figure 13.3.3 has drastically wider confidence bands around the regression line than the graph for the confidence interval for the mean value in Figure 13.3.2. A high price has been paid in order to account for individual variability.

Using the model for prediction outside the range of the x-values used to create the model can be very inaccurate. The nature of the relationship may not be linear outside of the range of the x's used to define the model. In the Jeep Cherokee example, the range of x-values spans from 1 to 6 years. Notice in Figure 13.3.3 that as you approach the edges of the data, the confidence interval widens somewhat. Using the model to predict the price of a 10-year-old Jeep Cherokee would no doubt have sizable error. Inferential methods are not valid outside the range of the data used to estimate the model.

Technology

In order to graph the regression line, along with the prediction interval bands, using Minitab, please visit stat.hawkeslearning.com and navigate to **Discovering Statistics and Data, Fourth Edition > Technology Instructions > Regression > Linear Regression Fitted Line Plot with Prediction Interval**.

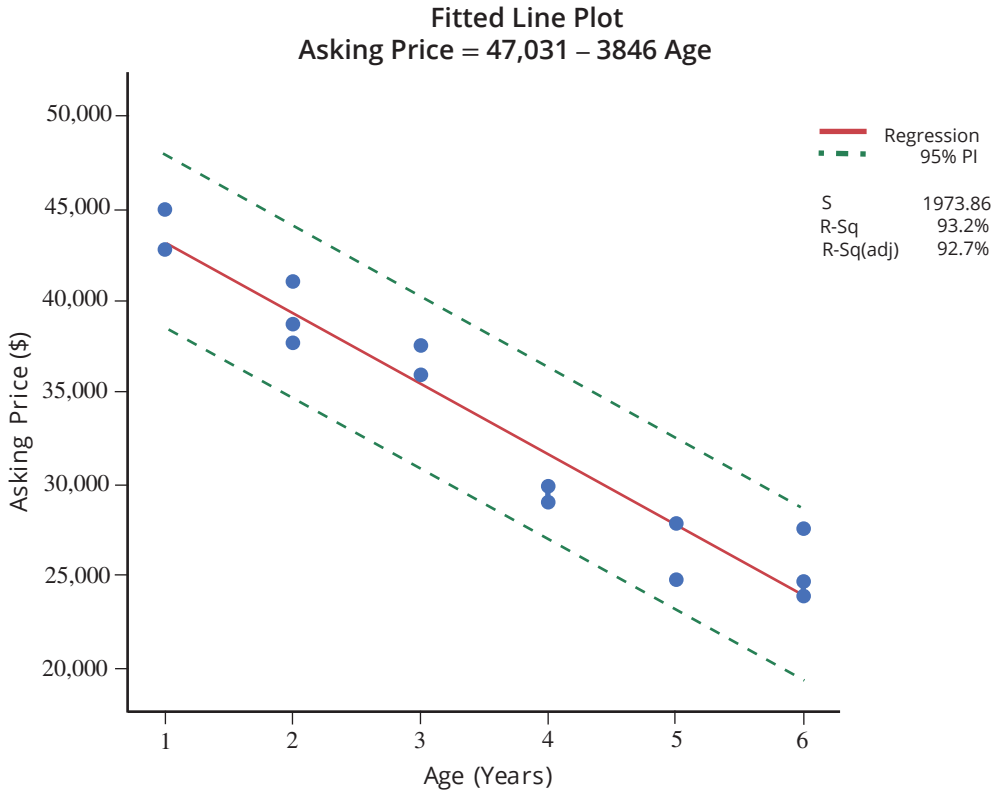


Figure 13.3.3

13.3 Exercises

Basic Concepts

1. Describe the difference in the interpretation of confidence intervals for the mean value of y given x and the predicted value of y given x.
2. Given a confidence interval and a prediction interval, which interval is wider? Explain why.
3. Why should you be cautious when using a regression model to predict outside the range of the x-values used to create the model?

Exercises

4. In Nevada, many forms of gambling are legal and very profitable. Sports betting amounts to billions of dollars annually. In football, a customer will bet on one of the teams to win the contest. However, in an attempt to even the game (from a betting point of view) one of the teams is selected as the favorite. The favorite's score in the game is reduced by an amount called the line. For example, if the Cowboys are favored over the Falcons by four points, then four points are subtracted from the Cowboys' score to determine the outcome of the game for betting purposes. Thus, if the Cowboys defeat the Falcons 32 to 30, in so far as settling any bets, the Cowboys score would be reduced by the spread and the Cowboys would be the loser 32 – 4 = 28 to 30. Where does the betting line come from? The line is created by a betting market. If too many people are betting on the Cowboys before the game starts, the bookmaker will try to make the game more attractive to potential Falcon bettors by increasing the spread say from four points to five points. On the other hand, if too many people are betting on the Falcons, the spread will diminish from four to perhaps three points. How accurate is the betting spread at predicting the actual spread, which is the actual difference in points between the favorite and the underdog? In the example of the Cowboys and the Falcons, the actual spread was +2 (32 – 30). To examine this question, we want to build the following model:

$$Actual\ Point\ Spread = \beta_0 + \beta_1 Betting\ Spread + \varepsilon_i.$$

If the betting spread is a good predictor of the actual spread, it should be able to account for a substantial portion of the variation in the actual spreads. The following table contains betting and actual spreads from 15 randomly selected football games.

Betting vs. Actual Spreads

Betting	4	1	3	2	1	2	5	5	3	4	2	3	5	7	6
Actual	12	−2	6	7	3	1	14	3	−7	5	14	9	2	21	8

a. Draw a scatterplot of the data. Describe the relationship you observe between actual point spread and the betting spread.

b. Estimate the parameters of the model using statistical software.

c. Is there evidence at the 0.05 level of a linear relationship between the betting spread and the actual spread?

d. What proportion of the variation in the actual point spread is explained by the betting spread?

e. Interpret the coefficient of the betting spread in the model (β_1).

f. Construct and interpret a 95% confidence interval for β_1.

g. If the betting spread is five, what is the predicted actual spread?

h. Construct and interpret a 95% prediction interval for a betting spread of five.

i. Construct a 95% confidence interval for the average value of the actual spread when the betting spread is five.

5. Net income is the level of actual profit that a company reports for the year. Net sales is the total sales less adjustment for returns. What is the relationship between net income and net sales for large corporations? Suppose a random sample of 27

large corporations has been selected, and the net income and net sales have been recorded. A regression analysis has been performed to estimate the model, and the output is given.

$$Net\ Income = \beta_0 + \beta_1 Net\ Sales + \varepsilon_i$$

Regression Analysis: Income versus Sales

The regression equation is
Income = 84 + 18.4 Sales

Predictor	Coef	SE Coef	T	P
Constant	83.6	118.1	0.71	0.486
Sales	18.434	4.446	4.15	0.000

S = 372.478 R-Sq = 40.7% R-Sq(adj) = 38.4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	2384660	2384660	17.19	0.000
Residual Error	25	3468497	138740		
Total	26	5853157			

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	1005.3	147.1	(702.4, 1308.2)	(180.5, 1830.0)

Values of Predictors for New Observations

New Obs	Sales
1	50.0

a. Find and interpret the standard deviation of the error terms in the output.

b. Interpret the slope coefficient. (The data used to estimate the model was in millions of dollars.)

c. What proportion of the variation in net income is explained by net sales?

d. Is there evidence of a linear relationship between net income and net sales? Test at the 0.05 level.

e. Construct and interpret a 95% confidence interval for β_1, the slope of the line.

f. The output also contains a predicted value for net income when sales are \$50,000,000. Find the predicted value of net income when sales are \$50,000,000. (Note that in the original data all observations were measured in millions of dollars. Thus, a predicted value of 10,000,000 would be displayed in the output as 10.)

g. Find and interpret the 95% confidence interval for the average value of net income given that sales are \$50,000,000.

h. Suppose your firm generated \$50,000,000 in sales. What would be the 95% prediction interval for your firm's net income?

i. Use the model to predict net income for a company with \$60,000,000 in sales. (Note that you must compute this manually.)

6. The personnel director of a large hospital is interested in determining the relationship (if any) between an employee's age and the number of sick days the employee takes per year. The director randomly selects eight employees and records their age and the number of sick days which they took in the previous year.

Sick Days and Age								
Employee	1	2	3	4	5	6	7	8
Age	30	50	40	55	30	28	60	25
Sick Days	7	4	3	2	9	10	0	8

A regression analysis has been performed to estimate the model and the output is given.

$$Sick\ Days = \beta_0 + \beta_1 Age + \varepsilon_i$$

Regression Analysis: Sick Days versus Age

The regression equation is
Sick Days = 15.2 - 0.247 Age

Predictor	Coef	SE Coef	T	P
Constant	15.186	1.713	8.86	0.000
Age	-0.24681	0.04105	-6.01	0.001

S = 1.47652 R-Sq = 85.8% R-Sq(adj) = 83.4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	78.794	78.794	36.14	0.001
Residual Error	6	13.081	2.180		
Total	7	91.875			

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	6.547	0.557	(5.184, 7.911)	(2.686, 10.409)

Values of Predictors for New Observations

New Obs	Age
1	35.0

a. Draw a scatterplot of the data. Describe the relationship you observe between the number of sick days and age.

b. Find and interpret the standard deviation of the error terms in the output.

c. Interpret the slope coefficient.

d. What proportion of the variation in the number of sick days an employee takes per year is explained by age?

e. Is there evidence of a linear relationship between the number of sick days an employee takes per year and age? Test at the 0.05 level.

f. Construct and interpret a 95% confidence interval for β_1, the slope of the line.

g. Find the predicted value of the number of sick days an employee will take per year if the employee is 35 years old.

h. Find and interpret the 95% confidence interval for the average number of sick days an employee will take per year, given the employee is 35.

i. Suppose a new employee is 35. Find a 95% prediction interval for the number of sick days this employee will take this year.

j. Use the model to predict the number of sick days per year for an employee who is 45 years old. Round to the nearest whole number.

7. A manufacturing company that produces automobiles is interested in studying the relationship between the number of hours of training that an automotive painter receives and the number of paint defects per auto produced. Ten employees are randomly selected. The number of hours of training each automotive painter has received is recorded and the number of paint defects on the most recent auto that they painted is determined. The results are as follows.

Training Hours and Paint Defects										
Hours of Training	1	4	7	3	2	2	5	5	1	6
Defects per Auto	1	4	0	3	5	4	3	2	5	1

A regression analysis has been performed to estimate the model, and the following output is produced.

$$Paint\ Defects\ per\ Auto = \beta_0 + \beta_1 Hours\ of\ Training + \varepsilon_i$$

Regression Analysis: Paint Defects per Auto versus Hours of Training

The regression equation is
Paint Defects per Auto = 4.65 - 0.515 Hours of Training

Predictor	Coef	SE Coef	T	P
Constant	4.6535	0.9426	4.94	0.001
Hours of Training	-0.5149	0.2286	-2.25	0.054

S = 1.45306 R-Sq = 38.8% R-Sq(adj) = 31.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	10.709	10.709	5.07	0.054
Residual Error	8	16.891	2.111		
Total	9	27.600			

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	2.594	0.469	(1.514, 3.674)	(-0.927, 6.115)

Values of Predictors for New Observations

New Obs	Hours of Training
1	4.00

a. Draw a scatterplot of the data. Describe the relationship you observe between the number of paint defects per auto and hours of training. Are there any unusual observations?

b. Find and interpret the standard deviation of the error terms in the output.

c. Interpret the slope coefficient.

d. What proportion of the variation in the number of paint defects per auto is explained by the hours of training? What other factors might affect the number of paint defects?

e. Is there evidence of a linear relationship between the number of hours of training and the number of paint defects per auto? Test at the 0.05 level and the 0.10 level.

f. Construct and interpret a 95% confidence interval for β_1, the slope coefficient.

g. Find the predicted value of the number of paint defects per auto for an automotive painter who has had 4 hours of training.

h. Find and interpret the 95% confidence interval for the average number of paint defects per auto for automotive painters who have had 4 hours of training.

i. Suppose a new automotive painter has had 4 hours of training. What would be the 95% prediction interval for the number of paint defects per auto?

j. Use the model to predict the number of paint defects per auto for an automotive painter who has had 7 hours of training. Round your answer to the nearest whole number.

Data

The data set is available on stat.hawkeslearning.com under **Data Sets > Global Statistics by Country**.

8. Using the Global Statistics by Country data set, answer the following questions regarding birth rate and female literacy rate.

a. Find the predicted value of birth rate for a female literacy rate of 90.

b. Find and interpret the 99% confidence interval for the average birth rate for a female literacy rate of 90.

c. Find and interpret the 99% prediction interval for a female literacy rate of 90.

d. How do these two intervals compare?

CR Chapter Review

Key Terms and Ideas

- Simple Linear Regression Model
- Population Regression Line
- Sample Regression Line
- Assumptions about the Error Term
- Residual
- Confidence Interval for β_1
- Testing a Hypothesis Concerning β_1
- Mean Square Error
- Standard Error
- Confidence Interval for the Mean Value of y Given x
- Confidence Interval for the Predicted Value of y Given x

Key Formulas

Section		
13.1	**Simple Linear Regression Model** $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$	**Estimated Simple Linear Regression Equation** $\hat{y}_i = b_0 + b_1 x_i$
	Sum of Squared Errors $\text{SSE} = \sum \left(y_i - \hat{y}_i \right)^2 = \sum \left(y_i - \left(b_0 + b_1 x_i \right) \right)^2$	**Slope of the Least Squares Line** $b_1 = \dfrac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i \right)^2}.$

The coefficient, b_3, is the estimated change in *List Price* for an additional bedroom for a given square footage and age. Intuitively, we would expect this coefficient to be positive since additional finishing costs are associated with adding a bedroom. The coefficient for b_3 is 1284.92, which means that the price of a home will increase by $1284.92 for each bedroom, all else remaining equal. It is difficult to evaluate whether $1284.92 is a reasonable magnitude, so we will rely on statistical inference to evaluate the coefficients in the next section.

14.1 Exercises

Basic Concepts

1. What is the multiple regression model?
2. What are the assumptions about the error term in a multiple regression model? Are these different from the assumptions required for the simple linear model?
3. What method is used to find the estimated multiple regression equation? Is this method different from the one used to find the simple linear regression equation?
4. What is the greatest challenge in building a multiple regression model?
5. What are some questions that should be asked once a multiple regression model is estimated?
6. What two aspects of the model coefficients are usually analyzed first when studying a multiple regression model?
7. In the simple linear regression model, what is the interpretation of b_1? Does this interpretation change in the multiple regression model?
8. When interpreting the coefficient of an independent variable in a multiple regression model, what assumption are we making regarding the other independent variables?

Exercises

9. Consider the following computer output of a multiple regression analysis relating annual salary to years of education and years of work experience.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.566946595
R Square	0.321428441
Adjusted R Square	0.29192533
Standard Error	10909.996
Observations	49

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	2593556200	1296778100	10.89473033	0.000133875
Residual	46	5475288584	119028012.7		
Total	48	8068844784			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	11214.19915	5625.172956	1.993574106	0.052147881	-108.6867382	22537.08504
Education (Years)	2854.891271	689.6666061	4.139523715	0.000146836	1466.664395	4243.118147
Experience (Years)	839.6360369	261.7094444	3.208275646	0.002433357	312.842248	1366.429826

a. Identify the estimated values of the coefficients b_0, b_1, and b_2.

b. Write the estimated multiple regression equation.

c. Can you think of other independent variables that may be useful in predicting annual salary?

d. Use the model in part **b**. to predict the annual salary of someone with 12 years of education and 2 years of work experience.

10. The manager of a publishing company would like to conduct cost analysis on the most recent books the company has published. The manager would like to estimate a multiple regression model to relate the cost of printing (per book) to the number of pages in the book and the number of copies printed. A computer output of the multiple regression model for the manager's data is given in the following table.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.987606014
R Square	0.975365639
Adjusted R Square	0.972467479
Standard Error	0.445885396
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	133.8201656	66.91008281	336.5464936	2.12863E-14
Residual	17	3.379834375	0.198813787		
Total	19	137.2			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	6.134155476	3.993435752	1.536059638	0.142925974	-2.291257484	14.55956844
Number of Pages	0.010801	0.004147682	2.604105041	0.018522101	0.002050156	0.019551845
Number of Copies	-0.009954478	0.005271436	-1.888380579	0.07616193	-0.021076236	0.00116728

a. Identify the estimated regression coefficients.

b. Write the estimated multiple regression equation.

c. Do the magnitudes and signs of the coefficients seem reasonable? Explain.

d. What other variables do you think could be useful in explaining printing cost per book?

11. A nutritionist wishes to study body weight based on height, age, average calories consumed per day, and the average number of minutes spent exercising per day.

a. Write the multiple regression model the nutritionist is interested in in terms of weight, height, age, calories, and exercise. Assume the coefficients have not yet been estimated.

b. Identify the independent variables in the multiple regression model.

c. Predict the sign of the coefficient for each of the independent variables in the model. Explain your answers.

d. Can you think of any other variables that might be useful for the nutritionist to take into account before performing the regression analysis?

12. Suppose the CEO of an electronics company wants to study the effects of various business practices on annual revenue.

a. Make a list of independent variables the CEO might be interested in studying.

b. Suppose the CEO has narrowed his list of factors down, and decided he wants to mainly study the effects of research and development expenditures, advertising expenditures, and the average annual salary paid to employees. Write the multiple regression model in terms of the dependent and independent variables, assuming the coefficients have not yet been estimated.

c. Make a guess of the sign of the coefficient of research and development expenditures. Explain your prediction.

d. Why should the CEO be cautious when using this model for revenue estimation and prediction?

13. Compare the following two estimated multiple regression equations that relate housing prices in thousands of dollars to the number of bedrooms in the house (*Bedrooms*), size of the lot (*Acreage*), and size of the house (*Square Footage*).

$$\hat{y} = -69{,}280.13 + 142{,}935.73\ Bedrooms + 369{,}879.29\ Acreage,$$

$$\hat{y} = -28{,}520.81 - 34{,}641.71\ Bedrooms + 194{,}986.08\ Acreage + 240.21\ Square\ Footage.$$

What happened to the coefficient on the *Bedrooms* term when the additional variable *Square Footage* was added? Does this make sense? Explain.

14. Consider the following estimated multiple regression equation relating the number of study hours for the ACT and high school GPA to a student's ACT score.

$$ACT\ Score = 8.35 + 0.015\ Study\ Hours + 0.30\ GPA$$

a. Identify the values of b_0, b_1, and b_2.

b. Interpret the value of b_0 in terms of the problem.

c. Interpret the value of b_1 in terms of the problem.

d. Interpret the value of b_2 in terms of the problem.

15. Consider the following estimated regression model relating annual salary to years of education and work experience, which was presented in Exercise 9.

$$Salary = 11{,}214.20 + 2854.89\ Education + 839.64\ Experience$$

a. Consider the coefficient for the *Education* variable. Do the sign and magnitude of the coefficient seem to make sense? Explain.

b. Consider the coefficient for the *Experience* variable. Do the sign and magnitude of the coefficient seem to make sense? Explain.

c. Interpret the regression coefficient for years of experience.

d. Suppose an employee with 8 years of education (note that education years are the number of years after 8th grade) has been with the company for 5 years. According to this model, what is their estimated annual salary?

e. How would you expect an employee's salary to change if they stay at the company for another year?

f. Suppose two employees at the company have been working there for five years. One has a bachelor's degree (8 years of education) and one has a master's degree (10 years of education). Which employee would you expect to earn a higher salary? What is the difference in salary between the two employees?

The Proof of the Pudding is in the Eating

How do you know if you have a useful predictive model?

One might think that a useful model would have a high R^2. But, is a high R^2 necessary to have a useful model? In the introduction to Chapter 14, I mentioned a story relating my first experience using regression analysis in predicting the speed of a horse in a race. The R^2 of that model was roughly 0.35. Yet the model predicted well enough to allow us to have a profitable betting experience. Later, this same friend and I would start a company that predicted stock prices. The R^2 associated with many of our models was less than 0.1. Yet, we were able to profitably trade substantial volumes of stock with these models.

To judge how effective a model is, you need to use it for its intended purpose. Thus, there are two questions for any predictive model. First, can you predict better with the model than without it? Second, can your model's predictions achieve the goals you have for the model? If the answer is yes to both of these questions, you have a useful model regardless of the value of R^2. Also note, a model with a large value of R^2 may not be a useful model by the two preceding criteria.

Adjusted R^2 (R_a^2)

Adding more independent variables to a regression model will always increase the R^2 value. R^2 will never decrease as variables are added because the SSE can never become smaller with the addition of independent variables, and the Total SS is always the same for a given set of responses. Since R^2 can be made larger by including a large number of independent variables, it is sometimes suggested that a modified measure be used that adjusts for the number of independent variables in the model. The **adjusted coefficient of determination** (denoted by R_a^2) adjusts R^2 by dividing each sum of squares by its associated degrees of freedom. Thus, R_a^2 is given by the following formula.

Adjusted R^2

The adjusted R^2 statistic takes into account the number of independent variables in the model by dividing each sum of squares by its associated degrees of freedom.

$$R_a^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\frac{\text{SSE}}{\text{Total SS}}$$

where n is the number of observations and k is the number of independent variables in the model.

DEFINITION

For example, if one were to fit a simple linear regression model to *List Price* using only *Square Footage* as the independent variable, we would get $R^2 = 0.8799$ and $R_a^2 = 0.8761$. However, when we add the other two independent variables to the model, we have $R^2 = 0.9634$ and $R_a^2 = 0.9598$. In this case, the value of R^2 increased by 0.0835, indicating that adding the other variables to the model helped explain more variability in *List Price*. On the other hand, the value of R_a^2 increased by slightly more (0.0837). R_a^2 is commonly used as a method of comparison between multiple regression models when one is attempting to find the model that best fits the data. Unlike the R^2 value, the adjusted coefficient of determination may actually become smaller when another independent variable is added to the model. Thus, the adjusted R^2 value is most useful when comparing multiple regression models with different numbers of independent variables.

14.2 Exercises

Basic Concepts

1. What does R^2 represent?
2. What range of values can the coefficient of determination take on?
3. Can you think of a way a model might have a large R^2 and not be useful for prediction? Explain.
4. Explain the difference between R^2 and adjusted R^2.
5. Explain why the adjusted R^2 statistic is sometimes a better measure to use to evaluate the fit of a regression model.
6. Will there ever be a situation in which the adjusted R^2 statistic is greater than the R^2 statistic? Explain your answer.

Exercises

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Mount Pleasant Real Estate Data.

7. Using the Mount Pleasant Real Estate data set, construct a multiple regression model relating housing prices (in thousands of dollars) to the number of bedrooms in the house, labeled *Bedrooms*, and the size of the lot on which the house was built, labeled *Acreage*.

a. Write the estimated regression equation.

b. Identify the values of SSR, SSE, and Total SS from the table.

c. What is the coefficient of determination for this model? Interpret this value in terms of the problem.

d. What is R_a^2 ? Interpret this value.

e. Compare the R^2 and R_a^2 values. Which value should be used to evaluate the fit of the multiple regression model? Explain why.

8. Add an additional variable, *Square Footage*, to the housing price model from Exercise 7.

a. Write the estimated regression equation.

b. What is R_a^2 for this model?

c. How does the adjusted R^2 value for this model compare to the adjusted R^2 value for the model in Exercise 7?

d. Do you think adding the additional independent variable, *Square Footage*, improved the model? Explain your answer.

9. The owner of a new pizzeria in town wants to study the relationship between weekly revenues and advertising expenditures. Both measures were recorded in thousands of dollars. The computer output for the simple linear regression model is given below.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.858179902
R Square	0.736472743
Adjusted R Square	0.692551534
Standard Error	1.058296197
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18.78005496	18.78005496	16.76804334	0.006394067
Residual	6	6.719945042	1.11999084		
Total	7	25.5			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	74.69887795	7.104358625	10.51451396	4.34789E-05	57.31513863	92.08261726
Advertising Expenditures	1.854820243	0.452960815	4.094880138	0.006394067	0.746465058	2.963175428

a. Write the estimated regression equation.

b. What is the coefficient of determination for this model? Interpret this value.

c. What is the value of the adjusted R^2 statistic? Is this statistic useful for the pizzeria owner as he studies this model? Explain.

d. Do you believe this model is useful in explaining revenues based on advertising expenditures? Explain your answer.

10. How could the restaurant owner improve this model? Are there other independent variables that he should consider including? The owner of the pizzeria discussed in Exercise 9 wishes to build on the model relating revenues to advertising expenditures by breaking the advertising expenditures into three categories: television advertising, newspaper advertising, and direct mail advertising.

a. Write the new regression model in terms of television, newspaper, and mail expenditures. Assume the coefficients have not yet been estimated.

b. Consider the following summary output for the new model. Write the estimated multiple regression equation.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.967040091
R Square	0.935166537
Adjusted R Square	0.88654144
Standard Error	0.64289449
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	23.8467467	7.948915566	19.23217829	0.007708883
Residual	4	1.653253302	0.413313326		
Total	7	25.5			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	73.93199827	4.523870838	16.34264127	8.20538E-05	61.37171922	86.49227731
Television	2.383047934	0.318133378	7.490719616	0.001698799	1.499768074	3.266327793
Newspaper	1.454439994	0.355820285	4.087569076	0.015004989	0.466524505	2.442355483
Mail	1.815990841	0.276487962	6.568064755	0.002780349	1.048337191	2.58364449

c. Interpret the coefficient for television advertising expenditures. Remember that revenues and expenditures are in thousands of dollars.

d. What is the adjusted coefficient of determination? Interpret this value.

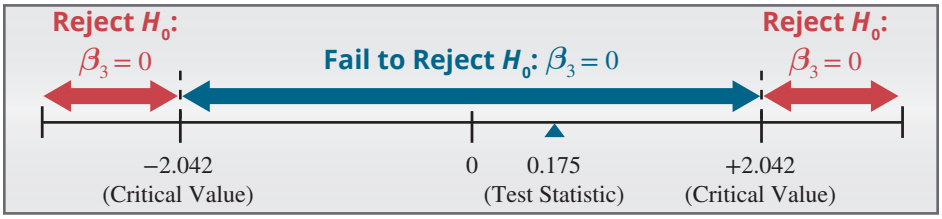
e. How does the coefficient of determination of this model compare to the coefficient of determination for the simple linear regression model in Exercise 9? Does this appear to be a more useful model? Explain.

f. What is the value of the R^2 statistic for this model? Should we use the R^2 value or the adjusted R^2 value when evaluating the usefulness of this model? Explain why.

Using the output we find that

$$t = \frac{b_3 - 0}{s_{b_3}} = \frac{b_3}{s_{b_3}} = \frac{1284.92}{7357.66} \approx 0.175.$$

The estimated value of b_3 is 0.175 standard deviations from zero. Is this persuasive evidence that $\beta_3 \neq 0$?



Step 4: Determine the critical value(s) or *P*-value.

This criteria is defined by the critical value of the test statistic. Since the test is two-tailed and $\alpha = 0.05$ then $\alpha/2 = 0.05/2 = 0.025$. The test statistic has a t-distribution with $df = 34 - (3 + 1) = 30$. The critical value corresponds to $t_{0.025,\ 30} = 2.042$. The P-value for the number of bedrooms is given in the output as 0.8625.

Step 5: Choose between the null and alternative hypotheses.

Since the value of the test statistic falls into the *Fail to Reject* region, there is insufficient evidence at the 0.05 level to reject the null hypothesis $\beta_3 = 0$. Alternatively, since the P-value of 0.8625 is greater than $\alpha = 0.05$, we fail to reject the null hypothesis.

Step 6: State the conclusion in terms of the original question.

Since we did not reject the null hypothesis H_0: $\beta_3 = 0$, then the variable *Bedrooms* is not a significant predictor of *List Price*, given the other variables currently in the model.

We apply the exact same t-test to the other variables in the model. Both b_1 and b_2, the coefficients of the variables *Square Footage* and *Age,* are significant. This suggests that a model with only two independent variables, *Square Footage* and *Age*, may produce a model almost as good as the one containing three variables.

14.3 Exercises

Basic Concepts

1. If the overall multiple regression model is not useful, what does this tell us about the coefficients of the independent variables?
2. What is the hypothesis being tested when we test to determine if the overall multiple regression model is useful?
3. When testing the overall model, describe the null and alternative hypotheses in plain English.
4. What is the test statistic used in a hypothesis test to determine if an overall model is significant? What is the distribution of this test statistic?

5. Explain the significance of the ratio of the mean square regression to the mean square error.
6. True or false: Even if there is no relationship between any of the independent variables and the dependent variable, sampling variation will explain some portion of the variation in the dependent variable.
7. How are the degrees of freedom calculated for a multiple regression model?
8. When testing the overall model for significance, do you perform a one or two-tailed test?
9. What is the rejection rule in tests of hypothesis for model significance?
10. What is the expression for a confidence interval for an individual coefficient, β_i?
11. Outline the three pieces of information needed to compute a confidence interval for an individual coefficient.
12. What is the test statistic used to test a hypothesis about an individual coefficient in a multiple regression model? How many degrees of freedom are associated with this test statistic?
13. If we fail to reject the null hypothesis in a hypothesis test about an individual coefficient, should this variable remain in the regression model? Explain.
14. Does a low R^2 imply that a model will not be useful for prediction?

Exercises

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Moneyball

15. In Lesson 5.3 Exercise 24, we used the Moneyball data set (1962-2001) to look at the individual relationships between runs scored (*RS*) and on-base percentage (*OBP*) and runs scored (*RS*) and slugging percentage (*SLG*). On-base percentage (*OBP*) and slugging percentage (*SLG*) were determined to be two of the most statistically significant variables that contributed to the number of runs scored. Remember that a run differential of 135 runs was identified as necessary to make the MLB playoffs. Use the Moneyball data set, subsetted to only the years 1962-2001, to perform the following.
    a. Build a single model to predict runs scored (*RS*) using on-base percentage (*OBP*) and slugging percentage (*SLG*). Write the estimated regression equation.
    b. Is the overall model significant at the 1% level?
    c. What percent of variation in the runs scored (*RS*) is explained by on-base percentage (*OBP*) and slugging percentage (*SLG*)?
    d. Determine if each independent variable is related to the dependent variable at the 0.01 level of significance.
    e. Should we consider removing any independent variables from this regression model? If yes, identify the variable(s) that should be removed and explain why.
16. Consider the model from Exercise 9 in Section 14.1 relating annual salary to years of work experience and years of education.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.566946595
R Square	0.321428441
Adjusted R Square	0.29192533
Standard Error	10909.996
Observations	49

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	2593556200	1296778100	10.89473033	0.000133875
Residual	46	5475288584	119028012.7		
Total	48	8068844784			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	11214.19915	5625.172956	1.993574106	0.052147881	-108.6867382	22537.08504
Education (Years)	2854.891271	689.6666061	4.139523715	0.000146836	1466.664395	4243.118147
Experience (Years)	839.6360369	261.7094444	3.208275646	0.002433357	312.842248	1366.429826

a. Formulate the hypotheses for testing the multiple regression model for overall significance.

b. Find the value of the test statistic for a hypothesis test about the overall model.

c. Is there evidence that the overall model is useful in predicting annual salary?

d. Consider the coefficient for years of education. Find a 95% confidence interval for the value of β_1. Interpret this interval.

e. Formulate the hypotheses for testing the significance of the coefficient β_1.

f. Is there sufficient evidence at the 0.05 level that years of education is useful in predicting annual salary?

17. Consider the printing cost model discussed in Exercise 10 of Section 14.1.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.987606014
R Square	0.975365639
Adjusted R Square	0.972467479
Standard Error	0.445885396
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	133.8201656	66.91008281	336.5464936	2.12863E-14
Residual	17	3.379834375	0.198813787		
Total	19	137.2			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	6.134155476	3.993435752	1.536059638	0.142925974	-2.291257484	14.55956844
Number of Pages	0.010801	0.004147682	2.604105041	0.018522101	0.002050156	0.019551845
Number of Copies	-0.009954478	0.005271436	-1.888380579	0.07616193	-0.021076236	0.00116728

a. What percentage of the variation in *Printing Cost is* explained by the two independent variables *Number of Pages* and *Number of Copies*?

b. Is the overall model significant at the 1% level?

c. Consider the estimated regression coefficient for the number of pages. Construct a 99% confidence interval for β_1. Interpret this interval.

d. Is the *Number of Pages* variable useful in predicting *Printing Cost* at the 5% level? Would the decision change at the 1% level?

e. Construct a 95% confidence interval for β_2. Interpret this interval.

f. Is the number of copies useful in explaining the variation in printing cost at the 5% level of significance? Do you think the publisher should consider removing this variable from the model? Explain your answer.

18. The following table contains US Census Bureau data from selected cities regarding rental rates of two-bedroom apartments, city populations, and median incomes.[1] Monthly rent is given in dollars, population is given in thousands of people, and median income is given in thousands of dollars. Suppose we wish to build a multiple regression model to predict the cost of rent based on population and median income.

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > City Population Data

Monthly Rent, Population, and Median Income in Selected Cities

City	Monthly Rent ($)	2020 Population (Thousands)	2020 Median Income
Denver, CO	1397	715.522	$72,661
Birmingham, AL	870	200.733	$38,832
San Diego, CA	1770	1386.932	$83,454
Gainesville, FL	965	141.085	$38,028
Winston-Salem, NC	827	249.545	$47,269
Memphis, TN	915	633.104	$41,864
Austin, TX	1346	961.855	$75,752
Seattle, WA	1702	737.015	$97,185
Richmond, VA	1070	226.610	$51,421
Charleston, SC	1318	150.227	$72,071
College Park, MD	1583	34.740	$68,825
Savannah, GA	1049	147.780	$46,149
Minneapolis, MN	1078	429.954	$66,068
Detroit, MI	850	639.111	$32,498
Baton Rouge, LA	886	227.470	$44,177

a. Write the multiple regression model using population and median income to predict rent. Assume the regression coefficients have not yet been estimated.

b. Predict the signs of the coefficients β_1 and β_2. Explain your answers.

c. Using statistical software, estimate the multiple regression equation. Identify the values of b_0, b_1, and b_2 and write the estimated multiple regression equation. Interpret the estimated coefficients.

d. At the 1% level of significance, is the overall model useful in predicting monthly rent? Identify the test statistic for this test.

e. Find a 95% confidence interval for β_2. Interpret this interval.

f. Determine if each independent variable is related to the dependent variable at the 0.05 level of significance.

g. Should we consider removing any independent variables from this regression model? If yes, identify the variable(s) that should be removed and explain why.

19. Using the information from Exercise 18, estimate the simple linear regression equation using median income to predict rent.

a. Write the estimated simple regression equation.

b. Is the simple linear regression model significant at $\alpha = 0.01$?

c. Is median income related to the monthly rental rate at $\alpha = 0.01$? Identify the test statistic used in this hypothesis test.

d. What percent of the variation in monthly rent is explained by median income? Compare this to the percent of variation in monthly rent explained by both population and median income in Exercise 18.

e. Which model do you think is a better model to use to predict monthly rental rates? Explain your answer.

14.4 Inference Concerning the Model's Prediction

Many regression models are developed solely to predict the dependent variable. To use the multiple regression model for prediction, insert the values of the independent variables in the model and calculate the predicted value. For a house with 2500 square feet that is ten years old with four bedrooms, the model would predict the price to be

$$\begin{aligned} \textit{List Price} &= 163579.06 + 108.24(2500) - 6318.47(10) + 1284.92(4) \\ &\approx \$376{,}134.04 \end{aligned}$$

This is the point estimate. How good is this estimate? The answer to this question depends on what you are trying to predict. Are you trying to predict the average price for all 2500 square foot homes that are ten years old with four bedrooms, or are you trying to predict the price of a particular home of this type?

Note

The value of \$376,134 for a 2500 square foot home that is 10 years old and has 4 bedrooms that is estimated using the Home Price model differs slightly from the value calculated in the Minitab output (\$376,142) due to rounding.

Confidence Intervals for the Mean Value of *y* Given *x*

In Section 13.3 we discussed a confidence interval for the mean (or average) value of y given $x = x_p$ for the simple linear regression model. In our multiple regression model, the point estimate, \$376,134, is the mean value of y given $x_1 = 2500$, $x_2 = 10$, and $x_3 = 4$. In other words, the price of \$376,134 is the estimated average price for all ten-year-old, 2500 square feet homes with four bedrooms. Since we do not have all homes in the sample, the predicted average price of \$376,134 is only an estimate of the true average value. How good is the estimate? For multiple regression, the expression for the confidence interval of the mean value of y given x is beyond the scope of an introductory text.

profits if trades were made using the model's prediction of future prices of a given stock. In case you are wondering, the model eventually worked. Before the company was sold it was trading six percent of all trades on the New York Stock Exchange and the NASDAQ. Regression modeling is an incredibly effective tool for modeling real-world phenomena, but if you're interested in a career in model building, it's also worth exploring machine learning. Both machine learning and statistical modeling are data-driven, but what sets machine learning apart, particularly in terms of modeling, is its emphasis on a model's ability to perform predictive tasks and its capacity to utilize multiple modeling techniques to accomplish this objective. The ultimate goal of machine learning is to determine the optimal model for a given predictive task.

14.4 Exercises

Basic Concepts

1. What is a point estimate for a multiple regression model?
2. Explain how a point estimate is interpreted as an "average" value.
3. Distinguish between a confidence interval and a prediction interval for a multiple regression model.
4. What is the price that is paid when making predictions regarding individual values?
5. Suppose an estimated multiple regression model, $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$, produces a 95% confidence interval of (3.292, 7.072) and a 95% prediction interval of (0.364, 10.000) when $x_1 = 6$ and $x_2 = 6$. Interpret both of these intervals.

Exercises

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > SAT Scores and Graduating GPA

6. Use the SAT Scores and Graduating GPA data set of 30 students that was discussed in Example 13.2.3. In that example, we estimated a model using only total SAT score to predict graduating GPA. However, with multiple regression, the *SAT Verbal* and *SAT Math* scores can be treated as separate variables in the model. Computer output of the model

$$College\ GPA = \beta_0 + \beta_1\, SAT\ Verbal + \beta_2\, SAT\ Math + \varepsilon$$

is given.

Regression Analysis: College GPA versus SAT Verbal, SAT Math

Analysis of Variance

Source	DF	Adj SS	Adj Ms	F-Value	P-Value
Regression	2	0.8763	0.4382	2.643	0.0895
Error	27	4.4760	0.1658		
Total	29	5.3524			

Model Summary

S	R-sq	R-sq(adj)
0.407160	16.37%	10.18%

Coefficient

Term	Coef	SE Coef	T-Value	P-Value
Constant	0.13128	1.15417	0.114	0.9103
SAT Verbal	0.00179	0.00137	1.311	0.2009
SAT Math	0.00273	0.00187	1.457	0.1566

Regression Equation

College GPA = 0.13128 + 0.00179 SAT Verbal + 0.00273 SAT Math

Prediction for College GPA

Settings

Variable	Setting
SAT Verbal	500
SAT Math	500

Prediction

Fit	SE Fit	95% CI	95% PI
2.39	0.18	(2.03, 2.75)	(1.48, 3.30)

a. What is the estimated regression model?

b. Use the output provided to determine the standard deviation of the error terms.

c. Interpret the coefficient of *SAT Verbal*. What would it mean if the coefficient was negative?

d. Determine if the overall model is useful in explaining *College GPA*. Test at the 0.05 level.

e. What proportion of the variation in GPA is explained by the model?

f. Determine if the *SAT Verbal* variable is a useful predictor of *College GPA*. Test at the 0.05 level.

g. The output includes a predicted GPA for someone scoring 500 on both the SAT Verbal and SAT Math portions. Find the predicted value in the output.

h. What is the model's estimate of the average GPA for individuals who scored 500 on both the SAT Verbal and SAT Math sections? Find the 95% confidence interval for this average. Interpret this interval.

i. Suppose your nephew scored 500 on both the SAT Verbal and SAT Math sections. What would be the model's prediction for his graduating GPA? Find the 95% prediction interval for your nephew in the output. Interpret this interval.

j. Why is the prediction interval in part **i**. so much wider than the confidence interval in part **h.**?

k. Summarize the strengths and weaknesses of the estimated model.

7. How tall will your child be? A researcher has collected a random sample of heights of parents and their female children (all heights are in inches). The heights of the mother, father, and daughter are recorded in the following table.

Heights of Parents and Daughters (Inches)													
Mother	64	66	62	70	70	58	66	66	64	67	65	66	68
Father	73	70	72	72	72	63	75	75	72	69	77	70	74
Daughter	65	65	61	69	67	59	69	70	68	70	70	65	70

a. Create two scatterplots using the mother with the daughter and the father with the daughter. Does there appear to be a linear relationship in either of the plots?

b. Using statistical software, estimate the parameters of the following regression model.

$$Daughter\ Height = \beta_0 + \beta_1\ Mother\ Height + \beta_2\ Father\ Height + \varepsilon$$

c. Is the overall model useful in explaining the variation in daughter height? Test at the 0.05 level.

d. Is the father's height useful in explaining the daughter's height? Test at the 0.05 level.

e. Is the mother's height useful in explaining the daughter's height? Test at the 0.01 level.

f. Interpret each of the regression coefficients.

g. Construct and interpret 95% confidence intervals for β_1 and β_2. Interpret these intervals.

h. Predict the height of a daughter whose father is six feet two inches tall and whose mother is five feet four inches tall.

i. Find a 95% prediction interval for the height of a daughter whose father is six feet two inches tall and whose mother is five feet four inches tall. Interpret this interval.

j. Find a 95% confidence interval for the average height of a daughter whose father is six feet two inches tall and whose mother is five feet four inches tall.

8. On Sunday, September 12, 2021, 14 games were played in the National Football League. The number of rushing yards, passing yards, first downs, and points for the 28 teams participating in these games is given in the table.[2]

Note

As of the date of this publication, Excel does not perform 95% confidence intervals for the mean of y given x nor will it perform 95% prediction intervals. Minitab, R, and Rguroo can perform these analyses.

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > NFL Statistics 09/12/2021

Team Data: September 12, 2021									
Team	**Rushing Yards**	**Passing Yards**	**First Downs**	**Points**	**Team**	**Rushing Yards**	**Passing Yards**	**First Downs**	**Points**
Jacksonville Jaguars	76	319	20	21	**Houston Texans**	160	289	22	37
Seattke SeaHawks	140	241	18	28	**Indianapolis Colts**	113	223	23	16
Los Angeles Chargers	90	334	27	20	**Washington**	126	133	15	16
NY Jets	45	207	16	14	**Carolina Panthers**	111	270	18	19
Minnesota Vikings	67	336	24	24	**Cincinnati Bengals**	149	217	20	27
Arizona Cardinals	136	280	22	38	**Tennessee Titans**	86	162	17	13
San Francisco 49ers	131	311	21	41	**Detroit Lions**	116	314	31	33
Pittsburgh Steelers	75	177	16	23	**Buffalo Bills**	117	254	22	16
Philadelphia Eagles	173	261	24	32	**Atlanta Falcons**	124	136	19	6
Cleveland Browns	153	304	24	29	**Kansas City Chiefs**	73	324	21	33
Green Bay Packers	43	186	14	3	**New Orleans Saints**	171	151	22	38
Denver Broncos	165	255	24	27	**NY Giants**	60	254	19	13
Miami Dolphins	74	185	16	17	**New England Patriots**	125	268	24	16
Chicago Bears	134	188	24	14	**Los Angeles Rams**	74	312	18	34

a. In order to predict a team's points from rushing yards, passing yards, and first downs, a multiple regression model is constructed. The associated regression output is given. Write the estimated regression equation for predicting points based on the three predictor variables.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.7545798
R Square	0.5693906
Adjusted R Square	0.5155645
Standard Error	7.0592337
Observations	28

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	1581.442	527.1473	10.57832	0.000127
Residual	24	1195.987	49.83278		
Total	27	2777.429			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-8.465112	7.655233	-1.10579	0.279778	-24.2647	7.33451	-24.2647	7.33451
Rushing Yards	0.1723691	0.042912	4.016838	0.000505	0.083804	0.26093	0.0838	0.26093
Passing Yards	0.1128868	0.028515	3.958808	0.000585	0.054034	0.17174	0.05403	0.17174
First Downs	-0.737402	0.515136	-1.43147	0.16519	-1.80059	0.32579	-1.80059	0.32579

b. Use the regression equation to predict the points scored by a team that rushed for 152 yards, had 190 passing yards, and 21 first downs.

c. Find the standard deviation of the error terms in the output.

d. Determine if the overall model is useful in predicting points scored. Use $\alpha = 0.05$.

e. What fraction of the total variation in points is explained by the model?

f. Is the *Rushing Yards* variable useful in predicting points scored at the 0.01 level?

g. Is the *Passing Yards* variable useful in predicting points scored at the 0.01 level?

h. Is the *First Downs* variable useful in predicting points scored at the 0.01 level?

i. The coefficient of *Rushing Yards* in the regression equation is 0.1724. Interpret this value.

j. Should any variables be removed from this model? Explain.

9. In the previous exercise, total points was predicted based on rushing yards, passing yards, and first downs. It is noted from the summary output that both *Rushing Yards* and *Passing Yards* have *P*-values of less than 0.01. However, *First Downs* does not appear to be significant as an independent variable. Perhaps a simpler model would be better.

a. Using the data from the previous exercise, estimate the regression equation

$$Points = \beta_0 + \beta_1\, Rushing\ Yards + \beta_2\, Passing\ Yards + \varepsilon.$$

b. Is the overall model significant in predicting total points? Test at $\alpha = 0.01$.

c. What percentage of the variation in total points is explained by Rushing Yards and passing yards? Compare this to the percentage of the variation in total points that was explained by the three independent variables *Rushing Yards*, *Passing Yards*, and *First Downs*.

d. Which model do you think would be better to use for estimation and prediction of total points; the model from Exercise 8 or the model in this exercise? Explain your answer.

e. Suppose that in preparation for the upcoming game against Miami, the coach of Buffalo wishes to predict the points that will be scored. He has studied Miami's defense in previous games, and predicts that the Buffalo offense will have approximately 102 rushing yards and 263 passing yards. How many points, according to the model, should Buffalo score in the next game?

f. Construct at 95% confidence interval for the average number of points that will be scored in the game against Miami. Interpret this interval.

g. Construct a 95% prediction interval for the number of points that will be scored in the game against Miami. Interpret this interval.

10. A personnel director is interested in studying the effects that age and work experience have on annual salary. Eight employees are randomly selected, and each employee's salary, age, and years of work experience are recorded.

Employee Data		
Salary	**Age**	**Experience**
$43,500	25	3
$75,000	55	20
$72,000	47	15
$52,500	30	7
$40,500	22	2
$87,000	62	26
$34,500	19	1
$64,500	44	10

a. Using statistical software, estimate the multiple regression equation for $Salary = \beta_0 + \beta_1 Age + \beta_2 Experience + \varepsilon$.

b. Determine if the overall model is useful in explaining salary at the 0.05 level of significance. What is the test statistic for the hypothesis test?

c. Is the *Experience* variable useful in predicting annual *Salary* at the 0.05 significance level?

14.5 Multiple Regression Models with Qualitative Independent Variables

Throughout Chapter 13 and this chapter, we have discussed quantitative variables in the regression models. Quantitative variables take on values on a well-defined scale, such as number of pizzas, miles to destination, income, age, and temperature. Many variables of interest, however, are not quantitative, but qualitative. Examples of qualitative variables are type of college (public or private), season of the year (spring, summer, fall, and winter), and type of investment (stocks, mutual funds, or bonds).

Lastly, to compare mall and downtown locations, we look at

$$\left(\beta_0 + \beta_1 x_1 + \beta_2\right) - \left(\beta_0 + \beta_1 x_1 + \beta_3\right) = \beta_2 - \beta_3.$$

The estimated difference between β_2 and β_3 is given by $\beta_2 - \beta_3$ (21.1001) which represents the difference between the average annual return for shops in the mall locations having x_1 households in the area and the average annual return for shops in the downtown locations having x_1 households in the area. In terms of the problem, we can say that for any number of households in a given area, the average annual return in a mall area will be \$21,100.10 greater than the average annual return in a downtown location.

There are three potential issues to keep in mind when using these regression results.

1. These results are only meaningful within the relevant range of the data that was used to estimate the regression equation. For example, using the model in Example 14.5.2 to predict annual return for a shop with 1000 households or 1,000,000 households in the surrounding area would likely yield an unreliable point estimate.
2. The regression lines for the three locations in Example 14.5.2 are assumed to have the same slope, but in reality they could have very different slopes. Using a regression model with **interaction terms** allows the slopes for the regression lines to differ.
3. The regression lines estimated in this example are all linear. This implies that annual return increases by the same amount for each additional thousand households within 15 miles of the shops. This assumption is sometimes unrealistic. This issue can be addressed using **polynomial (or nonlinear) regression models**.

Regression models with interaction terms and polynomial regression models are beyond the scope of this text and are not discussed in detail. Multiple regression is a complex topic that involves many methods of estimation. We only present the basics in this text.

14.5 Exercises

Basic Concepts

1. Give three examples of qualitative independent variables that may be of interest to someone performing regression analysis to predict annual salary.

2. Explain how qualitative variables are transformed into quantitative variables in order to estimate a regression model.

3. If a qualitative variable has c classes, how many indicator (dummy) variables will there be in the model? Explain why this is the case.

4. When an indicator (dummy) variable is equal to one, does this represent a difference in the slope or the intercept of the model? Explain.

5. What is a base level variable? Interpret the value of an estimated coefficient for an indicator variable in terms of the base level variable.

6. Identify three potential issues to keep in mind when constructing regression models involving indicator variables. Also suggest how these issues can be addressed.

Exercises

7. **a.** How many indicator variables would it take to construct a qualitative variable with 5 states?

b. Assume the states were: very small, small, medium, large, very large. Develop the indicator variables to represent this qualitative variable in a regression model.

c. Given your design what is the meaning of the constant term in your model?

8. **a.** How many indicator variables would it take to construct a qualitative variable with 4 educational levels?

b. Assume the educational levels are: non high school graduate, high school graduate, some college, and college graduate. Develop the indicator variables to represent this qualitative variable in a regression model.

c. Given the design you have chosen, what is the meaning of the constant term in your model?

9. Consider the following estimated multiple regression model relating GPA to the number of classes attended and the final exam score in a particular class, and if the student is a freshman (= 1 if freshman, = 0 otherwise).

$$\textit{Cumulative GPA} = -0.8777 + 0.0672\ \textit{Attendance} + 0.0678\ \textit{Exam Score} - 0.1436\ \textit{Freshman}$$

a. Are the signs of the estimated coefficients what you would expect for these three independent variables? Explain.

b. Interpret the coefficient for the *Attendance* variable.

c. Interpret the coefficient for the *Exam Score* variable.

d. Interpret the coefficient for the *Freshman* variable.

e. Suppose two students, one a freshman and one a senior, attended the same number of classes and both got a score of 88 on the final exam. What would be the expected difference in GPA for the two students?

10. Consider the following computer output for the multiple regression model discussed in the previous exercise.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.714589997
R Square	0.510638864
Adjusted R Square	0.508467143
Standard Error	0.516416069

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	188.1180981	62.70603271	235.1309671	1.8485E-104
Residual	676	180.2794359	0.266685556		
Total	679	368.397534			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-0.877712645	0.138557037	-6.334666683	4.34037E-10	-1.149766538
Attendance	0.067163994	0.003669275	18.3044333	4.30384E-61	0.059959449
Exam Score	0.067820136	0.004265782	15.89864106	1.37161E-48	0.059444361
Freshman	-0.143623671	0.047077779	-3.050774156	0.002371853	-0.236059922

a. Test the usefulness of the overall model in predicting *Cumulative GPA* using a 5% significance level.

b. What percentage of the variation in *Cumulative GPA* is explained by the three independent variables?

c. Is the qualitative independent variable, *Freshman*, useful in predicting Cumulative GPA? Use $\alpha = 0.05$.

d. Can you think of other variables that could be added to the model? Name one quantitative variable and one qualitative variable that might be useful.

11. A personnel director is interested in studying the effects which age, experience, and education level have on salary. Eight employees are randomly selected and each employee's salary, age, experience, and education level (0 if high school degree or below, 1 if college degree or above) are recorded.

Employee Data

Salary ($)	Age	Experience (Years)	Education Level
48,600	25	2	1
90,000	55	20	0
86,400	27	5	0
63,000	30	7	1
52,200	22	3	1
104,400	33	8	1
41,400	19	1	0
77,400	45	15	1

a. Create three scatterplots using salary with age, salary with experience, and salary with education level. Does each of the plots have a linear relationship?

b. Using statistical software, estimate the parameters of the following regression model:

$$Salary = \beta_0 + \beta_1\, Age + \beta_2\, Experience + \beta_3\, Education\ Level + \varepsilon.$$

c. Is the overall model useful in explaining salary? Test at the 0.05 level.

d. Is age useful in explaining salary? Test at the 0.05 level.

e. Is experience useful in explaining salary? Test at the 0.01 level.

f. Is education level useful in explaining salary? Test at the 0.10 level.

g. Interpret each of the regression coefficients.

h. Predict the salary of an employee with a college degree who is 35 years old with 10 years of experience.

i. Construct and interpret a 95% prediction interval for an employee with a Master's degree who is 35 years old with 10 years of experience. How useful is this interval?

j. Construct and interpret a 95% confidence interval for the average salary of an employee with a PhD who is 35 years old with 10 years of experience. How useful is this interval?

Note

To complete parts **i.** and **j.**, you will need to use Minitab, R, or Rguroo.

12. Consider the following crime data from select college campuses. The table contains the number of crimes committed, the number of campus police employed on campus, the total enrollment of the college, and whether or not the college is private. The full data set is available on the companion site.

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Campus Crime

Campus Crime Data				
School	Number of Crimes	Number of Police	Total Enrollment	Private School
1	64	12	1131	Yes
2	138	21	12,954	No
3	141	32	16,009	No
4	84	22	1682	Yes
5	86	35	2888	Yes
...				

a. Create an indicator (dummy) variable for whether or not the college is private. Let *Private* = 1 if the school is private and *Private* = 0 if the school is public.

b. Suppose education officials wish to predict the number of crimes on college campuses based on the number of police employed and total enrollment. They would also like to know whether there are fewer crimes committed on private campuses than public ones. Use statistical software to estimate the following regression model.

$$Crimes = \beta_0 + \beta_1\, Police + \beta_2\, Enrollment + \beta_3\, Private + \varepsilon$$

Write the estimated multiple regression equation.

c. Is the overall model useful in predicting the number of crimes? Use $\alpha = 0.05$.

d. Are the signs of the coefficients of the independent variables what you would expect for these data? Explain.

e. Is there evidence to support the officials' belief that there are fewer crimes committed at private schools than at public schools? Test using $\alpha = 0.05$. Would this decision change if $\alpha = 0.01$?

13. You wish to develop a model to analyze if the manufacturer influences the price of used cars, using data on six-year-old cars produced by three Japanese manufacturers: Honda, Nissan, and Toyota. Your data consists of number of doors (2 or 4), curb weight, engine size, city mpg, highway mpg, and price.

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Car Prices

Car Price Data							
Car	Manufacturer	Number of Doors	Curb Weight	Engine Size	City MPG	Highway MPG	Price
1	Toyota	2	1985	92	35	39	5348
2	Honda	2	1837	79	38	42	5399
3	Nissan	2	1889	97	31	37	5499
4	Toyota	2	2040	92	31	38	6338
5	Honda	2	1713	92	49	54	6479
6	Toyota	4	2015	92	31	38	6488
©HAWKES LEARNING ...							

a. Define indicator variable(s) that could be used to code the 3 Japanese manufacturers.

b. Write the proposed model using all the variables in the data including the indicator variable(s) developed in part **a.**

c. What variables in the model are significant at the 0.05 level? Which are not significant?

d. What proportion of the variation in car prices is explained by the model?

e. If you eliminate the variables that are not significant and rebuild the model, how much of the variation in price will you explain?

14. Consider the following sales data regarding weekly sales, the number of sales reps, and whether or not the sales were made in the first, second, third, or fourth quarter of the year. For each column containing an indicator variable, the variable is equal to 1 if that particular week was in that particular quarter, and equal to zero otherwise. For example, if the weekly data were recorded in January, the 1st quarter indicator variable would be equal to 1 and the indicator variables for the 2nd, 3rd, and 4th quarters would be equal to zero. The first quarter comprises January through March, the second quarter April through June, the third quarter July through September, and the fourth quarter October through December.

Weekly Sales by Quarter

Weekly Sales ($)	Number of Sales Reps	Quarter
4272.90	3	1
5069.70	9	1
6067.70	11	1
6680.55	17	1
9725.05	20	1
4107.10	3	2
7520.25	9	2
12,135.00	11	2
13,016.55	17	2
13,673.90	20	2
3272.05	3	3
5074.40	9	3
7505.45	11	3
8272.75	17	3
10,020.40	20	3
4925.75	3	4
10,018.10	9	4
12,505.85	11	4
15,329.05	17	4
19,477.20	20	4

a. How many indicator variables should be included in the multiple regression model relating weekly sales to the number of sales reps and the quarter of the year? Explain why.

b. What sign would you expect the coefficient for the sales reps variable to have? Explain your reasoning.

c. Using statistical software, estimate the following multiple regression model.
$\text{Sales} = \beta_0 + \beta_1(\text{Reps}) + \beta_2(\text{Quarter 1}) + \beta_3(\text{Quarter 2}) + \beta_4(\text{Quarter 3}) + \varepsilon_i$
Write the estimated multiple regression equation.

d. Interpret the coefficient of the indicator variable representing the first quarter.

e. Is there sufficient evidence that sales in the second quarter tend to be different from the sales in the fourth quarter? Use $\alpha = 0.05$.

f. What concerns should we have when predicting weekly sales using this model?

CR Chapter Review

Key Terms and Ideas

- Multiple Regression
- Multiple Regression Model
- Method of Least Squares
- Estimated Multiple Regression Equation
- Coefficient of Determination (Multiple Coefficient of Determination)
- Adjusted R^2
- F-Distribution
- Numerator Degrees of Freedom
- Denominator Degrees of Freedom
- F-Statistic
- Sum of Squares of Regression
- Sum of Squared Errors
- Total Sum of Squares
- Mean Square Regression
- Mean Square Error
- Calculating Degrees of Freedom in Multiple Regression Models
- Hypothesis Tests Concerning Individual Coefficients
- Test Statistic for Testing the Hypothesis $\beta_i \neq 0$
- Confidence Intervals for Individual Coefficients
- Confidence Interval for the Mean Value of y Given x
- Confidence Interval for the Predicted Value of y Given x
- Indicator (Dummy) Variable
- Base Level Variable
- Interaction Terms
- Polynomial (Nonlinear) Regression Models

Key Formulas

Section	
14.1	**Multiple Regression Model** $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$ where β_0, β_1, β_2, ..., β_k are the model's parameters, x_1, x_2, ..., x_k are the independent variables, and ε is a random error.

15.1 Exercises

Basic Concepts

1. Give two examples where you might be interested in comparing several population means.
2. **a.** What are experimental units? **b.** In the agricultural experiment in the beginning of 15.1, what were the experimental units?
3. **a.** What is a treatment? **b.** In the agricultural experiment in the beginning of 15.1, what was the treatment?
4. Explain how box plots can be useful in analyzing data when comparing population means.
5. How is the total variation broken down in an analysis of variance?
6. What does the total sum of squares describe? What are its degrees of freedom?
7. What is the mathematical expression for the sum of squares for treatments?
8. What is the mean square for treatments?
9. What is the relationship between the Total Sum of Squares, SST, and SSE? Explain why this relationship makes sense.
10. Why is it important to validate the assumptions upon which a hypothesis test is based?
11. What are the assumptions for an ANOVA F-test?
12. If you found that MST is much larger than MSE, would you tend to think that the population means were similar or different? Explain how this ratio brings you to this conclusion.
13. If the variability among the sample means is very similar to the variability among the sample observations, what value will F be close to? Explain why.
14. Is the null hypothesis generally rejected for large or small values of the F-statistic? Explain why this is the case.
15. What are the null and alternative hypotheses for the one-way ANOVA F-test?
16. Explain the completely randomized experimental design.
17. Why would data derived from an experimental design likely be better data than observational data?

Exercises

18. Consider the following table containing daily production data from a particular week for three different employee shifts.

Items Produced			
	First Shift (7 AM-3 PM)	Second Shift (3 PM-11 PM)	Third Shift (11 PM-7 AM)
Monday	140	168	77
Tuesday	181	224	123
Wednesday	127	162	77

Thursday	172	182	101
Friday	161	219	147
Saturday	152	171	145
Sunday	173	217	111

a. Identify the experimental units and the treatment in the context of this problem.

b. Compute the mean and median numbers of items produced for each shift.

c. Compute the values of the minimum, maximum, first, and third quartiles for each shift.

d. Construct side-by-side box plots for the three shifts.

e. Based on the box plots, do you think that there may be a significant difference in the average numbers of items produced during the first and second shifts? Explain.

f. Based on the box plots, do you think that there may be a significant difference in the average numbers of items produced during the second and third shifts? Explain.

g. Based on the box plots, do you think that there may be a significant difference in the average numbers of items produced during the first and third shifts? Explain.

h. Based on your analysis, which shift would you say is the most productive, on average? Explain your answer.

19. Consider the production data given in Exercise 18.

a. What is the value of the grand mean, $\bar{\bar{x}}$?

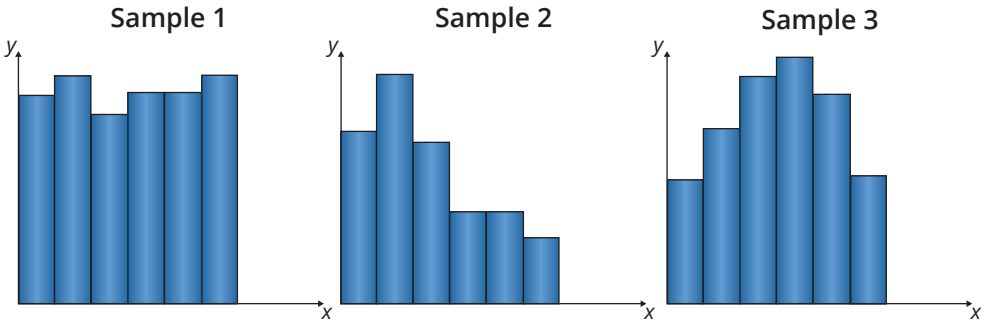
b. What is the value of n_2?

c. What is the value of k?

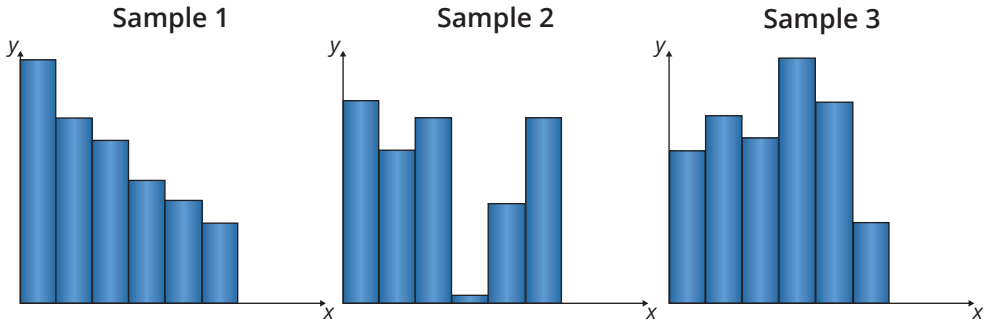
d. What is the value of N?

e. For this data, identify the degrees of freedom associated with the total sum of squares, the degrees of freedom associated with the sum of squares for treatments, and the degrees of freedom associated with the sum of squares for error. Verify that the relationship between the degrees of freedom (Total = Treatment + Error) holds.

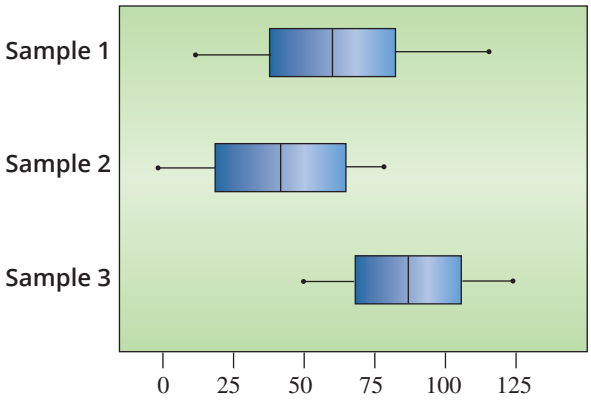
20. For each of the following histograms of sample data, decide whether or not you think it is reasonable to assume that the data was drawn from a population that has an approximately normal distribution.



21. For each of the following histograms of sample data, decide whether you think it is reasonable to assume that the data was drawn from a population that has an approximately normal distribution.

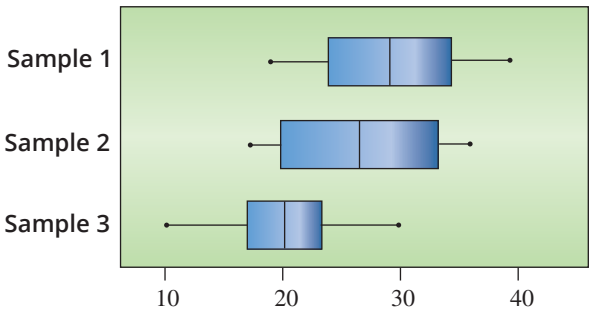


22. Consider the following box plots.



Do you think it is reasonable to assume that the three populations represented by the sample data in these box plots have equal variances? Explain.

23. Consider the following box plots.



Do you think it is reasonable to assume that the three populations represented by the sample data in these box plots have equal variances? Explain.

24. The results of a comparison of four popular minivans are reported in the following table. One of the features the researchers compared was the distance (in feet) required for the minivan to come to a complete stop when traveling at a speed of 60 miles per hour (braking distance). Suppose the braking distances were measured for five minivans of each type with the following results.

Braking Distances (Feet)			
Minivan A	Minivan B	Minivan C	Minivan D
150	153	155	167
152	150	150	164
151	156	157	169
149	151	158	162
153	155	155	173

a. Can the researchers conclude at $\alpha = 0.10$ that there is a difference among average braking distances for the four minivan models?

b. What assumptions did the researchers make in performing the test procedure in part **a.**? Does the data appear to satisfy these assumptions? Explain.

25. A steel company is considering the relocation of one of its manufacturing plants. The company's executives have selected four areas that they believe are suitable locations. However, they want to determine if the average wages are significantly different in any of the locations, since this could have a major impact on the cost of production. A survey of hourly wages of similar workers in each of the four areas is performed with the following results.

Hourly Wages ($)			
Area 1	Area 2	Area 3	Area 4
13	18	16	23
15	19	17	19
14	21	18	21
16	20	18	20
13	17	15	19

a. Does the data indicate a significant difference among the average hourly wages in the four areas at $\alpha = 0.05$?

b. What assumptions were made in performing the test in part **a.**? Does the data appear to satisfy these assumptions? Explain.

26. Consider the following table containing yields for mutual funds in different asset classes (small, mid, and large cap).

Fund Yield by Asset Class					
Small Cap		Mid Cap		Large Cap	
Fund	Yield (%)	Fund	Yield (%)	Fund	Yield (%)
Explorer Value	2.04	Capital Value	0.96	Equity Income	3.24
Small-Cap Value Index Admiral	2.46	Mid-Cap Value Index Admiral	1.57	High Dividend Yield Index	3.50
Small-Cap Index Admiral Shares	1.49	Extended Market Index Admiral Shares	1.22	500 Index Admiral Shares	1.57

Strategic Small-Cap Equity	0.38	Mid-Cap Index Admiral Shares	1.52	Diversified Equity	1.23
Explorer	0.17	Mid-Cap Growth	2.76	FTSE Social Index	1.42
Small-Cap Growth Index Admiral	0.21	Capital Value	0.32	Growth Equity	2.52
Explorer Value	2.55	Strategic Equity	1.54	U. S. Growth	0.37
Samll-Cap ETF	1.44	Capital Opportunity Admiral Shares	2.14	Windsor	1.64

Sum of squares for treatments ≈ 1.5986

Sum of squares for error ≈ 18.4205

a. What are the degrees of freedom associated with the sum of squares for treatments?

b. Find the mean square for treatments. Round your answer to two decimal places, if necessary.

27. A physical trainer has four workouts that he recommends for his clients. The workouts have been designed so that the average maximum heart rate achieved is the same for each workout. To test this design, he randomly selects 12 people and randomly assigns three of them to use each of the workouts. During each workout, he measures the maximum heart rate in beats per minute with the following results.

Maximum Heart Rates (Beats per Minute)			
Workout #1	**Workout #2**	**Workout #3**	**Workout #4**
180	160	175	185
185	170	180	190
170	175	170	180

a. Can the physical trainer conclude at $\alpha = 0.05$ that there is a difference among the average maximum heart rates which are achieved during the four workouts?

b. What assumptions did the physical trainer make in performing the test procedure in part **a.**? Does the data appear to satisfy these assumptions? Explain.

28. The results of a survey comparing the costs of staying one night in a full-service hotel (including food, beverages, and telephone calls, but not taxes or gratuities) for several major cities are given in the following table.

Hotel Costs per Night ($)				
New York	**Los Angeles**	**Atlanta**	**Houston**	**Phoenix**
300	240	190	195	238
320	250	198	190	240
325	230	185	200	236
350	245	195	192	248
275	235	182	198	228

a. Does the data suggest that there is a significant difference among the average costs of one night in a full-service hotel for the five major cities at $\alpha = 0.05$?

b. What assumptions were made in performing the test procedure in part **a.**? Does the data appear to satisfy these assumptions? Explain.

c. Based on the analysis you performed in part **b.**, which cities, if any, do you think have significantly different average costs for a one-night stay in a full-service hotel? Explain.

29. Consider the following information regarding the dividends paid per share by companies in the banking, transportation, and energy industries.

Dividends per Share ($)		
Banking	**Transportation**	**Energy**
1.52	1.00	2.08
3.12	1.20	2.68
1.32	0.20	0.70
0.60	0.40	2.00
1.20	1.09	1.91
1.00	0.61	1.60
1.19	0.35	1.28

a. Does the data provide sufficient evidence to conclude that there is a significant difference among the average dividends paid per share for the three different industries? Use $\alpha = 0.10$.

b. What assumptions were made in performing the test procedure in part **a.**? Does the data appear to satisfy these assumptions? Explain.

c. Based on the analysis you performed in part **b.**, which industries, if any, do you think pay significantly different average dividends per share? Explain.

30. The sales strategy data given below yields the following statistics for the sum of squares for treatments and the sum of squares for error.

SST = 24.875

SSE = 579.1

Sales by Strategy (Millions of Dollars)			
Strategy 1	**Strategy 2**	**Strategy 3**	**Strategy 4**
15	5	3	8
8	4	12	6
4	14	4	7
7	16	9	5
10	9	9	4
4	5	10	13
5	7	5	5
9	8	5	8
14	3	6	13
2	16	2	10

a. What are the degrees of freedom associated with the total sum of squares?
b. What are the degrees of freedom associated with the sum of squares for treatments, SST?
c. Find the mean square for treatments, MST. Round your answer to four decimal places.
d. Find the mean square for error, MSE. Round your answer to four decimal places.

15.2 Multiple Comparison Procedures

In the previous section, we used one-way ANOVA to test whether differences existed between population means. In the earlier examples, when we rejected the null hypothesis that all of the population means were equal, we were only testing if differences existed. However, the results of the one-way ANOVA test do not indicate which population means are different. To determine which population means are different, we need to perform more tests to determine if there are statistically significant differences between two population means such as $\mu_1 - \mu_2$, for example. Multiple comparison procedures present several options to the analyst when comparing means after finding significance when performing a one-way ANOVA. The ones we will consider are Fisher's Least Significance Difference (LSD) Method, Tukey's Honest Significant Difference (HSD), and performing *t*-tests to make pairwise comparisons between means.

Data were collected to study the use of media by three age groups–tweens (8–12 years old), teens (13–18 years old), and adults (over 18 years old). A sample of 50 participants in each age group was asked how frequently they engaged in activities such as time spent on cell phones, watching online videos, watching television, and playing mobile games. The table below presents some summary statistics from the data collected showing the sample size, the sample mean time spent on devices, and the sample standard deviation of the data by age group.

Table 15.2.1 – Screen Time Summary Statistics by Age Group

	Tweens 8–12 Years Old	Teens 13–18 Years Old	Adults Over 18 Years Old
n	50	50	50
Mean	128	137.48	100.12
Standard Deviation	20.90	25.57	20.32

Test to determine if there is a significant difference between average time spent on devices among the three age groups.

Let

μ_1 = population mean time spent on devices for participants 8–12 years old,

μ_2 = population mean time spent on devices for participants 13–18 years old, and

μ_3 = population mean time spent on devices for participants over 18 years old.

Data

This data set can be found on stat.hawkeslearning.com under **Discovering Statistics and Data, Fourth Edition > Data Sets > Screen Time by Age Group.**

Comparisons for all pairs using Tukey-Kramer HSD

Confidence Quantile

q*	Alpha
2.36773	0.05

HSD Threshold Matrix

Abs(Dif)-HSD

	13-18 Years Old	8-12 Years Old	Over 18 Years Old
13-18 Years Old	-10.601	-1.121	26.759
8-12 Years Old	-1.121	-10.601	17.279
Over 18 Years Old	26.759	17.279	-10.601

Positive values show pairs of means that are significantly different.

Connecting Letters Report

Level		Mean
13-18 Years Old	A	137.48000
8-12 Years Old	A	128.00000
Over 18 Years Old	B	100.12000

Levels not connected by same letter are significantly different.

Ordered Differences Report

Level	-Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
13-18 Years Old	Over 18 Years Old	37.36000	4.477358	26.7588	47.96118	< .0001*
8-12 Years Old	Over 18 Years Old	27.88000	4.477358	17.2788	38.48118	< .0001*
Over 18 Years Old	8-12 Years Old	9.48000	4.477358	-1.1212	20.08118	0.0898

Figure 15.2.2

It is interesting to note that the two multiple comparison procedures (Fisher's LSD and Tukey's HSD) in the aforementioned example do not produce the same results. This is not uncommon. Specifically, as the number of comparisons increases, the probability of committing a Type I Error increases when using Fisher's LSD. However, regardless of the number of comparisons being made, the probability of committing a Type I Error remains the same (i.e., α) when using Tukey's HSD.

Note

Note that the confidence intervals in the Ordered Differences Report vary slightly from our previous calculations due to the exact q-distribution critical value being used instead of the value from the table with error $df = \infty$.

15.2 Exercises

Basic Concepts

1. What is the purpose of multiple comparison procedures?
2. When should multiple comparison procedures be used?
3. What are the hypotheses tested if there are four population means in the ANOVA?
4. Define the concepts of balanced and unbalanced data when conducting a test to compare the pairwise sample means for a given set of samples.

Exercises

5. How many individual pairwise comparisons would need to be made if there are four population means in the ANOVA? What would be the probability of at least one Type I error if performing individual pairwise comparisons at a 0.01 significance level?
6. A two-sample t-test is conducted to test the pairwise differences in the mean number of candies consumed per family (average size of four family members) per day. The families belong to four different states. The following output is obtained (differences are computed in the given order of the states).

	Null hypothesis	Difference in Means	*t*-Test Statistic	*P*-value
Alabama and Los Angeles	$(H_0: \mu_A - \mu_{LA} = 0)$	5.6378	2.3479	0.046835
New York and Los Angeles	$(H_0: \mu_{NY} - \mu_{LA} = 0)$	−12.5798	7.8741	0.000049
Alabama and Texas	$(H_0: \mu_A - \mu_T = 0)$	41.2156	12.3721	0.000002
New York and Texas	$(H_0: \mu_{NY} - \mu_T = 0)$	0.4132	1.1553	0.281305
Texas and Los Angeles	$(H_0: \mu_T - \mu_{LA} = 0)$	−24.8714	9.2496	0.000015
Alabama and New York	$(H_0: \mu_A - \mu_{NY} = 0)$	32.6741	11.7420	0.000003

Assuming a significance level of $\alpha = 0.05$, answer the following questions.

a. Is there evidence to conclude that, on average, families in Alabama consume more candies per day than families in New York?

b. Which state appears to have the highest candy consumption per family per day according to the output.

7. Fisher's Least Significant Difference method examines the pairwise difference in the mean values of four treatment groups at a 0.05 level of significance. Determine the critical value if the total number of observations in all the samples is 30.

8. The number of paint defects found in a sample of 50 cars produced by three different car manufacturers (labeled A, B and C) are studied. The analysis of variance was significant at the 0.05 level indicating a difference in the average number of paint defects among the car manufacturers. Determine which car manufacturers are different using Fisher's Least Significant Difference method. Assume that the value calculated for Fisher's LSD is 4.4763, which is the same for each pair.

The following table shows the sample mean number of paint defects for each of the manufacturers.

Manufacturer	Mean Number of Paint Defects
A	7
B	12
C	9

9. The mean effect of three treatments on fasting blood sugar levels for three samples of 10 patients are shown below.

Treatments	Mean Fasting Blood Glucose Levels (mg/dL)
A	87.5
B	86.5
C	78.2

The ANOVA output for this experiment using R is as follows.

	df	Sum of Squares	Mean Square	*F*-value	*Pr(>F)*
Treatment	2	521.3	260.6	2.549	0.0968
Residuals	27	2760.6	102.2		

Assuming the level of significance is $\alpha = 0.10$, compare the pairwise differences in the mean blood glucose level for the three treatments using Fisher's Least Significant Difference method.

10. List one advantage of Tukey's HSD method over the two-sample *t*-test when the pairwise differences between the sample means are to be examined.

11. Compute the studentized range value for conducting Tukey's HSD test when the level of significance is equal to 0.05, the number of treatments is equal to 4, and the sample size of each of the four samples is equal to 16.

12. The cholesterol level of a total of 45 subjects is measured. The subjects were randomly divided into three groups and given different doses of medication (0 mg, 5 mg, 10 mg).

The one-way ANOVA table for testing if there is a significant difference in the mean cholesterol level for the different doses of medication is shown below.

	df	Sum of Squares	Mean Square	*F*-value	*Pr*(>*F*)
Dosage	2	53402	26701	3.57566	0.036813
Residuals	42	313632	7467.42857		

Is it wise to conduct a Tukey's HSD test to compare the difference in the mean cholesterol level at the following levels of significance?

a. 1%

b. 5%

13. Consider the test scores of a group of 15 students divided into three samples based on the type of curriculum studied. The following output is obtained after conducting a one-way ANOVA test.

	df	Sum of Squares	Mean Square	*F*-value	*Pr*(>*F*)
Curriculum	2	1301.7	650.9	28.18	2.93 E-05
Residuals	12	277.2	23.1		

The mean scores for the three samples are tabulated below.

Type of Curriculum	Mean Test Score
A	87.2
B	76.6
C	64.4

Determine if the mean tests scores are different for the following curriculum types using the confidence interval approach for Tukey's HSD with a 0.05 level of significance.

a. Sample A and Sample B

b. Sample B and Sample C

15.3 Two-Way ANOVA: The Randomized Block Design

In two-way ANOVA (analysis of variance), there are two independent variables that are being analyzed to determine their effects on a dependent variable. These two independent variables are often referred to as **factors**, which are called **treatments** and **blocks**. Treatments represent the main variable (factor) in the study. Researchers may be interested in the effects of both independent variables but are frequently interested in only one of them (treatments) and use the **randomized block design** to eliminate

15.3 Exercises

Basic Concepts

1. What is a completely randomized design? Give an example.
2. What are blocks? What is their purpose?
3. What is a randomized block design? How is it different from a completely randomized design?
4. What are the null and alternative hypotheses when comparing means using a randomized block design?
5. What is the breakdown of the sum of squares for a randomized block design? Does this breakdown make sense? Explain.
6. If blocking is successful, how does the value of SSE change?
7. What are the assumptions when performing a two-way ANOVA for a randomized block design?
8. What is the rationale for the test statistic used for the randomized block design?

Exercises

9. A car dealer is interested in comparing the average gas mileages of four different car models. The dealer believes that the average gas mileage of a particular car will vary depending on the person who is driving the car due to different driving styles. Because of this, he decides to use a randomized block design. He randomly selects six drivers and asks them to drive each of the cars. He then determines the average gas mileage for each car and each driver. The results of the study are as follows.

Gas Mileage (MPG)				
	Car A	**Car B**	**Car C**	**Car D**
Driver 1	33	29	27	37
Driver 2	36	32	30	40
Driver 3	34	30	28	38
Driver 4	31	27	25	35
Driver 5	33	29	27	37
Driver 6	35	33	31	41

   a. Identify the dependent variable, the treatment variable, and the blocking variable.

   b. Do you think a randomized block design is appropriate for the car dealer's study? Explain.

   c. The results of the two-way ANOVA for the dealer's survey of the average gas mileages of the different car models are given in the following table. Can the dealer conclude that there is a significant difference in average gas mileages of the four car models? Use $\alpha = 0.05$.

ANOVA

Source of Variation	*SS*	*df*	*MS*
Rows	84.8333	5	16.9667
Columns	348.5000	3	116.1667
Error	2.5000	15	0.1667
Total	435.8333	23	

d. Was the dealer able to significantly reduce variation among the observed gas mileages by blocking? Use $\alpha = 0.05$.

10. A banana grower has three fertilizers from which to choose. He would like to determine which fertilizer produces banana trees with the largest yield (measured in pounds of bananas produced). The banana grower has noticed that there is a difference in the average yields of the banana trees depending on which side of the farm they are planted (South Side, North Side, West Side, or East Side). Because of the variation in yields among the areas on the farm, he has decided to randomly select three trees within each area and then randomly assign the fertilizers to the trees. After harvesting the bananas, he calculates the yields of the trees within each of the areas. The results are as follows.

Banana Yields (Pounds)			
	Fertilizer A	**Fertilizer B**	**Fertilizer C**
South Side	53	51	58
North Side	48	47	53
West Side	50	48	56
East Side	50	47	54

a. Identify the dependent variable, the treatment variable, and the blocking variable.

b. Do you think a randomized block design is appropriate for the banana grower's study? Explain.

c. The results of the two-way ANOVA for the banana grower's study are given in the following table. Can the banana grower conclude that there is a significant difference among the average yields of the banana trees for the three fertilizers? Use $\alpha = 0.10$.

ANOVA

Source of Variation	*SS*	*df*	*MS*
Rows	36.2500	3	12.0833
Columns	104.0000	2	52.0000
Error	2.0000	6	0.3333
Total	142.2500	11	

d. Was the banana grower able to significantly reduce variation among the observed yields by blocking? Use $\alpha = 0.10$.

11. The FAA is interested in knowing if there is a difference in the average numbers of on-time arrivals for four of the major airlines. The FAA believes that the number of on-time arrivals varies by airport. To control for this variation, they randomly select 100 flights for each of the major airlines at each of four randomly selected airports and record the number of on-time flights. The results of the study are as follows.

On-Time Flights				
	Airline A	**Airline B**	**Airline C**	**Airline D**
Airport A	87	82	79	81
Airport B	88	84	81	82
Airport C	89	84	83	82
Airport D	90	86	85	83

a. Identify the dependent variable, the treatment variable, and the blocking variable.

b. Do you think a randomized block design is appropriate for the FAA's study? Explain.

c. The results of the two-way ANOVA for the FAA's study are given in the following table. Can the FAA conclude that there is a significant difference among the average number of on-time arrivals for the four major airlines? Use $\alpha = 0.01$.

ANOVA

Source of Variation	*SS*	*df*	*MS*
Rows	29.2500	3	9.7500
Columns	112.7500	3	37.5833
Error	5.7500	9	0.6389
Total	147.7500	15	

d. Was the FAA able to significantly reduce variation among the observed number of on-time arrivals by blocking? Use $\alpha = 0.01$.

12. A psychologist is interested in determining if there is a difference in the average numbers of patients for several age groups. The psychologist believes that there may be some variation in the numbers of patients depending on the region of the country (Northeast, Northwest, Southeast, or Southwest). The psychologist randomly selects 100,000 individuals from each region of the country for each of the age groups of interest and determines the number of psychology patients. The results of the study are as follows.

Psychology Patients							
	Age 15–24	**Age 25–34**	**Age 35–44**	**Age 45–54**	**Age 55–64**	**Age 65–74**	**Age 75–84**
Northeast	15	17	16	17	55	22	27
Northwest	13	16	16	16	49	19	26
Southeast	12	14	15	15	47	17	24
Southwest	13	15	15	16	53	20	25

a. Identify the dependent variable, the treatment variable, and the blocking variable.

b. Do you think a randomized block design is appropriate for the psychologist's study? Explain.

c. The results of the two-way ANOVA for the psychologist's study are given in the following table. Can the psychologist conclude that there is a significant difference among the average number of patients for the different age groups? Use $\alpha = 0.10$.

ANOVA

Source of Variation	*SS*	*df*	*MS*
Rows	44.9643	3	14.9881
Columns	4223.3571	6	703.8929
Error	25.7857	18	1.4325
Total	4294.1071	27	

d. Was the psychologist able to significantly reduce variation among the observed number of psychology patients by blocking? Use $\alpha = 0.05$.

13. In an experiment designed to compare automated blood pressure devices with those of the standard cuff method, each person in a sample of six patients has their systolic blood pressure determined by three different automated devices and by the standard cuff method. The data is given in the following table.

Blood Pressure (mm Hg)				
	Device 1	**Device 2**	**Device 3**	**Standard Cuff**
Patient 1	126	128	132	131
Patient 2	134	138	137	140
Patient 3	145	144	150	152
Patient 4	129	134	132	136
Patient 5	154	160	162	160
Patient 6	144	144	148	145

a. Identify the dependent variable, the treatment variable, and the blocking variable.

b. Why was a randomized block design used in this experiment?

c. From the data, SST and SSE were computed to be 106.4583 and 53.2917, respectively. With $\alpha = 0.05$, can we conclude that the four different methods of determining systolic blood pressure have different mean readings?

d. SSBL was computed to be 2412.8750. With $\alpha = 0.05$, can we conclude that using people as blocks significantly reduced variation in this study?

15.4 Two-Way ANOVA: The Factorial Design

The randomized block test presented in the previous section is one example of a **two-way ANOVA**. There were two independent factors considered in the analysis, namely the block (the different weight classes) and the treatment (diets), and each level of the treatment occurred with each level of the block. However, there was only one factor which was truly of interest to our experimenter, the treatment (diets).

The techniques discussed for the randomized block design can be extended to the situation where there are two factors of interest. The main features of the factorial design are:

From Figure 15.4.4, the calculated value of the test statistic is 13.4435 and the P-value is 0.00038. Figure 15.4.7 shows the rejection region. Since 13.4435 is larger than 3.4903 and the P-value is less than $\alpha = 0.05$, we reject the null hypothesis that age has no effect on average salary. There is persuasive evidence at $\alpha = 0.05$ that the mean salaries are significantly affected by age.

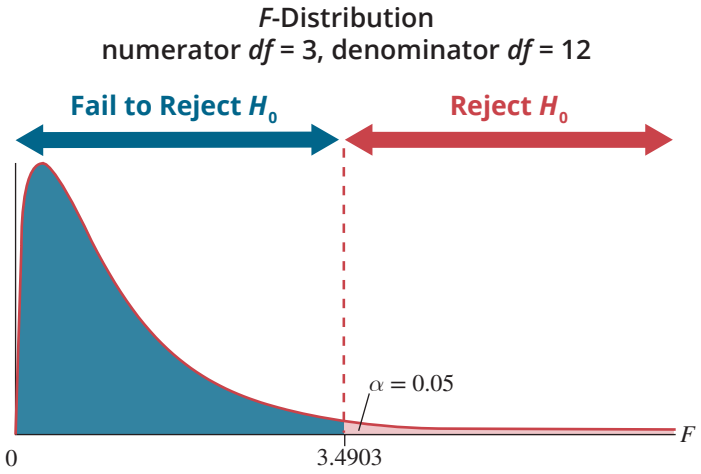


Figure 15.4.7

When using the two-way ANOVA, it is important to remember that the assumptions of normality, equal variances, and independent random samples should be satisfied in order for the test to produce meaningful results. As noted previously, the ANOVA test is robust with regards to the normality and equal variance assumptions.

An important concept in statistics related to ANOVA is the **design of experiments** (DOE). It is much easier to analyze a properly designed experiment as opposed to a poorly designed one. If you are planning on collecting data, you should involve a statistician at the onset, not just employ them to do the analysis once the data has been collected. Experimental design allows you to estimate the effects of several variables simultaneously, thus resulting in a more efficient collection of data that will be much easier to analyze. Design of experiments is a very extensive field of study for which entire books and courses have been developed, but is beyond the scope of this book.

15.4 Exercises

Basic Concepts

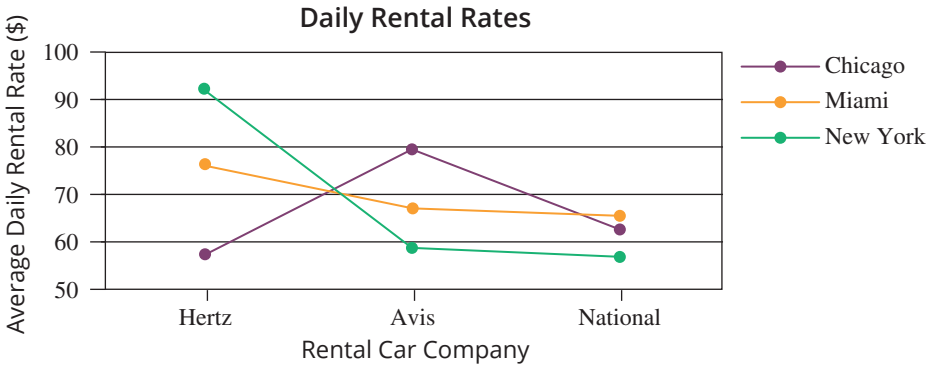
1. What is the difference between a randomized block design and a factorial design?
2. What is a complete factorial experiment?
3. What is an interaction plot? What kind of information does this plot give us?
4. Why is it important to determine if there is interaction between the two variables of interest in a factorial design?
5. Is it possible to perform a two-way analysis of variance if interaction exists between the two variables of interest? Explain why or why not.
6. Identify the four components that make up the total sum of squares in a complete factorial model.

7. Give the degrees of freedom associated with each component of the total sum of squares.
8. What is the test statistic for a test of interaction between factors? What are the degrees of freedom associated with this test statistic?
9. If there is enough evidence to reject the null hypothesis in a test for interaction, may we proceed with the main effects tests? Explain.
10. What is the test statistic for the main effects test for Factor A? What are the degrees of freedom associated with this test statistic?
11. What is the test statistic for the main effects test for Factor B? What are the degrees of freedom associated with this test statistic?
12. What are the rejection rules for the main effects tests? Can *P*-values be used as rejection criteria?

Exercises

13. The following table contains the results of a survey of daily rental rates of a mid-size car for three major rental car companies at three airport locations on three different days during the year.

Daily Rental Rates of Mid-Size Cars ($)			
	New York	**Chicago**	**Miami**
	93.99	54.99	71.99
	90.99	63.99	87.99
Hertz	96.99	57.99	68.99
	58.86	81.99	61.99
	52.10	85.99	70.99
Avis	68.98	71.99	66.99
	56.00	64.99	66.00
	63.00	67.00	58.99
National	52.00	52.99	71.99



a. Identify the dependent variable. Identify the two factors. In the associated computer output, which variable is Factor A and which variable is Factor B?

b. Consider the graph of the average daily rental rates for each of the major car rental companies by airport location. Does there appear to be any interaction between the variables airport location and major car rental company?

c. The results of the two-way ANOVA for the study are given in the following table. Perform a hypothesis test to determine if there is any interaction between the variables major rental car company and airport location at $\alpha = 0.05$. Does this agree with your observation in part **b.**?

ANOVA

Source of Variation	*SS*	*df*	*MS*
Sample	1011.7730	2	505.8865
Columns	58.7126	2	29.3563
Interaction	2514.3099	4	628.5775
Within	819.1289	18	45.5072
Total	4403.9244	26	

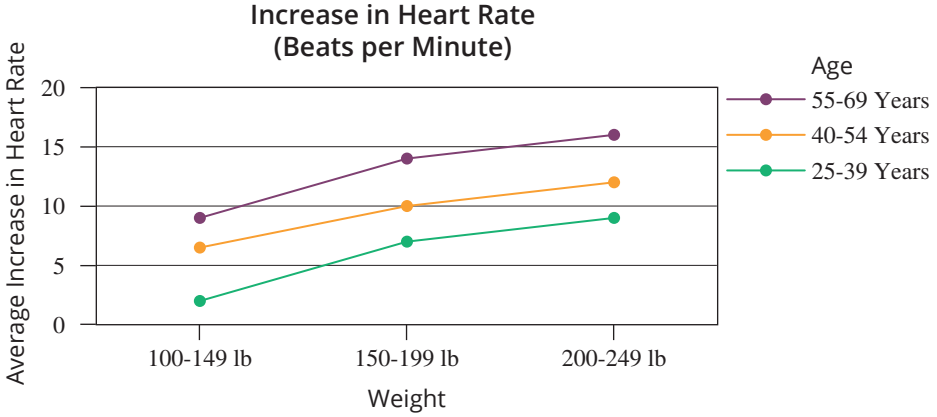
d. If there is no interaction found in part **c.**, is there sufficient evidence to conclude that there is a significant difference among the average daily rental rates for mid-size cars for the three rental car companies at the 0.05 level?

14. A doctor is interested in determining the increase in average heart rate caused by a medication used for treating high blood pressure. The doctor believes that the increase in heart rate will be related to two factors: the age of a person and the weight of a person. To test this theory, the doctor randomly selects two patients in each of the age and weight categories listed in the following table and determines the increase in heart rate (in beats per minute) of each patient 15 minutes after administering the drug. The results of the study are as follows.

Increase in Heart Rate (Beats per Minute)			
	25–39 Years	**40–54 Years**	**55–69 Years**
100–149 Pounds	2	7	11
	2	6	7
150–199 Pounds	7	11	16
	7	9	12
200–249 Pounds	10	13	18
	8	11	14

a. Identify the dependent variable. Identify the two factors. In the associated computer output, which variable is Factor A and which variable is Factor B?

b. Consider the following graph of the average increase in heart rate for each of the weight and age categories. Does there appear to be any interaction between the age and weight variables? Explain.



c. The results of the two-way ANOVA for the study are given in the following table. Perform a hypothesis test to determine if there is any interaction between the variables age and weight at $\alpha = 0.01$. Does this agree with your observation in part **b.**?

ANOVA

Source of Variation	*SS*	*df*	*MS*
Sample	133.0000	2	66.5000
Columns	147.0000	2	73.5000
Interaction	2.0000	4	0.5000
Within	30.5000	9	3.3889
Total	312.5000	17	

d. Is there sufficient evidence to conclude that there is a significant difference among the average increases in heart rate for the different weight categories? Use $\alpha = 0.01$.

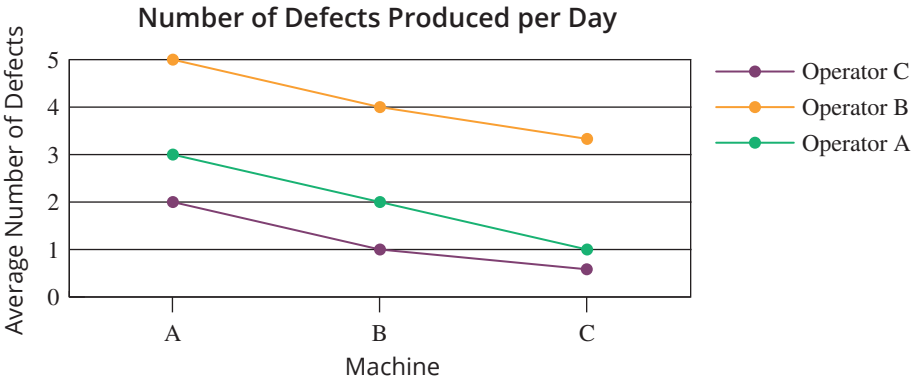
e. Is there sufficient evidence to conclude that there is a significant difference among the average increases in heart rate for the different age groups? Use $\alpha = 0.01$.

15. A supervisor of a manufacturing plant is interested in relating the average number of defects produced per day to two factors: the operator working the machine and the machine itself. The supervisor randomly assigns each operator to use each machine for three days and records the number of defects produced per day. The results of the study are as follows.

Number of Defects Produced per Day			
	Operator A	**Operator B**	**Operator C**
	3	7	3
	3	5	2
Machine A	3	3	1
	2	6	2
	2	4	1
Machine B	2	2	0
	1	5	1
	1	3	0
Machine C	1	2	1

a. Identify the dependent variable. Identify the two factors. In the associated computer output, which variable is Factor A and which variable is Factor B?

b. Consider the following graph of the average number of defects produced per day for each of the operators by machine. Does there appear to be any interaction between the variables operator and machine?



c. The results of the two-way ANOVA for the supervisor's survey of the number of defects produced per day are given in the following table. Perform a hypothesis test to determine if there is any interaction between the machine and operator variables. Use $\alpha = 0.10$. Does this agree with your observation in part **b.**?

ANOVA

Source of Variation	*SS*	*df*	*MS*
Sample	12.6667	2	6.3333
Columns	40.2222	2	20.1111
Interaction	0.4444	4	0.1111
Within	25.3333	18	1.4074
Total	78.6667	26	

d. Is there sufficient evidence to conclude that there is a significant difference among the average number of defects produced per day for the different machines? Use $\alpha = 0.10$.

e. Is there sufficient evidence to conclude that there is a significant difference among the average number of defects produced per day for the different operators? Use $\alpha = 0.10$.

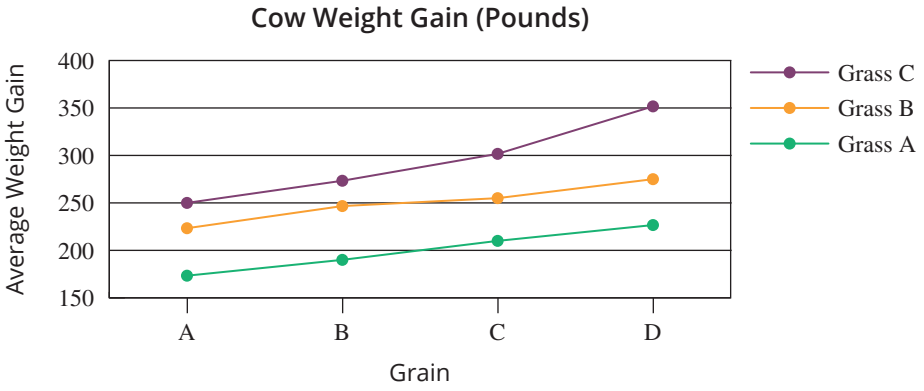
16. A dairy farmer thinks that the average weight gain of his cows depends on two factors: the type of grain that they are fed and the type of grass that they are fed. The dairy farmer has four different types of grain from which to choose and three different types of grass from which to choose. He would like to determine if there is a particular combination of grain and grass that would lead to the greatest weight gain on average for his cows. He randomly selects three one-year-old cows and assigns them to each of the possible combinations of grain and grass. After one year he records the weight gain for each cow (in pounds) with the following results.

Cow Weight Gain (Pounds)			
	Grass A	**Grass B**	**Grass C**
	175	225	250
	160	215	240
Grain A	185	230	260

	190	245	275
	185	240	260
Grain B	195	255	285
	210	255	300
	200	245	310
Grain C	220	265	295
	225	275	350
	235	270	360
Grain D	220	280	345

a. Identify the dependent variable. Identify the two factors. In the associated computer output, which variable is Factor A and which variable is Factor B?

b. Consider the following graph of the average weight gain of the cows for each of the possible combinations of grass and grain. Does there appear to be any interaction between the grass and grain variables?



c. The results of the two-way ANOVA for the farmer's study are given in the following table. Perform a hypothesis test to determine if there is any interaction between the variables grass and grain at $\alpha = 0.05$. Does this agree with your observation in part **b.**?

ANOVA

Source of Variation	*SS*	*df*	*MS*
Sample	23097.2222	3	7699.0741
Columns	53272.2222	2	26636.1111
Interaction	3127.7778	6	521.2963
Within	1916.6667	24	79.8611
Total	81413.8889	35	

d. If there is no interaction found in part **c.**, is there sufficient evidence to conclude that there is a significant difference in the average weight gains among the cows for the four different types of grain? Use $\alpha = 0.05$.

e. Is there sufficient evidence to conclude that there is a significant difference in the average weight gains among the cows for the three different types of grass? Use $\alpha = 0.05$.

17. The partially completed analysis of variance table given below is taken from the article, "Power and Status, Exchange, Attribution, and Expectation States (Small Group Research)."[5] The experimenters investigated the effects of power and knowledge on one's emotional reaction in a study involving 52 students selected from a large private university. Each of the factors was run at two levels, with 13 subjects at each of the four different factor combinations.

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*
Power	1.2700	1		
Knowledge	0.2500	1		
Interaction		1		
Error	4.1400	48		
Total	5.6700	51		

a. Identify the dependent variable. Identify the two factors. In the associated computer output, which variable is Factor A and which variable is Factor B?

b. Complete the ANOVA table.

c. Can we conclude, with $\alpha = 0.10$, that there is interaction between power and knowledge?

d. With $\alpha = 0.05$, can we conclude that there is a significant difference in the two levels of power?

e. With $\alpha = 0.05$, can we conclude that there is a significant difference in the two levels of knowledge?

16.1 Exercises

Basic Concepts

1. Describe the shape of the chi-square distribution.
2. What is the sampling distribution of the sample variance?
3. What are the degrees of freedom associated with the chi-square distribution?
4. Can a chi-square statistic ever be negative? Explain why or why not.
5. Describe how the chi-square distribution changes in shape as n becomes large.
6. Explain the meaning of $\chi^2_{\alpha,\,df}$.
7. Explain the procedure for determining chi-square critical values.

Exercises

8. Find the chi-square critical value for each of the following.

   a. $\alpha = 0.01,\ df = 14$

   b. $\alpha = 0.01,\ df = 26$

   c. $\alpha = 0.05,\ df = 4$

   d. $\alpha = 0.05,\ df = 9$

   e. $\alpha = 0.005,\ df = 12$

9. Find the chi-square critical value for each of the following.

   a. $\alpha = 0.005,\ df = 40$

   b. $\alpha = 0.025,\ df = 15$

   c. $\alpha = 0.025,\ df = 2$

   d. $\alpha = 0.10,\ df = 24$

   e. $\alpha = 0.10,\ df = 50$

10. Suppose that a marketing manager is studying sales data for products that are not available in stores and only sold online. She collects the following weekly sales data for 10 products not sold in stores. Assume the population standard deviation for this data is \$5000.

Weekly Online Sales										
Product	1	2	3	4	5	6	7	8	9	10
Sales (\$)	26,259	18,514	21,579	18,739	27,821	22,511	29,753	20,235	16,258	15,990

   a. Compute the sample standard deviation for this data. Round your answer to the nearest dollar.

   b. Compute the value of χ^2. Round your answer to three decimal places.

   c. How many degrees of freedom are associated with this chi-square distribution?

   d. What is the value of $\chi^2_{0.05,\,df}$ for this data?

11. Michael is interested in obtaining 30-year fixed mortgage rates in Myrtle Beach, SC. He obtained rate quotes from 8 lenders, and the APR rates that were quoted to him are given in the following table.

30-Year Fixed Mortgage Rates	
Lender	**APR (%)**
EverBank	5.375
AimLoan	5.875
Great Western	6.125
Greenlight	6.375
Flagstar	5.375
AuroraBank	5.750
Quicken	5.875
Roundpoint	5.375

a. Calculate the variance of the sample. Round your answer to six decimal places.

b. Assuming the population standard deviation for the rates is 0.1%, calculate the value of χ^2.

c. Determine the value of $\chi^2_{0.025,\,df}$ for this data.

12. A fitness club manager suspects that the pool heater is faulty as the temperature of the heated pool does not seem to stay consistent throughout the day. The manager decides to collect data to determine if a new pool heater is needed. The temperature of the pool is supposed to stay within a variance of 2 if the heater is operating properly. The temperature data collected over a 14-day period is given in the following table.

Daily Pool Temperature														
Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Temperature	76.2	75.4	77.9	78.0	79.2	77.8	75.1	78.3	77.9	79.6	78.2	76.2	75.9	74.9

a. Calculate the variance of the sample. Round your answer to three decimal places.

b. Assuming the population variance for the pool temperature is 2, calculate the value of χ^2.

c. How many degrees of freedom are associated with this chi-square distribution?

13. The manufacturer of a high blood pressure medication must ensure that each tablet contains 37.5 mg of the active ingredient. It is also crucial that the standard deviation of the active ingredient per dose be less than 0.2 mg. A sample of ten tablets is taken and the data are shown in the following table.

Amount of Active Ingredient in a High Blood Pressure Tablet										
Sample	1	2	3	4	5	6	7	8	9	10
Active Ingredient Amount (mg)	37.3	36.9	37.0	37.1	37.4	37.0	36.8	37.2	37.5	37.4

a. Calculate the standard deviation of the sample. Round your answer to six decimal places.

b. Assuming the population standard deviation for the active ingredient amount is 0.2 mg, calculate the value of χ^2.

c. How many degrees of freedom are associated with this chi-square distribution?

Step 6: State the conclusion in terms of the original problem.

At the 0.10 level, there is insufficient evidence to conclude that the distribution of the card types in the set differs from the company claim that the proportion of each card type is approximately equal.

16.2 Exercises

Basic Concepts

1. Describe what the test statistic for the chi-square test for goodness of fit measures.
2. What is a multinomial probability distribution? What more familiar probability distribution discussed previously in the text is a multinomial probability distribution related to?
3. List the four requirements for a multinomial experiment.
4. What are the null and alternative hypotheses for a chi-square test for goodness of fit?
5. What is the test statistic for a chi-square test for goodness of fit?
6. How many degrees of freedom does the test statistic for the chi-square test for goodness of fit have?
7. What assumptions are necessary for a chi-square test for goodness of fit?
8. How are the expected values determined in a chi-square test for goodness of fit?

Exercises

9. An internet provider claims that the service calls they receive are equally distributed among the five working days of the week. A survey of 85 randomly selected service calls produced the following results.

Service Calls					
	Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**
Number of Calls	15	20	25	15	10

   a. Is the company's claim refuted by the data at $\alpha = 0.05$?

   b. What assumptions were made in the test for part **a.**?

10. Suppose a consumer affairs representative for Mars Incorporated claims that M&M's plain chocolate candies are mixed such that each large production batch has "precisely" the following ratios of colored candies: 30% brown, 20% yellow, 20% red, 10% orange, 10% green, and 10% blue. To test this claim, a professor distributed small sample bags of M&M's to students and had them count the number of candies of each color. The counts of the students were then pooled with the following results.

Candy Colors							
	Brown	Yellow	Red	Orange	Green	Blue	Total
Number of Candies	84	79	75	49	36	47	370

a. If the representative's claim is true, what would be the expected number of candies in each of the color categories for 370 candies?

b. Is the representative's claim refuted by the data at $\alpha = 0.01$?

c. What assumptions were made in performing the test for part **b.**?

11. A highway department executive claims that the number of fatal accidents which occur in her state does not vary from month to month. A survey of 170 fatal accidents produced the following results.

Accidents												
	Jan.	Feb.	Mar.	Apr.	May	Jun.	July	Aug.	Sept.	Oct.	Nov.	Dec.
Accidents	18	16	7	5	8	12	15	18	15	11	20	25

a. Is the executive's claim refuted by the data at $\alpha = 0.01$?

b. What assumptions were made in the test for part **a.**?

12. A psychologist conducted an attitude survey of 200 randomly selected individuals several years ago. The individuals were asked to pick the one category which most accurately described their attitudes. The results of the survey were as follows.

1st Attitude Survey	
Attitude	Percent of Respondents
Optimistic	15%
Slightly Optimistic	30%
Slightly Pessimistic	30%
Pessimistic	25%

The psychologist believes that these attitudes have changed over time. To test this theory, he randomly selects 200 individuals and asks them the same questions. The results of the second survey are as follows.

2nd Attitude Survey	
Attitude	Percent of Respondents
Optimistic	20%
Slightly Optimistic	40%
Slightly Pessimistic	30%
Pessimistic	10%

a. Can the psychologist conclude that the attitudes have changed over time at $\alpha = 0.01$?

b. What assumptions were made in the test for part **a.**?

13. A manager for an insurance company believes that customers have the following preferences for life insurance products: 20% prefer Whole Life, 30% prefer Universal Life, and 50% prefer Life Annuities. The results of a survey of 240 customers were tabulated.

Life Insurance Preferences	
Product	**Number**
Whole Life	60
Universal Life	72
Life Annuities	108

a. State the null and alternative hypothesis for testing the insurance manager's claim.

b. Calculate the expected number of customers who prefer Whole Life policies.

c. Calculate the test statistic. (**Note:** keep intermediate calculations to six decimal places.)

d. Can we refute the insurance manager's claim for customers preferring each insurance product at a significance level of 0.05?

16.3 The Chi-Square Test for Association

Our interest sometimes extends beyond one variable to summarizing the relationship between two qualitative variables. For example, a radio executive might be interested in knowing if the marital status of an individual affects that person's preference for music. Here the qualitative variables of interest are marital status, which could take on values such as single, married, divorced, or widowed, and preference for music, which could take on the values Classical, Jazz, Easy Listening, Rock, Rap, and Country. Other examples of relationships between two qualitative variables which might be of interest are age and political preference, education level and job performance, income level and occupation, etc. When we are interested in this type of relationship, we often make use of a contingency table.

Contingency Table (or Two-Way Frequency Table)

A **contingency table** organizes data on two characteristics simultaneously. Each cell in a contingency table contains either a count or a proportion which represents the number of observations falling into that cell. It is important to note that contingency tables are composed of two variables, each satisfying the properties of the multinomial distribution.

DEFINITION

Table 16.3.1 shows the general form of a contingency table.

Table 16.3.1 - General Form of a Contingency Table

		Factor A		
		Level A1	**Level A2**	**Total**
Factor B	**Level B1**	n_1	n_2	$n_1 + n_2$
	Level B2	n_3	n_4	$n_3 + n_4$
Total		$n_1 + n_3$	$n_2 + n_4$	$n = n_1 + n_2 + n_3 + n_4$

Step 6: State the conclusion in terms of the original problem.

There is evidence at the 0.10 level to conclude that support for raising the national minimum wage and political affiliation are dependent. The difference in support is too great to believe it to be attributed to ordinary sampling variation alone.

16.3 Exercises

Basic Concepts

1. Explain the difference between the chi-square test for goodness of fit and the chi-square test for association.
2. What is a contingency table?
3. Describe the information that each cell in a contingency table provides.
4. What properties must the two categories of the contingency table possess?
5. What level(s) of measurement may the categories of a contingency table have?
6. Consider the variable income. Describe how this variable could be transformed to be included in a contingency table. Is information lost during the transformation?
7. Explain why a test for association is not valid if single data points are allowed to belong to more than one category.
8. Restate the multiplication rule for independent events. Explain how this rule pertains to the chi-square test for association.
9. State the null and alternative hypotheses for a chi-square test for association between two qualitative variables.

Exercises

10. A political analyst is interested in studying the relationship between age and political affiliation. The analyst randomly selects 200 people and determines their age and political affiliation. The number of responses in each of the categories is as follows.

Age and Political Affiliation

	Political Affiliation		
Age	**Democrat**	**Republican**	**Independent**
18–34	50	10	15
35–51	15	25	15
52–68	25	35	10

a. Can the analyst conclude that age and political affiliation are dependent at $\alpha = 0.05$?

b. What assumptions were made in the test for part **a.**?

11. A sociologist is interested in studying the relationship between education and crime. She randomly selects 450 people and asks their education level and whether or not they have ever been convicted of a felony. The following table displays the number of respondents in each category.

Education and Crime		
Have you ever been convicted of a felony?		
	Response	
Education Level	Yes	No
Less Than 9 Years	6	105
9 Years to 12 Years	12	93
12 Years to 16 Years	3	93
16+ Years	12	126

a. Can the sociologist conclude that education level and crime are dependent at $\alpha = 0.10$?

b. What assumptions were made in the test for part **a.**?

12. A psychologist is preparing a thesis on child abuse. He thinks that there may be a relationship between various types of child abuse and the marital status of the parents of the child. To study this, he randomly selects the records of 197 abused children and determines the marital status of the parents and the documented type of child abuse. The results of the study are as follows.

Child Abuse		
	Marital Status	
Type of Abuse	Married	Not Married
Neglect	50	50
Physical	20	30
Sexual	10	19
Emotional	10	8

a. Can the psychologist conclude that the type of child abuse and marital status of the child's parents are dependent at $\alpha = 0.05$?

b. What assumptions were made in the test for part **a.**?

13. The National Fire Protection Association is interested in studying the relationship between the causes of fires and the region of the country in which the fires occur. They randomly select 500 fires and determine the region of the country in which the fire occurred and cause of the fire with the following results.

Fires				
	Region			
Cause of Fire	North	South	East	West
Smoking	37	38	45	35
Heating Equipment	25	20	12	19
Arson	17	15	23	15
Electrical	12	13	25	13
Children at Play	10	11	8	11
Other	27	28	14	27

a. Can the association conclude that the cause of the fire and the region of the fire are dependent at $\alpha = 0.01$?

b. What assumptions were made in the test for part **a.**?

14. A market researcher believes that brand perception of one of the company's products may vary between different age groups. After interviewing 291 persons, the following data was compiled.

Market Research

	Brand Perception			
Age	**Favorable**	**Unfavorable**	**Neutral**	**Total**
18–29	64	20	23	107
30–45	49	14	18	81
Over 45	61	20	22	103
Total	174	54	63	291

a. Calculate the expected number of people from the 18-29 age group who respond favorably to the brand.

b. Calculate the expected number of people from the 30-45 age group who respond unfavorably to the brand.

c. Calculate the test statistic.

d. Can we conclude that brand perception is dependent on age at a significance level of 0.05?

15. An educational researcher wishes to know if there is a difference in academic performance for college freshmen that live on campus and those that commute. Data was collected from 232 students.

Educational Research

	Academic Performance			
Location	**Average**	**Below Average**	**Above Average**	**Total**
On campus	89	39	42	170
Off campus	31	16	15	62
Total	120	55	57	232

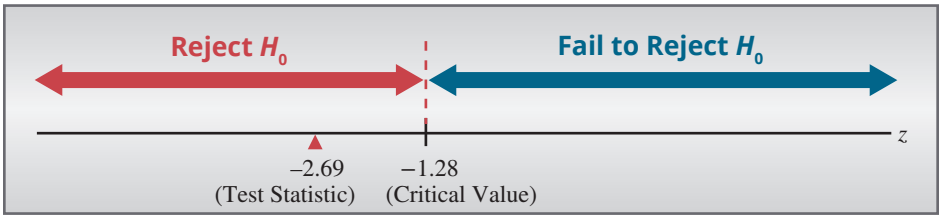
a. Calculate the expected number of college freshmen who live on campus and perform above average.

b. Calculate the expected number of college freshmen who live off campus and perform above average.

c. Calculate the test statistic.

d. Can we conclude that freshman housing location and academic performance are related at a significance level of 0.01?

Step 5: Choose between the null and alternative hypotheses.



As shown, the value of the test statistic falls in the rejection region (−2.69 is less than −1.28). It is unlikely that the difference between the observed value and the hypothesized value is due to ordinary sampling variation. Thus, we reject the null hypothesis at $\alpha = 0.10$.

Step 6: State the conclusion in terms of the original problem.

There is sufficient evidence for the Chamber of Commerce to conclude at $\alpha = 0.10$ that the median home price in Durham, North Carolina is significantly lower than the median home price in the United States of $428,700.

17.1 Exercises

Basic Concepts

1. What are parametric statistics?
2. Identify and explain the main disadvantage of the Sign Test.
3. Under what conditions are parametric statistical methods not appropriate for data analysis?
4. Identify the three characteristics of nonparametric statistical methods.
5. What are the disadvantages of nonparametric statistics?
6. The Sign Test and the Wilcoxon Signed-Rank Test are designed to conduct hypothesis tests involving which kind(s) of experiments? What is the corresponding parametric statistical technique used to analyze these types of experiments?
7. What assumptions are made when conducting the Sign Test?
8. Name the two ways that the Sign Test can be used to perform hypothesis tests.
9. How do the rejection regions for nonparametric tests differ from those for parametric tests? Explain.
10. What is done with measurements that have a difference of zero in a paired difference experiment? Why is this the case?
11. What are the null and alternative hypotheses associated with the Sign Test?
12. What is the test statistic for the Sign Test for small samples? How small is a *small* sample?
13. What is the test statistic for the Sign Test for large samples? How large is a *large* sample?
14. Identify the critical values and rejection rules for both small and large samples with regard to the Sign Test.

Exercises

15. Hurricane Hugo swept through the Lowcountry in South Carolina causing billions of dollars of damage. In the past, the median claim for homes damaged by hurricanes for an insurance company in the Lowcountry had been $25,000. The insurance company believes that the median claim will be significantly larger for homes damaged by Hugo than past hurricanes. In order to investigate this theory, the insurance company randomly selects 55 homes and sends adjusters to settle the claims. In the sample of 55 homes, 40 of the homes had a claim in excess of the historical median. Is there overwhelming evidence at $\alpha = 0.10$ that the median claim for home damage from Hurricane Hugo was greater than the historical median?

16. The manufacturer of Brand X floor polish is developing a new polish that they hope will dry faster than the competition's polish. The competition's polish is advertised to have an average (median) drying time of 10 minutes. In a random sample of 1000 polishes with the new polish, 700 of the polishes dried in less than 10 minutes. Based on the data, can the manufacturer conclude that the median drying time for Brand X is faster than the competition's brand at a 0.05 level of significance?

17. NarStor, a computer disk drive manufacturer, claims that the median time until failure for their hard drives is 14,400 hours. You work for a consumer group that has decided to examine this claim. Technicians ran 16 NarStor hard drives continuously for almost three years. Recently the last drive failed. The times to failure (in hours) are given in the following table.

Time Until Hard Drive Failure (Hours)							
330	620	1870	2410	4620	6396	7822	8102
8309	12,882	14,419	16,092	18,384	20,916	23,812	25,814

a. Is there overwhelming evidence that the median time until failure is less than the manufacturer claims? Use $\alpha = 0.05$.

b. What assumption did you make in performing the test in part **a.**?

18. A.C. Bone has developed a duck hunting boot which it claims can remain immersed for more than 12 hours without leaking. 15 of the boots are tested and the time until first leakage is measured. Nine of the boots last more than 12 hours without leaking.

a. Does the data substantiate A.C. Bone's claim at $\alpha = 0.05$?

b. What assumption did you make in performing the test in part **a.**?

19. Given that most textbooks can now be purchased online, one wonders if students can save money by comparison shopping for textbooks at online retailers and at their local bookstores. To investigate, students at Tech University randomly sampled 25 textbooks on the shelves of their local bookstores. The students then found the "best" available price for the same textbooks via online retailers. The prices for the textbooks are listed in the following table.

Textbook Prices					
	Price ($)			Price ($)	
Textbook	Bookstore	Online Retailer	Textbook	Bookstore	Online Retailer
1	70	60	14	85	75
2	38	36	15	100	85
3	88	89	16	68	62
4	165	149	17	67	69
5	80	136	18	140	142
6	103	95	19	49	40
7	42	50	20	149	127
8	98	111	21	126	130
9	89	65	22	92	93
10	97	86	23	144	129
11	140	130	24	98	84
12	40	30	25	40	52
13	175	150			

Using the data in the table, and without making any distributional assumptions, is it less expensive for the students to purchase textbooks from the online retailers than the local bookstores? Use $\alpha = 0.01$.

20. The management for a large grocery store chain would like to determine if a new scanner will enable cashiers to process a larger number of items on average than the scanner which they are currently using. Seven cashiers are randomly selected, and the number of grocery items which they can process in three minutes is measured for both the old scanner and the new scanner. The results of the test are as follows.

Number of Grocery Items Processed in Three Minutes							
Cashier	1	2	3	4	5	6	7
Old scanner	60	70	55	75	62	52	58
New scanner	65	71	55	75	65	57	57

Without making any assumptions about the distribution, can management conclude that the new scanner will allow cashiers to process a significantly larger number of items on average than the old scanner at $\alpha = 0.05$?

21. An auto dealer is marketing two different models of a high-end sedan. Since customers are particularly interested in the safety features of the sedans, the dealer would like to determine if there is a difference in the braking distance (the number of feet required to go from 60 mph to 0 mph) of the two sedans. Six drivers are randomly selected and asked to participate in a test to measure the braking distance for both models. Each driver is asked to drive both models and brake once they have reached exactly 60 mph. The distance required to come to a complete halt is then measured in feet. The results of the test are as follows.

Braking Distance of High-End Sedans (in Feet)						
Driver	1	2	3	4	5	6
Model A	150	145	160	155	152	153
Model B	152	146	160	157	154	155

Without making assumptions about the distribution of the data, can the auto dealer conclude that there is a significant difference in the braking distance of the two models of high-end sedans? Use $\alpha = 0.10$.

22. A nutritionist is interested in determining the decrease in cholesterol level which a person can achieve by following a particular diet which is low in fat and high in fiber. Seven subjects are randomly selected to try the diet for six months, and their cholesterol levels are measured both before and after the diet. The results of the study are as follows.

Cholesterol Levels							
Subject	1	2	3	4	5	6	7
Before Diet	155	170	145	200	162	180	160
After Diet	152	168	148	195	162	178	157

Can the nutritionist conclude that there is a significant decrease in average cholesterol level when the diet is used? We don't have any knowledge about the distribution of the data. Use $\alpha = 0.01$.

17.2 The Wilcoxon Signed-Rank Test

A disadvantage of the Sign Test, discussed in the previous section, is that it wastes information. The Sign Test merely counts the number of positive or negative signs in a paired difference experiment and ignores the magnitude of the differences. The **Wilcoxon Signed-Rank Test** is a nonparametric technique which can be used to evaluate a paired difference experiment where the magnitude of these differences is not ignored. However, the magnitude is not taken directly into account; instead, the ranks of the data are analyzed. This test is designed to detect populations whose centers are shifted to the right or the left of each other. As with the Sign Test, no distributional assumption is required. However, the pairs of data must have been selected in a random fashion, and it must be possible to rank the differences.

Ranking Procedure

First, take the differences of the ordered pairs. Then take the absolute value of the differences. Omit all pairs with an absolute difference of zero. So the value of n (the number of ordered pairs), will need to be reduced by the number of absolute differences that equal 0. If several pairs have absolute differences that are equal to each other (there are ties), assign to each of these the average of the ranks that would have been assigned. So, if two pairs had the same difference and they are tied for ranks 4 and 5, then we will average the ranks and give each of these pairs the average rank of 4.5 If three pairs are

Test Procedure for the Wilcoxon Signed-Rank Test

Assumptions:

1. Pairs of data have been selected in a random fashion.
2. Data are quantitative.

Hypotheses:

H_0: The probability distributions of the two populations of interest are the same.

H_a: $>$ One-Tailed: Population X is to the *right* of Population Y (Diff = $Y - X$).

$\neq$ Two-Tailed: Population X is to the *right* of or to the *left* of Population Y.

$<$ One-Tailed: Population X is to the *left* of Population Y (Diff = $Y - X$).

Test Statistic:

If $n \leq 25$, and

H_a: $>$, then $T = T_+$ = the sum of the ranks of the positive differences.

H_a: $\neq$, then $T = \text{Min}\,(T_+, T_-)$.

H_a: $<$, then $T = T_-$ = the sum of the ranks of the negative differences.

If $n > 25$, $T = \text{Min}\,(T_+, T_-)$, and the test statistic is given by

$$z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}.$$

Critical Value(s):

If $n \leq 25$, reject H_0 if $T \leq T_c$, the critical value in Table J.

If $n > 25$, and

H_a: $>$ or H_a: $<$, reject H_0 if $z \leq -z_\alpha$.

H_a: $\neq$, reject H_0 if $z \leq -z_{\alpha/2}$.

PROCEDURE

17.2 Exercises

Basic Concepts

1. What assumptions are required for the Wilcoxon Signed-Rank Test?

2. The Wilcoxon Signed-Rank Test is primarily used to perform hypothesis tests about what type of experiment?

3. What are the advantages and disadvantages of the Wilcoxon Signed-Rank Test?

4. Describe the procedure for assigning ranks to data in order to perform a Wilcoxon Signed-Rank Test. What is to be done when two values are the same?

5. Describe how to calculate the rank sums for a paired difference experiment in order to perform a Wilcoxon Signed-Rank Test.
6. If the sample size is less than or equal to 25, identify the three possible test statistics used for the Wilcoxon Signed-Rank Test. How do you choose which statistic to use?
7. What are the null and alternative hypotheses associated with the Wilcoxon Signed-Rank Test?
8. Explain why the population distributions are important when performing a Wilcoxon Signed-Rank Test.
9. What is the test statistic for the Wilcoxon Signed-Rank Test if the sample size is large? How large is *large* with regard to sample size?
10. Identify the critical values and rejection regions for both large and small samples with regard to the Wilcoxon Signed-Rank Test.

Exercises

11. Rank the following emerging markets mutual funds from lowest to highest price using the methodology presented for the Wilcoxon Signed-Rank Test.

Emerging Markets Mutual Funds

Mutual Fund	Price ($)
American Funds	49.30
Columbia Management	9.41
Morgan Stanley	88.50
Fidelity Investments	24.40
John Hancock	9.41
DWS Investments	15.57
UBS	12.15
Prudential Investments	9.23
Value Line Funds	32.82
The Vanguard Group	34.72

12. Rank the following consumer price indexes (CPI) for selected groups of goods and services in September 2011 using the methodology presented for the Signed-Rank Test. The data in the table represent the unadjusted percent change in price level from September 2010 to September 2011.[1]

Percent Change in CPI

Expenditure Category	CPI (% Change 9/10 to 9/11)
Food	4.7
Alcoholic Beverages	1.4
Housing	1.8
Apparel	3.5
Public Transportation	7.4
Medical Care	2.8

Percent Change in CPI	
Expenditure Category	**CPI (% Change 9/10 to 9/11)**
Education	4.4
Tobacco and Smoking Products	2.4
Gasoline	33.3
New and Used Motor Vehicles	3.6

13. A study conducted by the Orentreich Foundation found that women who practiced transcendental meditation (T.M.) for 20 minutes a day had high levels of DHEA-S, a hormone that may help prevent breast cancer and osteoporosis. Suppose eight women are randomly selected to participate in a study. The DHEA-S levels of the participants are measured prior to practicing transcendental meditation and then measured one year after practicing transcendental meditation for 20 minutes a day. The following table is a summary of the results of the study.

Study Results		
Study Participant	**DHEA-S Level Before T.M. (mg)**	**DHEA-S Level After T.M. (mg)**
A	20	25
B	25	25
C	18	20
D	27	26
E	19	20
F	24	26
G	20	21
H	30	29

a. Using the Sign Test, does the data indicate that the DHEA-S level of women increases after practicing transcendental meditation for 20 minutes per day for one year at $\alpha = 0.05$?

b. What assumptions were necessary to perform the Sign Test?

c. Using the Signed-Rank Test, does the data indicate that the DHEA-S level of women increases after practicing transcendental meditation for 20 minutes per day for one year at $\alpha = 0.05$?

d. What assumptions were necessary to perform the Signed-Rank Test?

e. Which test do you think produces more accurate results? Why?

14. The management for a large grocery store chain would like to determine if a new scanner will enable cashiers to process a larger number of items on average than the scanner which they are currently using. Seven cashiers are randomly selected, and the number of grocery items which they can process in three minutes is measured for both the old scanner and the new scanner. The results of the test are as follows.

Number of Grocery Items Processed in Three Minutes							
Cashier	1	2	3	4	5	6	7
Old scanner	60	70	55	75	62	52	58
New scanner	65	71	55	75	65	57	57

a. What assumption must be made in order to perform the test of hypothesis using the paired difference t-test?

b. Using the Signed-Rank Test, does the data provide conclusive evidence that the new scanner enables cashiers to process a significantly larger number of items than the old scanner at $\alpha = 0.05$?

c. What assumptions were made in performing the Signed-Rank Test?

d. How do the results of the Signed-Rank Test compare with the paired difference t-test performed in Section 12.3, Exercise 9?

15. An auto dealer is marketing two different models of a high-end sedan. Since customers are particularly interested in the safety features of the sedans, the dealer would like to determine if there is a difference in the braking distance (the number of feet required to go from 60 mph to 0 mph) of the two sedans. Six drivers are randomly selected and asked to participate in a test to measure the braking distance for both models. Each driver is asked to drive both models and brake once they have reached exactly 60 mph. The distance required to come to a complete halt is then measured in feet. The results of the test are as follows.

Braking Distance of High-End Sedans (in Feet)						
Driver	1	2	3	4	5	6
Model A	150	145	160	155	152	153
Model B	152	146	160	157	154	155

a. What assumption must be made in order to perform a test of hypothesis using the paired difference t-test?

b. Using the Signed-Rank Test, does the data provide conclusive evidence that there is a significant difference in the median braking distance of the two sedans at $\alpha = 0.10$?

c. What assumptions were made in performing the Signed-Rank Test?

d. How do the results of the Sign Test performed in Section 17.1, Exercise 21 and the signed-rank test performed in part **b.** compare with the paired difference t-test performed in Section 12.3, Exercise 10?

17.3 The Wilcoxon Rank-Sum Test

We discussed nonparametric procedures for testing claims about a paired difference experiment in the previous two sections. In this section we will discuss a nonparametric procedure for hypothesis tests in which an independent experimental design is used to compare two population medians.

17.3 Exercises

Basic Concepts

1. What type of data is the Wilcoxon Rank-Sum Test used to analyze?
2. What is the parametric test used to analyze the type of data that can also be analyzed by the Wilcoxon Rank-Sum Test? What assumptions are associated with this test, and why are they sometimes not reasonable?
3. What assumptions are required for the Wilcoxon Rank-Sum Test?
4. What levels of measurement may data possess in order for the Wilcoxon Rank-Sum Test to be performed?
5. Describe the procedure for ranking data in order to perform a Wilcoxon Rank-Sum Test.
6. What are the null and alternative hypotheses associated with the Wilcoxon Rank-Sum Test?
7. What is the test statistic associated with the Wilcoxon Rank-Sum Test for small samples? What does the test statistic depend on and how small is a *small* sample?
8. What is the test statistic associated with the Wilcoxon Rank-Sum Test for large samples?
9. Identify the critical values associated with the Wilcoxon Rank-Sum Test for both large and small samples.

Exercises

10. A luxury car dealer is considering two possible locations for a new auto mall. The rent on the south side of town is cheaper. However, the dealer believes that the average household income is significantly higher on the north side of town. The dealer has decided that he will locate the new auto mall on the north side of town if the results of a study which he has commissioned show that the median household income is significantly higher on the north side of town. The results of the study are as follows.

Household Incomes	
North Side (in thousands of dollars)	South Side (in thousands of dollars)
100	93
95	95
105	92
75	100
125	86
85	98
115	88
105	93
95	93

a. Use the Wilcoxon Rank-Sum Test to determine if the auto dealer should locate the new auto mall on the north side of town. Use $\alpha = 0.05$.

b. What assumptions were made in performing the hypothesis test in part **a.**?

11. An internal auditor for Tiger Enterprises has been asked to determine if there is a difference in the amount charged for daily expenses by two top salespersons, Mrs. Ellis and Mr. Ford. The auditor randomly selects seven days and determines the daily expenses for each salesperson, excluding hotel cost.

Daily Expenses							
Mrs. Ellis ($)	85	83	88	84	86	85	85
Mr. Ford ($)	90	85	95	80	100	85	95

a. Using the Wilcoxon Rank-Sum Test, can the auditor conclude that there is a difference in the median amount charged for daily expenses by the two top salespersons, Mrs. Ellis and Mr. Ford? Use $\alpha = 0.05$.

b. What assumptions were made in performing the test in part **a.**?

12. The Armed Forces have two different programs for training aircraft personnel. A government regulatory agency has been commissioned to evaluate any differences which may exist between the two programs. The agency administers a standardized test to randomly selected groups of students from the two programs. The results of the test for the students in each of the programs are as follows.

Standardized Test Scores							
Program A	85	95	75	100	70	90	80
Program B	87	96	78	100	74	92	82

a. Using the Wilcoxon Rank-Sum Test, can the agency conclude that there is a difference in the median test scores of students in the two programs? Use $\alpha = 0.10$.

b. What assumptions were made in performing the test in part **a.**?

13. A supply clerk with the Navy has been asked to determine if a new battery which has been offered to the Navy (at a reduced price) has a shorter life than the battery which they are currently using. He randomly selects batteries of each type and allows them to run continuously so that he can measure the time until failure for each battery. The results of the test are as follows.

Time Until Failure for Batteries (Hours)						
New Battery	655	730	670	715	685	745
Old Battery	745	675	730	690	760	660

a. Using the Wilcoxon Rank-Sum Test, does the data suggest at $\alpha = 0.05$ that the median time until failure for the new battery is significantly less than the median time until failure for the old battery?

b. What assumptions were made in performing the test in part **a.**?

14. A cereal manufacturer has advertised that its product, Fiber Oat Flakes, has a lower fat content than its competitor, Bran Flakes Plus. Because of the complaints from the manufacturer of Bran Flakes Plus, the FDA has decided to test the claim that Fiber Oat Flakes has a lower median fat content than Bran Flakes Plus. Several boxes of each cereal are selected and the fat content per serving is measured. The results of the study are as follows.

Fat Content of Cereals (Grams)									
Fiber Oat Flakes	5	6	4	7	3	5	5	6	4
Bran Flakes Plus	6	8	4	9	3	7	5	8	4

a. Using the Wilcoxon Rank-Sum Test, does the study performed by the FDA substantiate the claim made by the manufacturer of Fiber Oat Flakes at $\alpha = 0.05$?

b. What assumptions were made in performing the test in part **a.**?

15. A Hollywood studio believes that a movie which is considered a drama will draw a larger crowd on average than a movie which is a comedy. To test this theory, the studio randomly selects several movies which are classified as dramas and several movies which are classified as comedies and determines the box office revenue for each movie. The results of the survey are as follows.

Box Office Revenues (Millions of Dollars)					
Drama	279	206	243	181	277
Comedy	216	292	439	299	301

a. Using the Wilcoxon Rank-Sum Test, does the data substantiate the studio's belief that dramas will draw a larger crowd on average than comedies at $\alpha = 0.05$?

b. What assumptions were made in performing the test in part **a.**?

16. *Consumer Magazine* is reviewing the top selling amplifiers produced by two major stereo manufacturers. One of the most important qualities of the amplifiers is the maximum power output. Brand A has redone their internal design and claims to have a higher maximum power level than Brand B. To test this claim, *Consumer Magazine* randomly selects amplifiers from each brand and determines the maximum power output. The results of the test are as follows.

Maximum Power Output (Watts)							
Brand A	800	828	772	830	770	826	774
Brand B	780	805	755	807	753	803	757

a. Using the Wilcoxon Rank-Sum Test, does the data substantiate the claim that the Brand A amplifier has a higher median maximum power output than Brand B at $\alpha = 0.05$?

b. What assumptions were made in performing the test in part **a.**?

17. A state environmental board wants to compare pollution levels in two of its major cities. Sunshine City thrives on the tourist industry and Service City thrives on the service industry. The environmental board randomly selects several areas within the cities and measures the pollution levels in parts per million with the following results.

Pollution Levels (ppm)								
Sunshine City	8.50	9.00	8.00	9.07	7.93	9.14	7.86	8.50
Service City	7.90	8.35	7.45	8.40	7.40	8.45	7.35	7.90

a. Using the Wilcoxon Rank-Sum Test, can the state environmental board conclude at $\alpha = 0.05$ that Service City has a lower pollution level on average than Sunshine City?

b. What assumptions were made in performing the test in part **a.**?

17.4 Exercises

Basic Concepts

1. What is the correlation coefficient? How is this different from the Spearman rank correlation coefficient?
2. What is the formula for calculating Spearman's rho?
3. Can you calculate Spearman's rho if there are ties in the rank data?
4. Identify the difference in notation between Spearman's rho for population and sample data.
5. Explain the similarities in the behavior of the parametric correlation coefficient and Spearman's rho.
6. Identify one main advantage of the Spearman's rank correlation coefficient versus the parametric correlation coefficient.
7. Explain the procedure for ranking data when calculating Spearman's rho.
8. What are the null and alternative hypotheses for the rank correlation test?
9. Consider the value $r_s = 0.12$. Interpret this value in terms of the x and y variables used to calculate Spearman's rho.

Exercises

10. Chris is a new cashier assigned to a scanner in a supermarket. Each day a sample of purchases at that scanner is examined and a percent of pricing errors is recorded along with the total number of customers who used that scanner. Does the following data indicate an association between Chris' performance and how busy his scanner was? Use $\alpha = 0.05$.

% Pricing Errors and Total Customers			
Number of Customers	Errors (%)	Number of Customers	Errors (%)
57	4.2	67	2.5
44	5.5	71	2.9
32	5.7	69	2.6
60	3.9	56	1.0
55	3.2	51	2.0
59	4.1	70	1.7
63	3.3		

11. Twelve new runners were randomly assigned to different training programs, where they were required to run a certain number of miles every week for a year prior to a major race. After the training, the participants ran the race and their finishing times were recorded.

Miles of Training and Race Times			
Miles Logged	Race Time (Minutes)	Miles Logged	Race Time (Minutes)
35	198	30	189

Miles of Training and Race Times			
Miles Logged	Race Time (Minutes)	Miles Logged	Race Time (Minutes)
25	165	29	240
45	155	42	224
60	148	24	201
70	135	19	246
21	243	55	166

a. With 95% confidence, is there evidence that the number of miles logged in a week during training affects the runner's race time?

b. Can the linear correlation coefficient, r, be calculated in order to fit a least squares regression line to the data in the table in an effort to predict the finish time of runners based on the number of miles logged during training? Why or why not?

12. The following data consists of college rankings of five universities by two different magazines. Is there a correlation between the rankings of the magazines? Use $\alpha = 0.10$.

College Rankings by Magazines					
College	A	B	C	D	E
Magazine 1	1	4	2	3	5
Magazine 2	4	3	1	5	2

13. An anthropologist records the heights (in inches) of ten fathers and their sons. Does the following data support (at the 5% level) that taller fathers tend to have taller sons?

Heights of Fathers and Sons (Inches)			
Son's Height	Father's Height	Son's Height	Father's Height
72	70	65	71
68	73	70	78
74	72	69	67
66	68	67	65
71	69	80	66

14. After a mother-daughter golf tournament, mothers and daughters were ranked among themselves. Does the following data show (at the 5% level) a correlation between the daughters' and mothers' golf skills?

Golf Rankings			
Daughter's Ranking	Mother's Ranking	Daughter's Ranking	Mother's Ranking
1	5	5	3
9	4	3	6
10	8	7	7
2	2	6	10
4	1	8	9

Example 17.5.3

Detecting Randomness of a Set of Numbers

Is the following data random?

16, 25, 52, 11, 38, 47, 12, 98, 4

Solution

How do you test randomness with a numerical data set? Create a new data set comparing each value to the median value. To do this, substitute each value in the original data set with an A if it is above the median value, a B if it is below the median value, and eliminate any values that equal the median.

H_0: The data is random.

H_a: The data is not random.

Median = 25

16, 25, 52, 11, 38, 47, 12, 98, 4

B, Ø, A, B, A, A, B, A, B

m (the number of A's) = 4

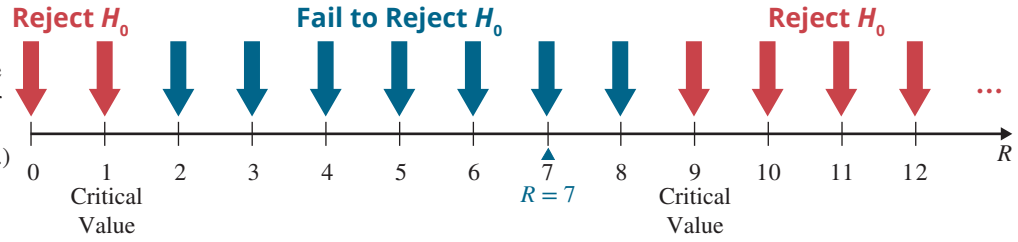
n (the number of B's) = 4

R (the number of runs) = 7

critical values = 1, 9

Fail to reject H_0; there is no evidence of nonrandomness.

(Reject H_0 for values less than or equal to the smaller critical value or greater than or equal to the larger critical value.)



17.5 Exercises

Basic Concepts

1. Describe in your own words what is being tested with the runs test.
2. Consider the following sequence of 10 coin tosses.

   H, H, T, T, H, H, H, T, T, H

   Without performing any kind of test, do you believe this sequence is random? Explain why or why not.
3. What are the null and alternative hypotheses associated with the runs test?
4. What parameters need to be calculated in order to perform a runs test?
5. What is the rejection rule for a small sample runs test? How small is a *small* sample?
6. What is the rejection rule for a large sample runs test? How large is a *large* sample?
7. If a numerical set of data is under consideration, which parameter are the data points compared to in order to perform the runs test?

Exercises

8. In the state of Tennessee, the number of deaths due to traffic accidents from 2010 to 2020 are shown in the following table. Use the runs test to examine non-randomness at the 0.05 level.

Traffic Fatalities in Tennessee, 2010-2020	
Year	**Number of Traffic Fatalities**
2010	1032
2011	937
2012	1014
2013	995
2014	963
2015	962
2016	1037
2017	1024
2018	1040
2019	1148
2020	1221

9. A sociologist designs a study that involves a procedure of selecting individuals randomly from an email list and then contacting them to determine if they own or rent their residence. The results are recorded in the order of phone calls (O = Own, R = Rent).

O O O R R O R O R R O R R R R O R R R O O R R R O R

Does the sociologist have a random sequence of residential data at the 0.05 level?

10. A car tire manufacturer keeps track of the tires produced by one of the production lines. They observe the following sequence (D for defective items and N for non-defective items).

D D D N N D N D N D D D

Test the quality control manager's claim that there is no pattern in producing defective tires at the 0.05 level.

11. A marathon runner tries to run every day except when it is raining during the month of July. He observes the rainy (R) days and sunny (S) days to be able to predict the weather as follows.

S S S R R S S S R R R R S R S R R S S R S R S R R S R S R S S

Are the rainy days randomly scattered in the month of July at the 0.05 level?

Braking Distances (in feet) with Ranks						
	Brake Pad A	**Rank**	**Brake Pad B**	**Rank**	**Brake Pad C**	**Rank**
	94	2	103	9	96	4
	95	3	127	14	126	13
Rank Sum		**21**		**55**		**44**

The hypotheses for this test can be written as follows.

H_0: The braking distances for the three pads are the same.

H_a: At least one of the braking distances is different.

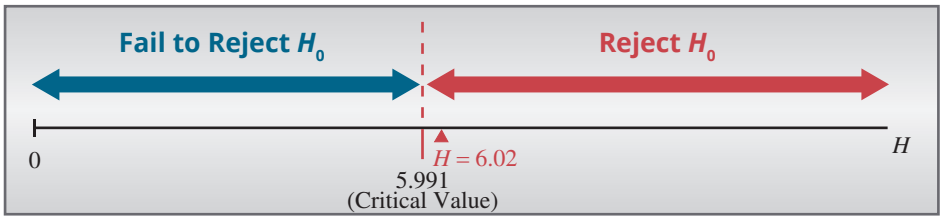
The ranks of the observations are given in the table. The sum of the ranks for each type of brake pad is found in the last row of the table.

$$R_1 = 21 \quad R_2 = 55 \quad R_3 = 44$$

We can now calculate the test statistic,

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1).$$

We find that $H = \frac{12}{15(15+1)}\left(\frac{21^2}{5} + \frac{55^2}{5} + \frac{44^2}{5}\right) - 3(15+1) = 6.02.$



Referring to Appendix A, Table G, we see that $\chi^2_{0.05,2} = 5.991$. Because the test statistic exceeds the critical value ($6.02 > 5.991$) we reject the null hypothesis and conclude that at least one of the braking distances is sufficiently different.

17.6 Exercises

Basic Concepts

1. Which parametric test corresponds to the nonparametric Kruskal-Wallis Test?
2. What are the null and alternative hypotheses associated with the Kruskal-Wallis Test?
3. What are the assumptions associated with the Kruskal-Wallis Test?
4. How is the Kruskal-Wallis test similar to the Wilcoxon Rank-Sum Test?
5. What is the test statistic for the Kruskal-Wallis Test? How is it calculated?
6. What is the rejection rule for the Kruskal-Wallis Test?
7. How many populations can be compared using the Kruskal-Wallis Test?

Exercises

8. An Internet service provider is considering four different servers for purchase. Potentially, the company would be purchasing hundreds of these servers, so it wants to make sure it is making the best decision. Initially, five of each type of server are borrowed, and each is randomly assigned to one of the 20 technicians (all technicians are similar in skill). Each server is then put through a series of tasks and rated using a standardized test. The higher the score on the test, the better the performance of the server. The data is as follows.

Server Test Scores			
Server 1	**Server 2**	**Server 3**	**Server 4**
48.5	56.4	52.1	64.3
46.5	68.2	56.3	68.3
52.4	68.5	48.3	72.2
54.1	64.2	52.2	70.6
58.9	60.1	54.8	56.5

Perform a Kruskal-Wallis Test on this data using $\alpha = 0.10$. Are there differences between the servers?

9. The following summary is obtained from an experiment where groups of cows were fed according to one of the four different feeding schedules, and their milk productions were recorded. The data given shows the daily milk production in gallons for each cow. Test at $\alpha = 0.10$ to examine whether or not the milk production for all four schedules is the same.

Milk Production by Schedule (Gallons)					
Schedule 1	11.5	12.7	12.9	10.1	10.5
Schedule 2	9.1	10.7	9.5	10.9	10.4
Schedule 3	12.4	11.9	10.0	11.4	12.1
Schedule 4	12.8	12.6	11.7	11.3	10.9

10. The following data set contains the reading speed (in words per minute) of second grade students.

Reading Speeds (wpm)		
Public School	**Private School**	**Home School**
54	66	65
67	55	64
63	62	60
105	69	72
61	71	68

Is there sufficient evidence at the 0.01 level of significance to conclude that the reading speeds vary by school type?