Where Did the Term *Statistics* Come From?

Latin: The evolution of the word statistics comes from the Latin phrase *statisticum collegium* (lecture about state affairs).

German: The German term *Statistik* was first used by Gottfried Achenwall in 1749.

English: The anglicized form of the German term was introduced by John Sinclair (a Presbyterian pastor) who published the *Statistical Account of Scotland*, a 21-volume compilation. This work was published at various periods during the 1790s.

As Table 1.1.1 demonstrates, none of the estimates of the percentage of effectiveness were exactly correct, but most were very close. If we cannot determine how much faith we should place in our estimates, it will be difficult to use the estimates to make decisions. The process of selecting samples and determining the reliability of our estimates is a large part of what statistics is about.

1.1 Exercises

Basic Concepts

1. What are three objectives of statistical methods?
2. Briefly describe the role of statistics in managerial decision making.
3. What is a controlled experiment?
4. Explain the purpose of a control group.
5. Explain the purpose of a treatment.
6. How is it possible to know the results of a presidential election before Election Day?
7. What is a population?
8. What is a frame?
9. What is a population parameter?
10. Why is it often difficult to determine the exact value of a population parameter?
11. What is a sample?
12. What is a statistic?
13. Describe the relationships between populations, samples, parameters, and statistics.

Exercises

14. Determine whether the statement describes a population or sample.

    The salaries of all students that graduate from State University with degrees in Information Technology.

15. Determine whether the statement describes a population or sample.

    The number of billable man-hours logged per week by employees at Deloitte.

16. Determine whether the statement describes a population or sample.

    The number of times 6 out of 50 CEOs of technology companies in the Fortune 500 meet with their board of directors during a calendar year.

17. Determine whether the statement describes a population or sample.

    The final rankings of 5 candidates out of the 22 who applied for the open CFO position in your organization.

18. Identify the population being studied.

    The salaries of 13 out of the 30 employees who work during the night shift at a manufacturing plant.

19. Identify the sample chosen for the study.

    The price of homes of a sample of 35 employees who work at a company in Silicon Valley.

20. Identify the population being studied and the sample chosen.

The annual tuition paid by students in a sample of 34 from your class.

21. Determine if the numerical value describes a population parameter or a sample statistic.

The average number of hours students in your statistics class study per week is 15.8.

22. Determine if the numerical value describes a parameter or a statistic.

A survey of 1910 people in the U.S. revealed that 73% of those surveyed work a full-time job.

1.2 Statistics and Quality

"Statistical thinking is critical to improvement of a system."

—Mary Walton, *The Deming Management Method*

Definition

Process

A **process** is a series of actions that changes inputs to outputs.

The idea of a **process** is closely tied to quality control. We encounter processes in all facets of our lives. A simple credit card transaction is a process—the customer inserts or swipes the card, the number is digitally read from the card, there is a credit authorization procedure, and then finally the credit card is approved or rejected for the amount of money that the customer intended to spend. In a business context, a process is a series of steps that produces a product or service. Closely monitoring and continuously improving processes produce high quality products. Monitoring the process means taking measurements of key variables over time. Improving processes means reducing process variation by finding the causes of variation and eliminating them.

Definition

Statistical Process Control

Statistical Process Control (SPC) is a group of statistical methods designed to monitor and control processes.

In order to improve a process there must be an understanding of how the process is currently performing. This requires definition and monitoring of the process. Statistics helps with decisions about how the data will be collected, what data will be needed, and the analysis of the data. In addition to ferreting out production problems, **Statistical Process Control** (SPC) is a group of statistical methods designed to monitor and control processes. SPC is helpful in detecting problems in a process before they create a defective product or service. We will study this subject more extensively in chapter 18.

1.2 Exercises

Basic Concepts

1. Describe the role of statistics in the quality movement.
2. What is a process?
3. How are processes improved?
4. What is SPC?

Exercises

5. **a.** Describe a process at your school or place of employment.

   b. In your opinion, how could this process be improved?

   c. What type of data could you collect to use in analyzing this process?

1.3 Descriptive Statistics versus Inferential Statistics

The science of statistics is divided into two categories, **descriptive** and **inferential**. Descriptive methods describe and summarize data, while inferential methods aid in drawing conclusions and making decisions and predictions about populations and processes for which it is impractical to obtain measurements on each member.

Descriptive Statistics

The emphasis in **descriptive statistics** is analyzing observed measurements, usually from a sample. With descriptive statistics we try to answer questions such as:

Inferential Statistics

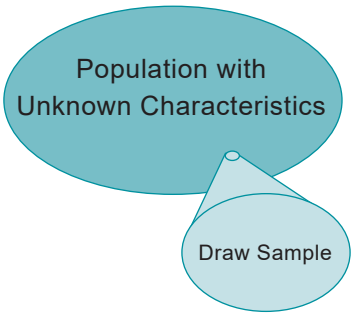
Definition

Inferential Statistics

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

It would be preferable to have measurements of the entire population, but in most cases these data are either not obtainable or would be much too costly to obtain. For example, to be absolutely certain that all car air bags will inflate in head-on collisions would require each new car to be crash tested in a head-on crash. If 100 percent inspection were a requirement, cars would be a scarce commodity. Fortunately for automobile manufacturers, statistical sampling techniques can reliably estimate, with a relatively small sample, the fraction of air bags that will inflate.

Use Sample Data to Make Inferences about Unknown Population Characteristics



Example 1.3.1

Differentiating Between Descriptive and Inferential Statistics

The Michelin tire company has a feature called "Track Connect" that will assist racecar drivers in getting the maximum performance out of their tires when on a racetrack. Michelin has an app that will give personalized advice before, during, and after a driver takes laps around a track. The app will make suggestions for optimal tire pressure and temperature so that the car is handled efficiently as they navigate the track (road track or oval). To evaluate the app, Michelin randomly collected data from 30 drivers that took laps around various track surfaces using several car models, different tires (those with and without the sensors), and under different weather conditions. Using the data from the drivers' experiences on the tracks, Michelin has concluded that the tires lasted longer (i.e., less wear) and the cars had improved gas mileage. Were the results of this experiment an example of descriptive or inferential statistics?

Source: michelinman.com/trackconnect.html

SOLUTION

The Michelin tire company has collected data on tire performance from a random sample of 30 racecar drivers who took laps around various racetracks, under various track conditions. The primary data collected from the tires were air pressure and temperature. They also collected gas mileage (miles per gallon) data for each of the racecars. Using the collected data, Michelin was able to conclude that the tires lasted longer and had better gas mileage than tires without the sensors. This is a case of inferential statistics.

1.3 Exercises

Basic Concepts

1. What is the difference between descriptive and inferential statistics?
2. Name three questions that a descriptive statistic can be used to answer.

Exercises

3. Determine whether the statement describes a descriptive or inferential statistic.

The average price of a car at the new car dealership in town is $28,200.

4. Determine whether the statement describes a descriptive or inferential statistic.

A survey of 885 people revealed that 51% have a college degree; therefore, it can be assumed that 51% of the U.S. population has a college degree.

1.4 The Value of Statistical Literacy

Part of being an intelligent human being is the desire to learn the truth about the world we live in. But as Oscar Wilde said:

"The truth is rarely pure and never simple."

—Oscar Wilde, *The Importance of Being Earnest*

Being statistically illiterate puts one (or one's organization) at a competitive disadvantage compared to companies that possess and use statistical knowledge and analytical tools. Statistics and its uses cannot be avoided. Therefore, learning and using statistical tools will give you and your organization more flexibility when making decisions.

To intelligently appreciate or produce statistical information, you must be statistically literate to defend yourself from a persuasive but fallacious statistical argument, to decrease your vulnerability to pseudo-sciences, and to diminish the chances of making poor and sometimes injurious business decisions.

A statistically literate person understands the language of statistics and understands statistical concepts and reasoning. To become statistically literate, one should be able to think "statistically". This will involve asking questions like:

- Where did the data come from?
- How was the sample taken and is the sample large enough?
- How reliable or accurate were the measures used to generate the reported data?
- Are the reported statistics appropriate for this kind of data?
- Is a graph drawn appropriately?
- How was this probabilistic statement calculated?
- Do the claims make sense?
- Should there be additional information?
- Are there alternative interpretations?

Liar or Statistician?

In his book *How to Tell the Liars from the Statisticians*, Robert Hooke sheds light on our exposure to misleading statistics in everyday life. In the preface he writes, "The science of statistics has made great progress in this century, but progress has been accompanied by a corresponding increase in the misuse of statistics. The public, whether it gets its information from television, newspapers, or news magazines, is not well prepared to defend itself against those who would manipulate it with statistical arguments. Many people either believe everything they hear or come to believe in nothing statistical, which is even worse." Throughout the remaining chapters, Hooke uses examples from politics, economics, entertainment, and the medical community to illustrate the dangers of being statistically illiterate. You might be surprised to learn the ways in which the misuse of statistics affects you every day. In order to digest the plethora of statistical information you encounter, you must first become statistically literate.

Source: Hooke, Robert. *How to Tell the Liars from the Statisticians.* New York, New York: Marcel Dekker Inc., 1983. Print.

1.4 Exercises

Basic Concepts

1. What are the consequences of being statistically illiterate? How could this put you at a disadvantage in business?

2. What kinds of questions would a statistically literate person ask?

Exercises

3. Determine whether the statement describes a descriptive or inferential statistic.

The average price of a car at the new car dealership in town is $28,200.

4. Determine whether the statement describes a descriptive or inferential statistic.

A survey of 885 people revealed that 51% have a college degree; therefore, it can be assumed that 51% of the U.S. population has a college degree.

1.4 The Value of Statistical Literacy

Part of being an intelligent human being is the desire to learn the truth about the world we live in. But as Oscar Wilde said:

"The truth is rarely pure and never simple."

—Oscar Wilde, *The Importance of Being Earnest*

Being statistically illiterate puts one (or one's organization) at a competitive disadvantage compared to companies that possess and use statistical knowledge and analytical tools. Statistics and its uses cannot be avoided. Therefore, learning and using statistical tools will give you and your organization more flexibility when making decisions.

To intelligently appreciate or produce statistical information, you must be statistically literate to defend yourself from a persuasive but fallacious statistical argument, to decrease your vulnerability to pseudo-sciences, and to diminish the chances of making poor and sometimes injurious business decisions.

A statistically literate person understands the language of statistics and understands statistical concepts and reasoning. To become statistically literate, one should be able to think "statistically". This will involve asking questions like:

- Where did the data come from?
- How was the sample taken and is the sample large enough?
- How reliable or accurate were the measures used to generate the reported data?
- Are the reported statistics appropriate for this kind of data?
- Is a graph drawn appropriately?
- How was this probabilistic statement calculated?
- Do the claims make sense?
- Should there be additional information?
- Are there alternative interpretations?

Liar or Statistician?

In his book *How to Tell the Liars from the Statisticians*, Robert Hooke sheds light on our exposure to misleading statistics in everyday life. In the preface he writes, "The science of statistics has made great progress in this century, but progress has been accompanied by a corresponding increase in the misuse of statistics. The public, whether it gets its information from television, newspapers, or news magazines, is not well prepared to defend itself against those who would manipulate it with statistical arguments. Many people either believe everything they hear or come to believe in nothing statistical, which is even worse." Throughout the remaining chapters, Hooke uses examples from politics, economics, entertainment, and the medical community to illustrate the dangers of being statistically illiterate. You might be surprised to learn the ways in which the misuse of statistics affects you every day. In order to digest the plethora of statistical information you encounter, you must first become statistically literate.

Source: Hooke, Robert. *How to Tell the Liars from the Statisticians.* New York, New York: Marcel Dekker Inc., 1983. Print.

1.4 Exercises

Basic Concepts

1. What are the consequences of being statistically illiterate? How could this put you at a disadvantage in business?

2. What kinds of questions would a statistically literate person ask?

Exercises

3. Do some research on the internet and locate an advertisement for a product or service that you suspect may be making a false claim.

a. What leads you to suspect the claim is false?

b. Does the ad include data or statistics in its claim? If so, do the data or statistics reported seem accurate?

c. Does the ad reveal the source of the data and how it was collected?

d. Are there any figures or graphs included in the ad? If so, is the graph appropriate and does it make sense?

2.1 Exercises

Basic Concepts

1. What are the two fundamental problems of measurement?
2. When measurements are used to help solve a problem, what desirable characteristics should the measurements possess?
3. Name and briefly describe three measurement systems commonly used in business.
4. When you encounter any type of data, what three questions should you ask to determine the quality of the measurements?
5. What is the scientific method?
6. What is a confounding variable?
7. How does statistics interact with the steps in the scientific method?
8. Name and briefly describe the two main branches of statistics.
9. What is the decision-making method?
10. What is different between the scientific method and the decision-making method?
11. Are problems that can be solved by collecting data always the result of a system malfunction? Explain.
12. Give an example of how statistics can be used to improve a process.
13. What are fuzzy concepts? What are the measurement problems associated with fuzzy concepts?
14. Give an example of a tool that has been widely accepted as an instrument used to measure a fuzzy concept.
15. What are the two ways of obtaining data?
16. What are the dangers of making conclusions based on poorly collected data?
17. How do you treat the problem of a confounding variable?
18. Explain the difference between the control group and the experimental group in a controlled experiment.
19. What is an explanatory variable?
20. What is a response variable?
21. What is bias? How can it be controlled?
22. What is a completely randomized design? What are the advantages of using a completely randomized design?
23. What is a before and after study?
24. What is the placebo effect? Give an example.
25. What is a double blind study?
26. How do observational studies differ from controlled experiments?
27. What kinds of problems can be associated with an observational study?
28. Researchers use surveys for two main purposes. Name and give an example of each.

Exercises

29. Specify whether the following variables are well-defined or not. Justify your answer.

- **a.** Height
- **b.** Weight
- **c.** Hot
- **d.** Temperature
- **e.** Beauty

30. A researcher has developed a test that reportedly measures intelligence. The test includes questions such as:

$$\textit{What is the lowest common denominator of the fractions } \frac{5}{32} \textit{ and } \frac{6}{9}?$$

Who invented the digital computer?

Is it reasonable to measure intelligence with these questions? Discuss.

31. 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.

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

33. The manager of an electronics company was interested in determining the reason for the increase in sales volume over the last three years. The manager randomly selected data on the advertising budget, number of salespeople, and average product costs. When examining the data, the manager found that her average product costs were fairly stable but the advertising budget steadily increased over the last two years along with the number of salespeople. Are there any confounding variables in this study? If so, what are they and why do you consider them confounding?

34. A company that produces bulbs for projectors wanted to conduct an experiment to determine the length of life of its bulbs. The company's leading competitor's bulbs have an average life of 1000 hours. The company sampled its bulbs and found that the average life of the bulbs was 1200 hours. Thus, the company has concluded and advertises that its bulbs last longer than the competition by at least 100 hours. Were the results of this experiment an example of descriptive or inferential statistics? Explain your answer.

35. 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*. 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 were gathered from an observational study instead of from a controlled experiment. How would this affect the conclusions that might be made from the study?

36. 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. 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 were gathered from an observational study instead of from a controlled experiment. How would this affect the conclusions that might be made from the study?

37. 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.

Source: Science, Vol. 259, No. 5099

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?

38. 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.

Source: Self magazine

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?

39. 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.

40. 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.

Source: Self magazine

- **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?

41. A religious group conducted a survey with two of the questions asking "Do you go to church?" and "Are you happy?" After conducting the survey, the group concluded that those who go to church are generally happier than those that do not go to church. Do you think going to church makes one happier? Describe how confounding variables could play a role with the conclusion drawn by the religious group.

42. In May 2011, Internet Explorer reversed its trend in the United States and gained usage share (the percentage of users using a particular Internet browser). In June of 2011, the trend reversal became global. Internet Explorer gained 0.57% in June across all operating systems with Internet Explorer 8.0 gaining 0.86% globally. The gains for Internet Explorer came primarily at the expense of Mozilla Firefox (−0.51%). Google Chrome's pace of usage share gains slowed to +0.2% for June. The gains for IE were the largest in Europe and Asia:

Internet Explorer in Europe: +0.88%

Internet Explorer in Asia: +0.81%

This increase may be the result of a marketing campaign. In early June, Microsoft launched their "Confidence" campaign aimed at showing the security features of Internet Explorer 8.

Source: netmarketshare.com

- **a.** Are the results stated above likely to have come from an observational study?
- **b.** How can Microsoft (and other companies) benefit from this information?

43. 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?

companies can gain a deeper understanding of their customers to better meet their needs. Companies also use analytics to develop models to drive insights into the past, present, and future.

There are primarily two types of analytics—predictive and prescriptive. We use both types of analytics to help us gain insight into the future. **Predictive analytics** answers the question of what could happen. **Prescriptive analytics** answers the question of what should happen.

Retail stores use predictive analytics to predict products that customers will buy, times that they will log on to specific sites, or even the amount of time that a customer may spend on a site. Being able to make these predictions allow retailers to better tend to customer needs as well as make the business more profitable.

After prediction, prescriptive analytics focuses on how to take advantage of future opportunities of the decision-making process. Prescriptive analytics is considered the future of business analytics. It provides an organization with adaptive, automated, and time-dependent courses of action to take advantage of business opportunities. For example, in finance, prescriptive analytics can be used to help investors select which investments to purchase. In sports, prescriptive analytics can help teams determine which player to draft or trade.

Analytics plays a major role in the past, present, and future of the decision-making processes of many organizations. Businesses of all kinds that use Big Data and analytics can improve their decision-making processes. Applications such as banking, health care, etc., need predictive and prescriptive analytics to improve their standards and quality to help their customers and themselves.

Definition

Predictive Analytics

Predictive analytics uses past data to develop models that can help determine what future events are most likely to happen.

Definition

Prescriptive Analytics

Prescriptive analytics is the development of models that help us answer the question, "What should we do moving forward?"

Data Resources

We live in a data rich society. Anyone with access to a personal computer can access thousands of different databases throughout the internet. These databases are packed full of observational data. The website for this textbook, stat.hawkeslearning.com, provides numerous links to data resources. However, some of the largest, most credible, and most commonly used databases are:

- Amazon Web Services
- Centers for Disease Control and Prevention (CDC)
- Data.gov (Data regarding the U.S.)
- Federal Reserve Economic Data (FRED)
- Organization of Economic Cooperation and Development (OECD)
- The World Bank
- The World Factbook (CIA)
- UNdata (United Nations)
- United States Census Bureau
- World Health Organization

2.2 Exercises

Basic Concepts

1. Where is Big Data used?
2. What are the three (sometimes four) characteristics of Big Data?
3. Give another example of where we can find semi-structured data.
4. What is business analytics?
5. How is business analytics used?
6. What is the difference between predictive and prescriptive analytics?
7. Give an example of a company using predictive analytics to make business decisions.
8. Give an example of a company using prescriptive analytics to help their organization moving forward.

Exercises

9. GOES satellites (GOES-16 & GOES-17) provide continuous weather imagery and monitoring of meteorological and space environment data across North America. These satellites provide the kind of continuous monitoring necessary for intensive data analysis. They hover continuously over one position on the surface. Describe three characteristics of Big Data that would be produced by these satellites.
10. For the data in question 9, would the data collected be described as: structured, unstructured, or semi-structured? Explain your choice of answer.

11. The following sample of data about BMI (Body Mass Index) was obtained by the WHO (World Health Organization). What kind of data analytics can be done on these data?

Mean BMI (kg/m²) [age-standardized estimate] 18+ years, 2016

Country	Both sexes	Male	Female
Afghanistan	23.4 [22.0–24.8]	22.6 [20.1–25.1]	24.1 [23.0–25.3]
Albania	26.7 [25.8–27.5]	27.0 [25.8–28.2]	26.3 [25.0–27.6]
Algeria	25.5 [24.5–26.5]	24.7 [23.4–26.1]	26.4 [24.9–27.8]
Andorra	26.7 [24.6–28.7]	27.3 [24.8–29.8]	26.1 [22.8–29.5]
Angola	23.3 [21.2–25.6]	22.3 [19.7–25.0]	24.3 [20.9–27.7]
Antigua and Barbuda	26.7 [24.6–28.8]	25.7 [23.2–28.2]	27.7 [24.4–31.0]
Argentina	27.7 [26.8–28.6]	27.8 [26.6–29.0]	27.6 [26.3–28.8]
Armenia	26.3 [25.8–26.9]	25.6 [24.8–26.3]	27.0 [26.1–27.8]
Australia	27.1 [26.6–27.6]	27.6 [26.9–28.2]	26.7 [26.0–27.4]
Austria	25.6 [24.3–26.8]	26.5 [24.8–28.2]	24.6 [22.6–26.5]

Source: World Health Organization https://apps.who.int/gho/data/view.main.CTRY12461?lang=en

12. JetBlue Airlines collects data which includes passenger ID number, name, date of birth, country of birth, country of residence, frequent flyer number, departure airport, destination airport, airfare paid, and flight number. How can this information be used to make business decisions?

13. Mobile phone companies use the GPS feature to determine the location of users and to provide location-based services (LBS) such as information, entertainment, and security. Describe how a marketing company can use location-based services analytics to provide targeted ads about a nearby retail store to a user.

14. Describe how prescriptive analytics can be used in sports to determine which player to draft or trade.

15. Credit card companies are some of the biggest users of data analytics. How can location-based services (LBS) be used to prevent credit card fraud?

16. Many hospitals and health care providers are now utilizing electronic health records (EHR). This of course generates an immense amount of patient data; hence the need for data analytics. Describe how this data can be used to improve service and better patient care.

SOLUTION

a. The number of gigabytes used has a meaningful zero point and the ratio of two values for gigabyte usage is meaningful. That is, one who uses 16 gigabytes when compared to one who uses 4 gigabytes, a ratio of 4 can be calculated indicating that the customer using 16 gigabytes has used four times as much data as the customer using 4 gigabytes. Thus, the number of gigabytes used is measured on the ratio scale.

b. The student's response on the faculty evaluation is measured on the ordinal scale since we can have some order associated with the responses but cannot perform arithmetic operations.

c. As stated earlier in the text, temperature (measured in degrees Celsius or Fahrenheit) is measured on the interval scale. In this case, 0 degrees Fahrenheit does not mean the absence of temperature which prevents us from calculating meaningful ratios.

d. The customer's response will be only the name of the shoe brand. We cannot perform any arithmetic operations or ranking. Additionally, by the name only, one brand cannot be considered more (or better) than the other. Thus, shoe brand is measured on the nominal scale.

2.3 Exercises

Basic Concepts

1. What are qualitative data? Give an example.

2. What are quantitative data? Give an example.

3. Which levels of measurement are associated with qualitative data? Which levels are associated with quantitative data?

4. If data are quantitative, they can be further classified into two categories. Name and briefly describe these categories.

5. What is the difference between discrete and continuous data?

6. What is a level of measurement?

7. What are the four levels of measurement? Give an example of each.

8. For which level(s) of measurement is arithmetic appropriate?

9. What is the primary difference between nominal and ordinal data?

10. What is an arbitrary zero value? Which level of measurement has this property?

11. What is the fundamental difference between interval and ratio data?

12. Decision makers usually prefer to consider data that possess which level of measurement?

Exercises

13. 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.

14. Identify the following variables as discrete or continuous.

- **a.** The number of on-time flights at the Hartsfield 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 U.S. cities.
- **e.** The number of alcoholics who are men.

15. 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*. The data are 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.** Why is some method of data summary necessary here?

16. 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. 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?

17. Under most states' auto lemon laws, dealers or car makers must replace defective autos that aren't successfully repaired after three attempts or that remain in the shop for 30 days. The table below shows data for Hawaii for the year 2010, weighing car makers' lemons against statewide market share. Assume the "lemon index" is the share of the complaints divided by the total market share for each manufacturer.

Lemon Index: Hawaii, 2010

Best	Lemon Index	Worst	Lemon Index
Toyota (includes Lexus)	0.212	Chrysler (includes Dodge and Jeep)	6.512
Honda	0.462	Kia	2.750
Ford (includes Lincoln)	0.868	GM (includes Chevrolet, GMC, Buick)	2.375
Nissan (includes Infinity)	1.056	BMW	2.129
Mazda	1.833	Hyundai	2.000

Source: Hawaii.gov

Answer the following questions for the variable "Lemon Index".

- **a.** Are the data quantitative or qualitative? Why?
- **b.** What is the highest level of measurement these data could have?

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

- **a.** The temperature (in degrees Fahrenheit) of patients with pneumonia.
- **b.** The age at which the average male marries.
- **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.

19. Determine the level of measurement (nominal, ordinal, interval, or ratio) 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, Never.
- **d.** The number of pounds of snack food eaten by an individual in his or her lifetime.

20. Given the table below on browser usage, what is the highest level of measurement that these data could have? Justify your answer.

Browser Usage Share (%)				
Month	Microsoft Internet Explorer	Mozilla Firefox	Google Chrome	Apple Safari
July 2010	60.74	22.91	7.16	5.09
August 2010	60.48	22.90	7.50	5.15
September 2010	59.62	22.97	7.99	5.27
October 2010	59.18	22.83	8.50	5.36
November 2010	58.44	22.76	9.26	5.55
December 2010	57.08	22.81	9.98	5.89
January 2011	56.00	22.75	10.70	6.30
February 2011	56.77	21.74	10.93	6.36
March 2011	55.92	21.80	11.57	6.61
April 2011	55.11	21.63	11.94	7.15
May 2011	54.27	21.71	12.52	7.28
June 2011	54.84	21.20	12.72	7.41

2.4 Time Series Data and Cross-Sectional Data

Time Series Data

Recall from Chapter 1, the science of statistics is divided into two categories: descriptive statistics and inferential statistics. Fundamental to the concept of statistical inference is the notion of population—the total collection of measurements. **Time series data** originate as measurements usually taken from some process over equally spaced intervals of time.

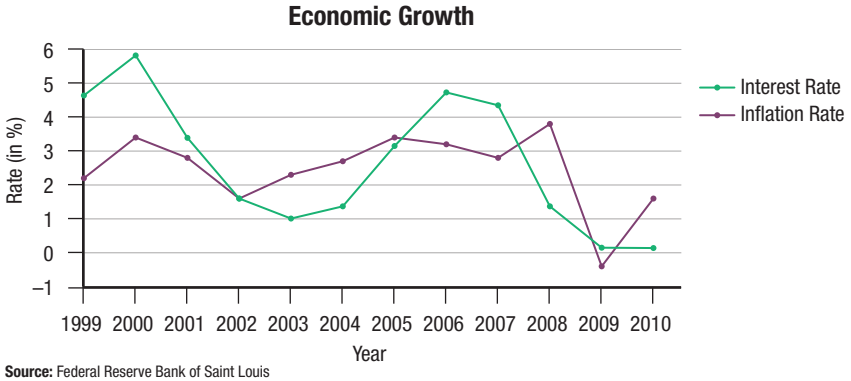
2.4 Exercises

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 are cross-sectional data?
7. What is the difference between cross-sectional data and time series data?

Exercises

8. Consider the following graph of long-term interest rates (10-year treasury notes) and inflation rates.



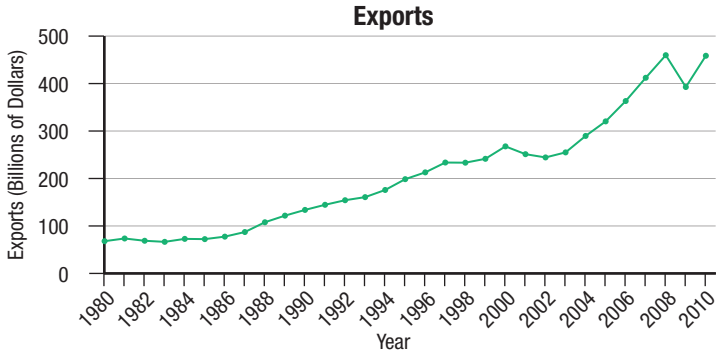
Source: Federal Reserve Bank of Saint Louis

   a. Are the interest rate data presented above time series data?

   b. Are the inflation rate data presented above time series data?

   c. For each of parts **a.** and **b.** if the data is time series data, does the series appear to be stationary or nonstationary?

9. Consider the following graph of total exports.



Source: Federal Reserve Bank of Saint Louis

   a. Are these data time series data?

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

10. Using a newspaper, journal, or website as your source, give an example of time series data. Be sure to reference your source and give a brief description of the data.

11. The following table shows the annual average crude oil price from 1946 through 2011. Prices are adjusted for inflation to April 2011 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 $102.26 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 $16.44. 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?

Annual Average Domestic Crude Oil Prices ($ per Barrel)

Year	Nominal	Inflation Adjusted (April 2011)	Year	Nominal	Inflation Adjusted (April 2011)
1946	1.63	18.49	1979	25.10	77.05
1947	2.16	21.73	1980	37.42	102.26
1948	2.77	25.92	1981	35.75	88.55
1949	2.77	26.17	1982	31.83	74.24
1950	2.77	25.90	1983	29.08	65.69
1951	2.77	24.00	1984	28.75	62.26
1952	2.77	23.47	1985	26.92	56.28
1953	2.92	24.50	1986	14.44	29.62
1954	2.99	25.04	1987	17.75	35.13
1955	2.93	24.57	1988	14.87	28.32
1956	2.94	24.35	1989	18.33	33.24
1957	3.14	25.12	1990	23.19	39.80
1958	3.00	23.38	1991	20.20	33.36
1959	3.00	23.15	1992	19.25	30.85
1960	2.91	22.15	1993	16.75	26.09
1961	2.85	21.44	1994	15.66	23.76
1962	2.85	21.19	1995	16.75	24.73
1963	2.91	21.39	1996	20.46	29.32
1964	3.00	21.75	1997	18.64	26.12
1965	3.01	21.47	1998	11.91	16.44
1966	3.10	21.48	1999	16.56	22.30
1967	3.12	21.04	2000	27.39	35.76
1968	3.18	20.53	2001	23.00	29.23
1969	3.32	20.36	2002	22.81	28.50
1970	3.39	19.65	2003	27.69	33.86
1971	3.60	20.00	2004	37.66	44.81
1972	3.60	21.44	2005	50.04	57.57
1973	4.75	23.87	2006	58.30	65.03
1974	9.35	42.58	2007	64.20	69.51
1975	12.21	51.00	2008	91.48	95.25
1976	13.10	51.78	2009	53.48	55.96
1977	14.40	53.41	2010	71.21	73.44
1978	14.95	51.58	2011 (Partial)	86.84	–

Source: www.inflationdata.com

12. Do you think the pay of executives working for digital companies increases/decreases as the company's stock price increases/decreases? Examine the following table.

CEO Compensation and Stock Performance

Exec	Salary/ Bonus ($)	Stock/ Options ($)	Other Non-Equity Compensation ($)	Total 2007 Compensation ($)	Change from 2006 Compensation (%)	2007 Stock Performance (%)
Tom Rogers (Tivo)	800,000	6,200,000	495,075	7,495,075	+102	+32
Mel Karmazin (Sirius)	5,250,000	–	18,743	5,268,743	+23	–23
Paul Sagan (Akamai)	403,651	3,554,264	497,362	4,455,277	–40	–48
Reed Hastings (Netflix)	850,000	1,568,307	270	2,418,577	+5	–6
Rob Glaser (RealNetworks)	1,169,384	643,400	354,200	2,166,984	–26	–45
Bobby Kotick (Activision Blizzard)	899,560	1,188,467	–	2,088,027	+6	+49
Magid M. Abraham (comScore)	421,952	1,125,000	–	1,546,952	+185	–16
Barry Diller (IAC)	500,000	–	927,429	1,427,429	+270	+21
John S. Riccitiello (Electronic Arts)	750,000	–	625,350	1,375,350	–37	–38
Steve Ballmer (Microsoft)	1,340,833	–	10,001	1,350,834	N/A	0
Wayne T. Gattinella (WebMD)	830,000	–	9214	839,214	+6	+10

Source: paidContent.org

What type of data is in the Salary/Bonus column? What do you think about executive salaries as a function of the company's stock performance? Justify your responses.

Table 3.1.6 – Frequency Distribution of Responses

"I am aware that I can reduce exposure to system compromise by restricting who uses my personal computer."	
Strongly Agree	520
Slightly Agree	435
Neutral	310
Slightly Disagree	115
Strongly Disagree	90

Definition

Relative Frequency Distribution

A **relative frequency distribution** summarizes data into classes and provides in tabular form a list of the classes along with the proportion (or percentage) of observations in each class.

In Table 3.1.7 and Table 3.1.8, **relative frequency distributions** are calculated. In these tables, the frequencies are converted into percentages. These are defined as **relative frequencies**, i.e., the proportion relative to the total. They are valuable in assessing the data quickly in terms we use frequently. (See Section 3.3 for an additional discussion of relative frequency distributions.)

Table 3.1.7 – Relative Frequency Distribution of Responses

"I am aware that there are measures that I can take to help protect my personal information on my personal computer."	
Strongly Agree	35%
Slightly Agree	27%
Neutral	21%
Slightly Disagree	10%
Strongly Disagree	7%

Table 3.1.8 – Relative Frequency Distribution of Responses

"I am aware that I can reduce exposure to system compromise by restricting who uses my personal computer."	
Strongly Agree	35%
Slightly Agree	30%
Neutral	21%
Slightly Disagree	8%
Strongly Disagree	6%

As one can see in Tables 3.1.5 through 3.1.8, the majority of students are aware that there are measures that they can take to protect the information on their personal computers. Summarizing the qualitative data (via a frequency distribution table or relative frequency distribution table) allows the researcher to make conclusions about the data without having to view each observation.

3.1 Exercises

Basic Concepts

1. From a comprehension standpoint, what are the advantages of visual images over the written word?
2. Describe two situations in which graphical displays are used in business.
3. Describe the purpose of a frequency distribution.
4. What are the basic questions to ask when examining the structure of a data set?
5. What are the two steps to constructing a frequency distribution?
6. In the construction of a frequency distribution, what are the two requirements that the classification categories must meet?

Exercises

7. In order to help him decide when and where to advertise, a local repairman decided to pull his invoices for the month of June and tally what types of machines he had worked on. There were forty-eight items repaired that month.

Office copier	Washing machine	Air conditioner
Air conditioner	Fan	Lawn mower
Lawn mower	Air conditioner	Fan
DVD Player	Fan	Air conditioner
Air conditioner	Lawn mower	Washing machine
Lawn mower	Air conditioner	Stereo
Exercise bike	DVD Player	Air conditioner
Air conditioner	Lawn mower	Lawn mower
Lawn mower	Air conditioner	Fan
Radio	Washing machine	Air conditioner
Air conditioner	Radio	Stereo
Fan	Air conditioner	Lawn mower
Washing machine	Lawn mower	Air conditioner
Air conditioner	Fan	Fan
Lawn mower	Air conditioner	DVD player
Washing machine	Washing machine	Air conditioner

a. What level of measurement do the data possess?

b. Are the data qualitative or quantitative?

c. Construct a frequency distribution for the data. Any machine types worked on three or fewer times are classified as miscellaneous.

8. 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*. 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 do 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?

9. 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? Discuss the implications of this when reviewing data collected and distributed by someone else for open answer questions.

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

3.2 Displaying Qualitative Data

Graphical analysis is a trade-off. We lose sight of the individual observations (the raw data). In return, we are able to see a representation of the totality of observations. 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. Fortunately, the development of graphics software has made the creation of sophisticated graphs quite easy.

Several types of graphs and tabular displays will be discussed in this chapter. Bar charts, stacked bar charts, 3-D bar charts, and pie charts are effective, visually appealing methods of graphically displaying qualitative data. An examination of publications such as *Time*, *USA Today*, *The Wall Street Journal*, *Scientific American*, or *Forbes* provides convincing evidence of the frequent and beneficial usage of these graphical display techniques.

Definition

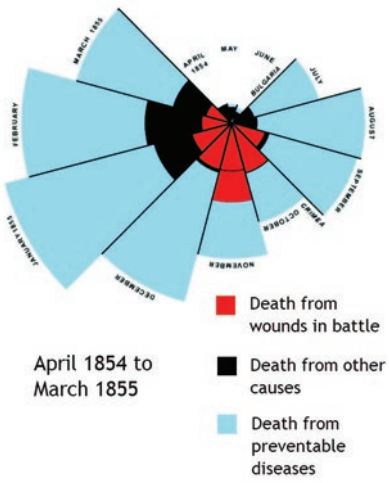
Bar Chart

The **bar chart** is a simple graphical display in which the length of each bar corresponds to the number of observations in a category.

Bar Charts

Bar charts are often used to illustrate a frequency distribution for qualitative data.

Bar charts are valuable as presentation tools and are especially effective at reinforcing differences in magnitudes, since they permit the visual comparison of data by displaying the magnitude of each category by a vertical or horizontal bar. Figure 3.2.1 is a bar chart constructed from majors of the students in a business statistics course.



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.

In order to determine from the pie chart how much was spent in a particular category, multiply the total amount by the proportion given for that category in the pie chart. For example, to find the dollar amount of government funds spent on Medicare and Medicaid, multiply the total amount of government expenditures ($4.4 trillion) by the proportion spent on Medicare and Medicaid. This means that $4.4(0.25) = $1.1 trillion of all government expenditures goes to Medicare and Medicaid.

Federal Government Spending, Fiscal Year 2019

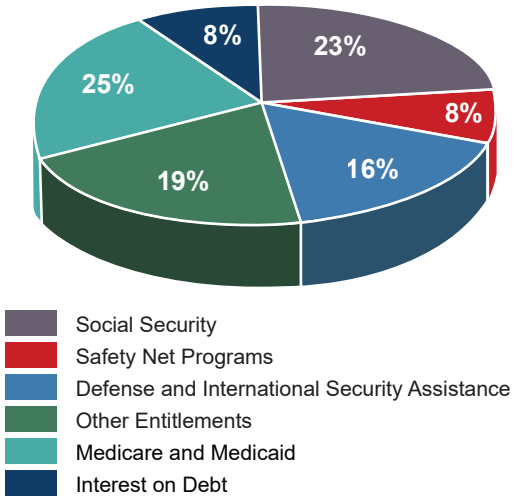


Figure 3.2.12

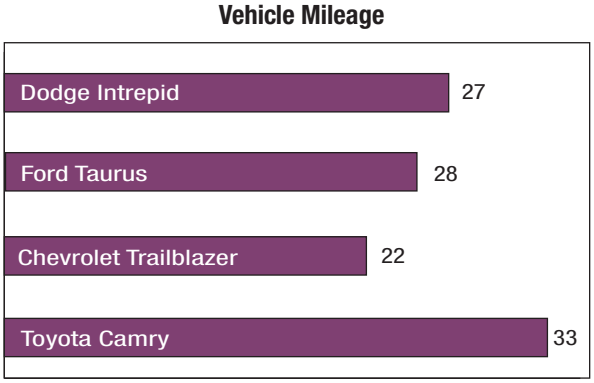
3.2 Exercises

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. Why would a stacked bar chart be preferred over a normal bar chart?
8. What is one disadvantage of using a 3-D chart?
9. What is a pie chart?
10. What is the main advantage of using a pie chart?

Exercises

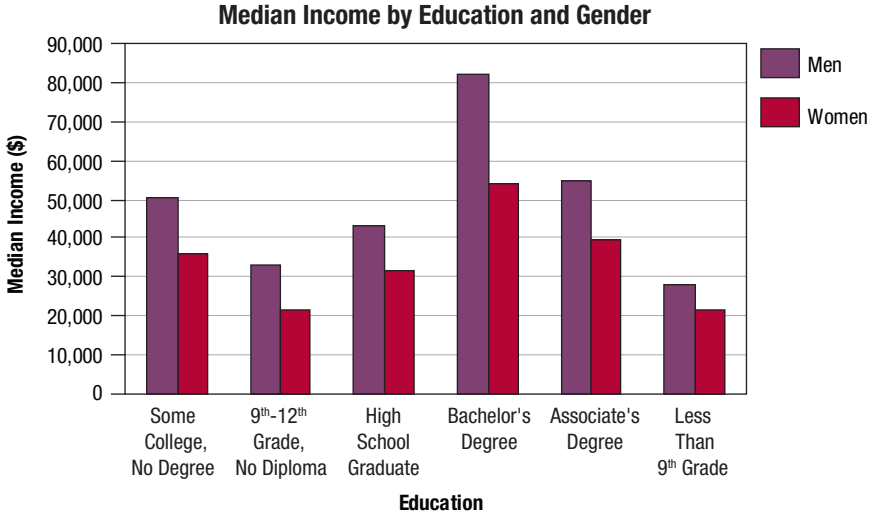
11. A consumer magazine uses bar charts to compare four popular brands of automobiles. This particular bar chart represents a comparison of the miles per gallon (mpg) for the four brands.



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.

12. The following bar chart presents the median income of U.S. employees by education level and gender. Evaluate the bar chart using the guidelines suggested in the section on the aesthetics of bar chart construction.



13. Consider the following data regarding the number of wildfires in the U.S. categorized by the size class in acres and the cause of the fire.

Number of Wildfires in the U.S.		
Size Class (Acres)	**Lightning-Caused**	**Person-Caused**
0.25 or less	4637	2367
0.26 – 9	1940	1904
10 – 99	219	571
100 – 299	44	103
300 – 999	26	52
1000 – 4999	43	17
5000 +	21	9
Total	6930	5023

a. Construct a bar chart for the number of wildfires caused by lightning.

b. Construct a bar chart for the number of wildfires caused by people.

c. Construct a stacked bar chart for the number of wildfires caused by lightning and the number of wildfires caused by people.

d. Construct a pie chart for the number of wildfires caused by lightning.

e. Construct a pie chart for the number of wildfires caused by people.

f. What did you learn from the charts created in parts **a.** through **e.**?

14. Consider the following data regarding the average spending on healthcare per person for various countries.

Healthcare Costs Around the World per Capita, 2009	
Country	**Average Cost ($)**
Australia	2886
Canada	2998
Denmark	2743
Finland	2104
France	3048
Germany	2983
Iceland	3159
Ireland	2455
Japan	2249
Sweden	2745
Switzerland	3847
United Kingdom	2317
United States	5711

Source: www.creditloan.com

a. Construct a bar chart for the average healthcare cost per person for the various countries.

b. What did you learn from the chart?

15. Consider the following data regarding professions with high projected growth rates for the years 2008 through 2018.

Occupation Growth Rates

Occupation	Projected Increase 2008 – 2018
Biomedical engineers	72%
Network systems and data communications analysts	53%
Home health aides	50%
Personal and home care aides	46%
Financial examiners	41%
Medical scientists, except epidemiologists	40%
Physician assistants	39%
Skin care specialists	38%

Source: Bureau of Labor Statistics

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

b. What did you learn from the chart?

16. Consider the following data regarding the methods which consumers use to pay for items purchased in a particular store.

Payment Methods

Method of Payment	Relative Frequency
Cash	81.1%
Checks	7.6%
General-Purpose Credit Cards	5.5%
Proprietary Credit Cards	5.3%
Debit Cards	0.5%

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.

Cumulative Frequency Distribution

The cumulative frequency distribution gives the reader an opportunity to look at any category and determine immediately the number of observations that belong to a particular category and all categories below it.

Definition

Cumulative Frequency

The **cumulative frequency** is the sum of the frequency of a particular class and all preceding classes.

Table 3.3.4 – Cumulative Frequency Distribution of Revenue Data

Revenue (Millions of Dollars)	Frequency	Cumulative Frequency
0 to 40	50	50
41 to 81	30	80
82 to 122	14	94
123 to 163	3	97
164 to 204	1	98
205 to 245	0	98
246 to 286	1	99
287 to 327	0	99
328 to 368	0	99
369 to 409	1	100

In this example, the reader can easily see in Table 3.3.4 that 97 out of 100 revenues are less than or equal to $163 million.

Cumulative Relative Frequency

To obtain the cumulative relative frequency, add the relative frequencies of all preceding classes to the relative frequency of the current class.

Definition

Cumulative Relative Frequency

The **cumulative relative frequency** is the proportion of observations in a particular class and all preceding classes.

Table 3.3.5 – Cumulative Relative Frequency Distribution of Revenue Data

Revenue (Millions of Dollars)	Frequency	Relative Frequency	Cumulative Relative Frequency
0 to 40	50	0.50	0.50
41 to 81	30	0.30	0.80
82 to 122	14	0.14	0.94
123 to 163	3	0.03	0.97
164 to 204	1	0.01	0.98
205 to 245	0	0.00	0.98
246 to 286	1	0.01	0.99
287 to 327	0	0.00	0.99
328 to 368	0	0.00	0.99
369 to 409	1	0.01	1.00

From the cumulative relative frequency in Table 3.3.5, it is easy to see that 97% of the revenues are less than or equal to $163 million.

3.3 Exercises

Basic Concepts

1. What are the fundamental decisions in constructing frequency distributions for quantitative data?
2. Describe the general guidelines for selecting the number of classes for a quantitative frequency distribution.
3. What is a good starting point for determining the class width?

4. What is a relative frequency distribution? How do you calculate relative frequencies from raw frequencies?
5. What is a cumulative frequency distribution?
6. What is a cumulative relative frequency distribution?

Exercises

7. 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	**Frequency**
0 – 6	15
7 – 13	21
14 – 20	27
21 – 27	9
28 – 34	2
35 and above	1

   a. Construct a relative frequency distribution.

   b. Construct a cumulative frequency distribution.

8. The closing prices (in pence) for selected stocks trading on the London Stock Exchange were as follows. Construct a frequency distribution for the stock prices.

Closing Prices	
Stock	**Closing Price (Pence)**
Allied Lyons	439
Babcock	208
Barclays Bank	543
Bass Ltd	992
British GE	238
Cadbury Sch	257
Guinness	379
Hanson Trust	169
Lucas Indus	655
Reed Int'l	467
STC	318
Tate & Lyle	833
Thorm EMI	741
Utd. Biscuit	326

9. Every year, the average temperatures of 100 selected U.S. cities are published by the National Oceanic and Atmospheric Administration. 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.

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.

10. Consider the assets (in billions of dollars) of the 10 largest life insurance companies listed in the following table.

Assets (Billions of Dollars)				
148.4	110.8	55.6	52.4	50.4
42.7	41.7	36.2	35.7	35.7

a. Construct a frequency distribution for the assets (in billions of dollars) of the 10 largest life insurance companies.

b. Construct a relative frequency distribution for the assets (in billions of dollars) of the 10 largest life insurance companies.

c. Construct a cumulative frequency distribution for the assets (in billions of dollars) of the 10 largest life insurance companies.

3.4 Graphical Displays of Quantitative Data

Several types of graphs and tabular displays will be discussed in this section such as histograms, line graphs, stem-and-leaf displays, and dot plots.

Histograms

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

Each of the classes in the frequency distribution is represented by a vertical bar whose height is proportional to the frequency of the interval. The horizontal boundaries of each vertical bar correspond to the class boundaries. Once the frequency distribution has been calculated, all the information necessary for plotting a histogram is available. In Figure 3.4.1, the histogram is created from the frequency distribution of the revenue data in Table 3.3.2.

Definition

Histogram

A **histogram** is a bar graph of a frequency or relative frequency distribution in which the height of each bar corresponds to the frequency or relative frequency of each class.

Table 3.4.13 – Percentage of Obese Adults by US County	
FIPS Code	Percentage of Obese Adults 2016
1001	30.5
1003	26.6
1005	37.3
1007	34.3
1009	30.4
…	…

Using our new data, we generate the following map.

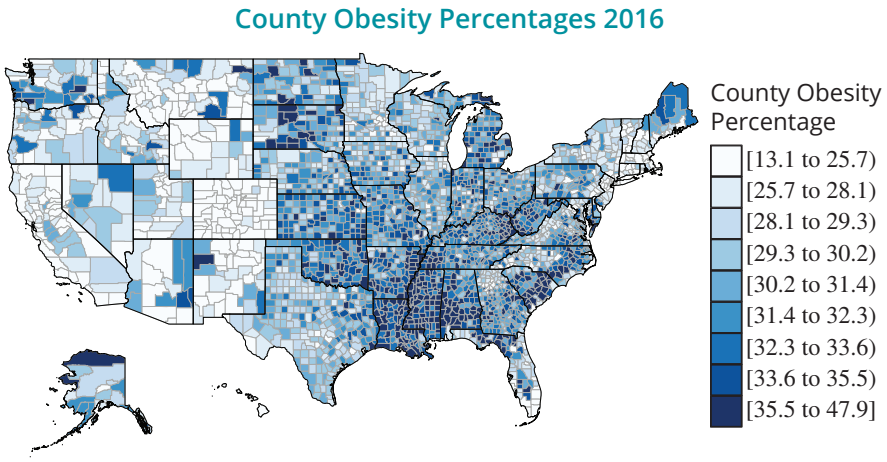
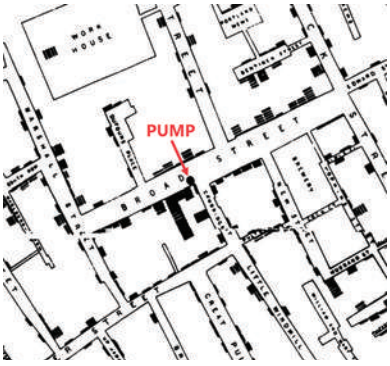


Figure 3.4.12

As you can see, Figure 3.4.12 is substantially different from Figure 3.4.11, and comes along with an entirely new set of conclusions. If we had used Figure 3.4.11 to come to a conclusion, we would have assumed that the Southwest and Northeast regions are the areas in the country that struggle with obesity the most. However, once we normalized our obesity variable by making it a ratio of the total county population, Figure 3.4.12 seems to suggest that it is actually the Southeast and the Midwest that struggle with obesity the most. The data shows that these regions have a higher percent of adults who are obese relative to their total population than the Northeast and Southwest do. This is a prime example of why it is necessary to normalize data when making comparisons.



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.

3.4 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 display is sometimes called a "hybrid graphical method."
6. Identify the advantages of a stem-and-leaf display.
7. Consider the following data value: 39. What would be the stem and leaf for this value if we identified the stem as the tens digit? What would be the steam and leaf if we identified the stem as the hundreds digit?

8. When constructing a stem-and-leaf display, 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. What is a dot plot?

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. Identify a variation on a time series plot that can make the data more visually interesting.

Exercises

17. The closing prices (in dollars) for selected stocks trading on the New York Stock Exchange and NASDAQ Exchange were as follows.

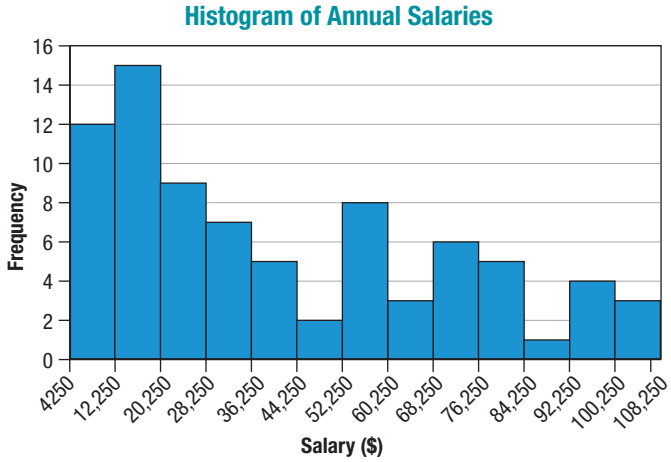
Closing Prices

Stock	Closing Price ($)
Citigroup (C)	34.70
Pfizer (PFE)	22.34
Herbalife (HLF)	69.72
JP Morgan Chase (JPM)	44.34
Intel (INTC)	28.07
WalMart (WMT)	60.67
Microsoft (MSFT)	31.52
PepsiCo (PEP)	66.17
General Motors (GM)	24.81
Verizon Communications (VZ)	37.66
Southwest Airlines (LUV)	8.31
Sprint Nextel (S)	2.76
Yahoo! Inc (YHOO)	15.07
International Business Machines (IBM)	205.47

a. Construct a frequency distribution for the closing prices.

b. Construct a histogram for the closing prices.

18. A sample of 80 laborers is selected from a large city and their annual salaries are determined. The following histogram summarizes the data.



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

b. How many of the laborers earn at least $36,250?

c. What percent of the laborers earn at most $28,249?

d. What percent of the laborers earn at least $84,250?

19. A nutritionist is interested in knowing the percent of calories from fat which Americans consume on a daily basis. To study this, the nutritionist randomly selects 25 Americans and evaluates the percent of calories from fat consumed in a typical day. The results of the study are as follows.

Calories from Fat per Day (%)				
34	18	33	25	30
42	40	33	39	40
45	35	45	25	27
23	32	33	47	23
27	32	30	28	36

a. Construct a frequency distribution for the percent of calories from fat.

b. Construct a relative frequency distribution for the percent of calories from fat.

c. Construct a histogram of the relative frequency distribution.

d. Comment on any information about the percent of calories from fat consumed by the participants in the study which you were able to ascertain by examining the distributions and the histogram.

20. Consider the assets (in billions of dollars) of the 10 largest commercial banks listed in the following table.

Assets (Billions of Dollars)				
216.9	138.9	115.5	110.3	103.5
98.2	76.4	64.0	53.5	49.0

a. Construct a frequency distribution for the assets (in billions of dollars) of the 10 largest commercial banks.

b. Construct a relative frequency distribution for the assets (in billions of dollars) of the 10 largest commercial banks.

c. Construct a histogram of the relative frequency distribution.

d. Comment on any information about the assets (in billions of dollars) of the 10 largest commercial banks which you were able to ascertain by examining the distributions and the histogram.

21. Fifty hospitals in a western state were polled as to their basic daily charges for a semi-private room. The results are listed in the following table, rounded to the nearest dollar.

Daily Charges for Semi-Private Rooms (Dollars)									
125	135	148	156	248	215	156	148	135	149
178	156	135	125	214	256	258	265	156	148
123	147	189	199	189	248	215	259	158	235
268	269	158	198	147	258	269	239	288	199
179	179	189	169	258	178	257	249	259	259

a. What level of measurement do 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 data in the following table are the toxic emissions (in thousands of tons) for 10 states in the United States.

Toxic Emissions (Thousands of Tons)									
206	147	441	128	127	133	422	152	114	134

Source: Toxics in the Community, U.S. Environmental Protection Agency

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

b. Comment on any information about the toxic emissions (in thousands of tons) of the 10 states that you were able to ascertain by examining the stem-and-leaf display.

23. Consider the following highway miles per gallon for 19 selected models of mini-compact, sub-compact, and compact cars.

Miles per Gallon									
26	46	36	31	28	28	27	38	42	36
37	33	23	29	37	34	29	40	28	

a. Construct a stem-and-leaf display for the data.

b. Comment on any information about the highway mpg of the selected models which you were able to ascertain by examining the stem-and-leaf display.

24. An instructor is interested in comparing exam scores for fraternity and non-fraternity males in her class. Meaningful comparisons between two sets of data can be made using a side-by-side stem-and-leaf display. To illustrate this, note the following display summarizing the scores.

Leaf (Non-Fraternity)	Stem	Leaf (Fraternity)
	0	9
2	1	4 0 8
	2	5 7 9 4 5 5 1
3 9	3	2 6 6 9 7 7 3 2 1 6 0
	4	2 7 5
5 6 4 8 9 9 0 2	5	4 7 6 7
4 4 7 8 1 0 3 2 2 6 8 9	6	6 8 9 9 5
5 4 7 8 4 3 8 8 9 1	7	3 4 2 7 8 6 7 4 3
2 9 7 4	8	4 5 3 8 9 9 6 4 2 1 1 4 5
4 2	9	4 3 5 1 6 7 7 0 3

Key: Non-Fraternity 2 | 9 = 92

Fraternity 9 | 4 = 94

a. What level of measurement do the data possess?

b. Based upon the stem-and-leaf display, compare the two groups. Think of the several ways in which this can be done.

c. Suppose that 60% is considered a passing score on the exam. What percent of the fraternity students passed the exam? Non-fraternity students?

d. If someone scores 90 or higher on the exam, they will be exempt from taking the next exam. What percent of the fraternity students will be exempt from taking the next exam? Non-fraternity students?

25. Microsoft's consumer PC sales growth for the last 16 quarters are listed in the following table. Examine the data (sales growth, in percentages) and answer the following questions.

PC Sales Growth

Quarter	Sales Growth (%)	Quarter	Sales Growth (%)
1	20	9	20
2	24	10	19
3	22	11	33
4	19	12	37
5	23	13	24
6	27	14	10
7	16	15	0
8	10	16	–4

Source: Citi Investment Research and Analysis IDC, Company Reports, May 2011

a. Construct an ordered array of the data in rank order.

b. What conclusions can you make based on the ordered array?

26. *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?

27. Construct a dot plot for the following set of data.

23	19	15	20	17
16	18	14	23	22
19	23	19	16	25
17	20	21	23	24

28. Listed in the following table is the number of passing attempts per game by Super Bowl champion Aaron Rodgers in the 2010 NFL season. Construct a dot plot for the data set.

Passing Attempts by Aaron Rodgers

31	29	45	17	46
33	34	34	34	31
35	30	11	37	28

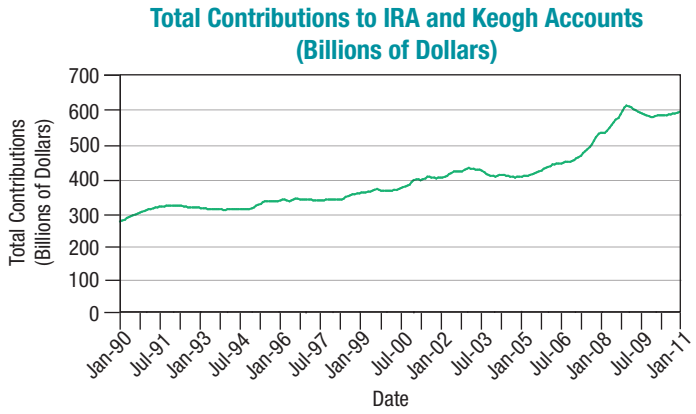
Source: ESPN

29. The following table contains the average monthly energy consumption (in kilowatt-hours) by household for nine South Atlantic states in 2007. Construct a dot plot for the data set.

Monthly Energy Consumption (Kilowatt-Hours)	
773	1143
960	1210
1163	1207
1171	1138
1086	

Source: U.S. Energy Information Administration

30. The following line graph displays the total IRA and Keogh accounts (in billions of dollars) in the U.S., charted from June 1990 to June 2011.



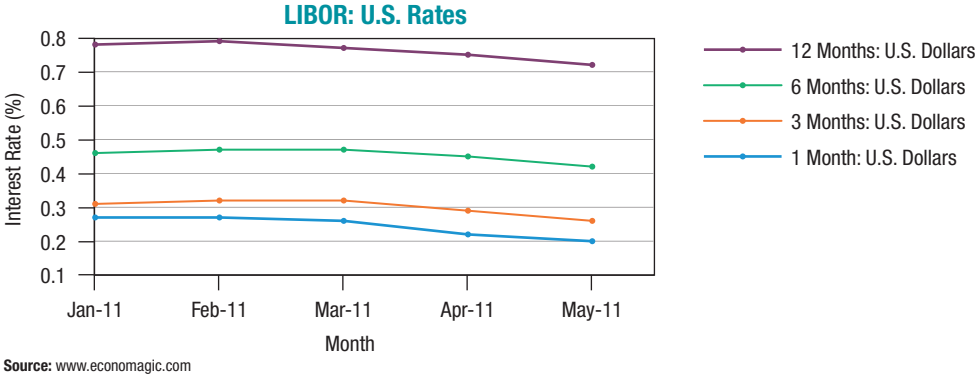
Source: www.economagic.com

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

b. Are the data time series data?

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

31. The following chart contains LIBOR (which stands for London Interbank Offered Rate) data for January 2011 through May 2011. LIBOR is the average interest rate that banks in London charge when lending funds to other banks. The line graphs in the figure represent 1-month, 3-month, 6-month, and 12-month interest rates.



Source: www.economagic.com

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

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

32. The Gallup Poll frequently obtains responses to the question, *At the present time, do you think religion as a whole is increasing its influence on American life or losing its influence?* The percentage of the respondents who answered "increasing" is given below for various polls.

Survey Responses										
Year	2001	1995	1992	1991	1990	1988	1986	1984	1982	1980
Percent	71	38	27	27	33	36	48	42	41	35
Year	1978	1977	1975	1974	1970	1969	1968	1965	1962	1957
Percent	37	37	39	31	14	14	19	33	45	69

a. What level of measurement do the responses to the question possess?

b. Construct a time series plot for the data.

c. What conclusions can you make from the plot?

33. The following table gives the number of immigrants (in thousands) and the average annual immigration rate per 1000 people in the U.S. population for the decade ending in the year given.

Annual Immigration (per 1000 People)											
Year	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
Number	3688	8795	5736	4107	528	1035	2515	3322	4493	7338	9095
Rate	5.3	10.4	5.7	3.5	0.4	0.7	1.5	1.7	2.1	2.9	3.2

a. What levels of measurement do the three variables in this exercise possess?

b. Construct a time series plot of the number of immigrants per decade.

c. Find the percent change in the number of immigrants from the decade ending in 1900 to the decade ending in 2000.

d. Find the percent change in the average annual immigration rate per 1000 people in the U.S. population from the decade ending in 1900 to the decade ending in 2000. Compare your answer to that which you obtained in part **c.** Can you explain why these answers are different?

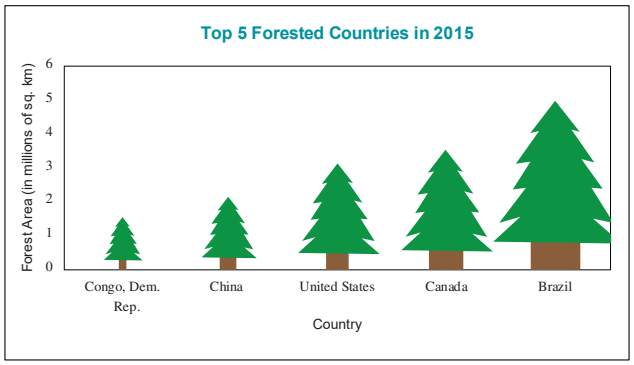
3.5 Analyzing Graphs

Graphs that help us visualize data can either be enlightening, in the sense that they gives 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 3.5.1 shown below, 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

Incorrect Pictograph



Correct Pictograph

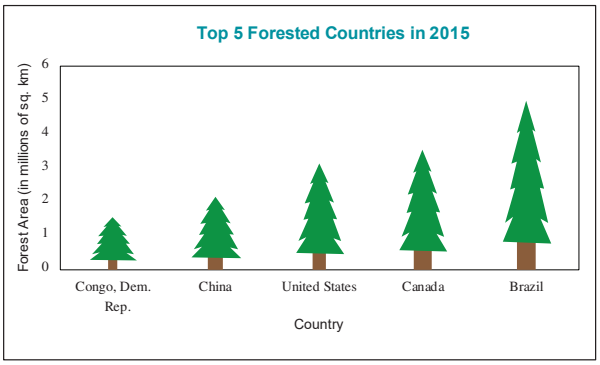


Figure 3.5.6

3.5 Exercises

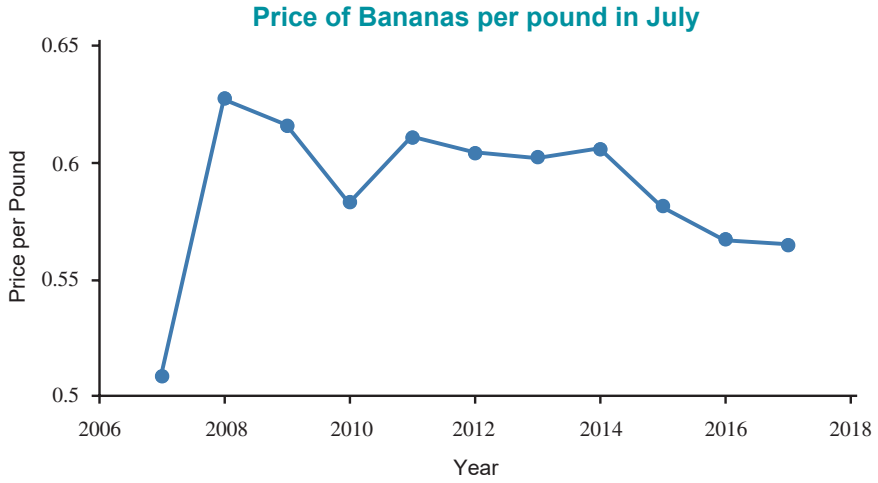
Basic Concepts

1. Why is it important to label and title graphs properly?
2. What types of sources are 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? Why or why not?

   Source: https://data.bls.gov/cgi-bin/surveymost



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
Discovering Business Statistics, Second Edition > Data Sets > San Francisco Salaries 2014

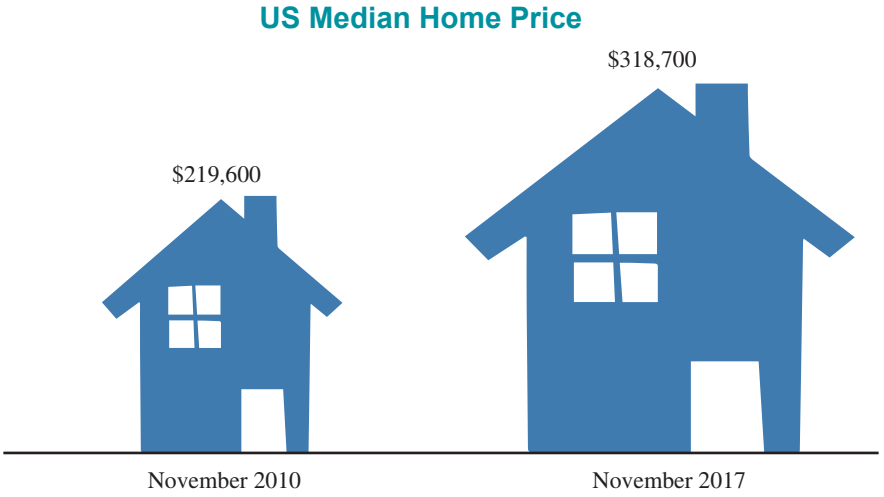
7. The US median home price increased from \$219,600 in November 2010 to \$318,700 in November 2017, as shown in the following pictograph.

Source: U.S. Census Bureau

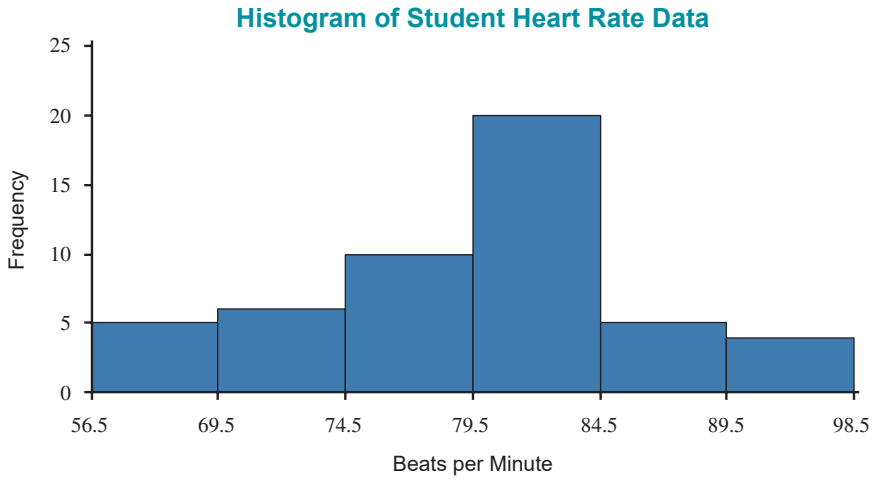
a. What was the percentage increase in US median home price between November 2010 and November 2017?

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 uses the heart rate data from Example 3.3.1 but has different classes than were used in the example. What errors can you find in the histogram?



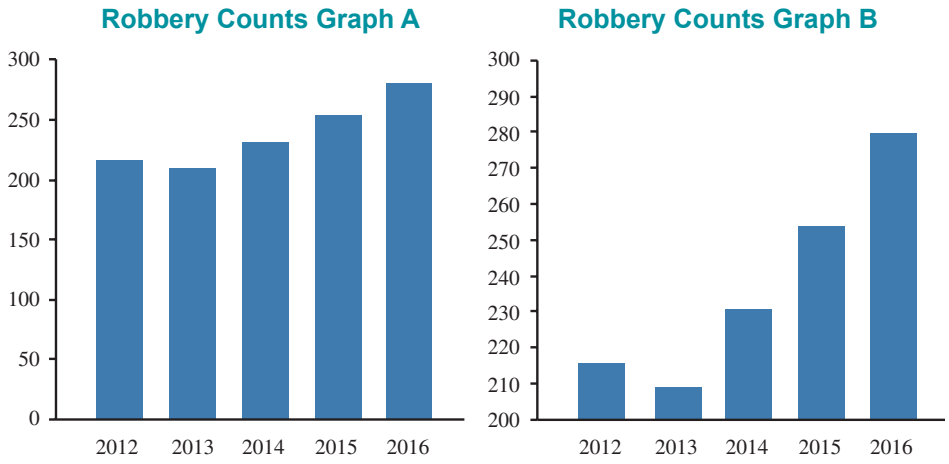
9. The number of robberies in North Charleston, SC is depicted in two different graphs below. Use these graphs to answer the following.

Source: northcharleston.org

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

b. If you lived in North Charleston, how concerned would each of these graphs make you feel? Explain.

c. Approximately how many times taller is the 2016 bar compared to the 2013 bar in Graph B? How many times more robberies were there actually in 2016 compared to 2013?



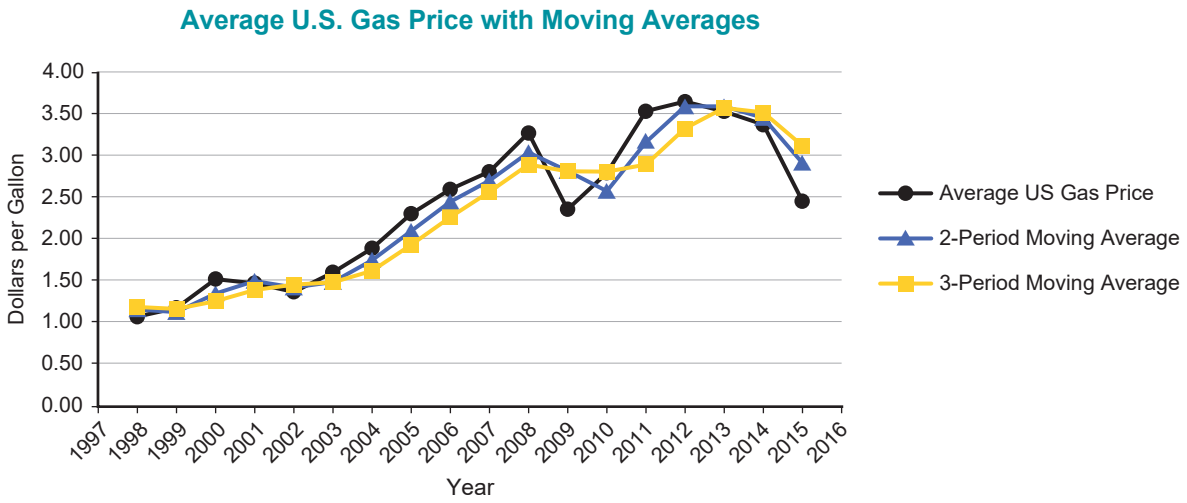


Figure 4.1.10

4.1 Exercises

Basic Concepts

1. Describe the difference between statistics and parameters.
2. Discuss 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. List the types of data that are appropriate for each of the measures of location discussed in the previous question.
6. Why is the mean a measure of central tendency?
7. What is a resistant measure?
8. Describe a situation in which using the weighted mean as a measure of location would be appropriate.
9. What does it mean if we say that a data set is positively skewed? Negatively skewed?
10. Explain why the mean should not be calculated for a nonstationary time series.
11. What is a moving average? When is it useful?

Exercises

12. The data in the table below represent the percentage growth of assets 20 years after the initial investment. Calculate the mean, median, 10% trimmed mean, and mode for percentage growth.

Percentage Growth of Assets									
90.25	93.83	91.41	92.27	90.89	99.12	92.88	97.74	96.28	95.33
91.16	94.30	95.51	92.27	97.63	95.94	90.95	94.76	92.27	92.88

13. A survey was taken of customers asking what percentage above wholesale price would they be willing to pay for a product considered to be a necessity. Calculate the mean, median, 20% trimmed mean, and mode for the percentage above wholesale price that the randomly selected customers are willing to pay.

Percentage Above Wholesale Price							
19	14	11	11	18	20	10	15
20	10	19	11	18	18	11	

14. Calculate the mean, median, 10% trimmed mean, and mode for the following data on the number of cars in line at noon at a favorite fast-food restaurant on 10 consecutive days.

Number of Cars in Line									
2	22	6	18	10	14	12	12	16	8

15. 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 containers that his 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 of 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?

16. 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 manufacturer be more interested in the mean, median, or mode of the ratings?

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?

17. The following table contains the daily high temperatures for a southern city in July (measured in degrees Fahrenheit).

High Temperatures in July (°F)									
84	85	84	88	94	100	97	102	97	89
89	90	88	95	91	95	99	93	97	99
90	94	90	88	91	88	106	99	102	85

a. Calculate the mean of the daily high temperatures.

b. Calculate the median of the daily high temperatures.

c. Calculate the mode of the daily high temperatures.

d. Calculate the 10% trimmed mean of the daily high temperatures.

e. Which measure of central tendency do you think best describes the center of the data set? Why?

18. A tour guide informs his group that the "average" temperature at their destination is 60 degrees Fahrenheit. Once they arrive, they discover that the daytime highs are about 120 degrees Fahrenheit and the nighttime lows are about 0 degrees Fahrenheit. Do you feel the tour guide accurately described the temperatures to the group? Discuss.

19. A worker is participating in a test on a new machine. Her daily production, measured in numbers of units, for the twenty-day test is listed in the following table. On days 4 and 5, the worker was ill and went home shortly after coming to work.

Daily Production										
Day	1	2	3	4	5	6	7	8	9	10
Units	100	104	117	20	20	111	105	106	115	101
Day	11	12	13	14	15	16	17	18	19	20
Units	101	102	115	116	113	103	104	119	118	108

a. What level of measurement does the data possess?

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

c. Considering the worker's illness, which measure computed in part **b.** best describes the production capability of the machine? Discuss.

20. Consider the following per capita greenhouse emissions (in tons of carbon dioxide equivalent per capita) for 10 randomly selected states.

Greenhouse Emissions per Capita (Tons)				
11.76	15.65	22.93	24.75	21.22
18.72	22.55	27.99	12.23	114.40

a. What level of measurement do the data possess?

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

c. Considering the data, which measure computed in part **b.** best describes the per capita greenhouse emissions? Discuss.

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

Clothing Store 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. Calculate the two-period moving averages for the data.

c. Calculate 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. Late in the summer of 1996, Tiger Woods became a professional golfer. This highly publicized event followed a sensational college career at Stanford University, where Tiger won three United States Amateur championships. Tiger was not a professional very long before he had his first win on the pro tour, the Las Vegas Invitational. He received a total of $297,000 for his accomplishment. Since becoming a professional, Tiger has won more than 82 times and has surmassed a net worth of more than $1 billion. The table below contains the prize money (in millions, US dollars) that Tiger has won on the golf course each year from 1996 through 2016.

Career Earnings of Tiger Woods from 1996 to 2016 (in Million U.S. Dollars)			
Year	**On Course**	**Year**	**On Course**
1996	0.89	2007	22.9
1997	2.38	2008	7.74
1998	2.93	2009	21.02
1999	7.68	2010	2.29
2000	11.03	2011	2.07
2001	7.77	2012	9.12
2002	8.29	2013	12.09
2003	6.7	2014	0.61
2004	6.37	2015	0.55
2005	11.99	2016	0.11
2006	11.94		

a. Find the mean.

b. Find the median.

c. Find the mode.

d. Find the 10% trimmed mean and compare it to the mean and the median.

e. Comment on the skewness of the distribution.

Two standard deviations above the mean is

$$\mu + 2\sigma = 7498 + 2(3639) = \$14{,}776$$

and two standard deviations below the mean is

$$\mu - 2\sigma = 7498 - 2(3639) = \$220.$$

Therefore, by Chebyshev's Theorem, we can say that at least 75% of the tuition and fees of colleges and universities in the United States is between \$220 and \$14,776 for 2019–2020.

Who Is King of the Hill?

In 1961, Wilt Chamberlain was the National Basketball Association (NBA) rebounding leader with 27 rebounds per game. In 1992, the colorful Dennis Rodman won the same honor with 18.7 rebounds per game. Common sense suggests that professional basketball in the 1990s is played at a much higher level than in the 1960s. So why has the rebounds per game for the rebounding leader fallen? Is it another case of "less is more"?

Researchers investigating this interesting puzzle considered two other variables: the number of rebounding opportunities (this had gone down since the field goal percentage has increased historically) and the average number of minutes played per game, which has also fallen.

Thus, when we adjust the actual rebounds obtained by the rebounding leaders to the number of minutes played and the total number of rebounding opportunities, we see a completely different picture. The adjusted rebound numbers for Chamberlain and Rodman are 35.42 and 51.06 respectively.

The Coefficient of Variation

Sometimes a data analyst wants to compare the variation of two or more data sets. The **coefficient of variation** is a unit-free statistical measure that enables the comparison of the variation in two or more data sets.

Formula

Coefficient of Variation

The coefficient of variation, another statistical measure, compares the variation in data sets.

For population data, the coefficient of variation is defined as

$$\text{CV} = \left(\frac{\sigma}{\mu} \cdot 100\right)\%,$$

and for sample data,

$$\text{CV} = \left(\frac{s}{\bar{x}} \cdot 100\right)\%.$$

When comparing the variation of data sets, many times the units of measure will be different. The coefficient of variation standardizes the variation measure by dividing it by the mean. The division has one interesting side effect: the unit of measure is removed from the statistic.

One of the primary focuses of quality control in manufacturing is the reduction in variation of the output of the process. A bolt manufacturer wants to compare the variability of two bolt manufacturing processes. One process creates bolts with a mean length of 2.5 cm and a standard deviation of 0.2 cm. Is this process more variable than one that produces a bolt that has a mean length of 1 inch and a standard deviation of 0.052 inches?

$$\text{CV}_{\text{Bolt 1}} = \frac{0.2}{2.5} \cdot 100 = 8.0\%$$

$$\text{CV}_{\text{Bolt 2}} = \frac{0.052}{1} \cdot 100 = 5.2\%$$

The coefficient of variation for Bolt 1 is 8.0% . This means that the variation is 8% of the mean value. The coefficient of variation for Bolt 2 is 5.2% of the mean. Therefore, the process used to make Bolt 2 is less variable than that for Bolt 1.

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. What is the empirical rule? When is it appropriate to use the empirical rule?
8. What is Chebyshev's Theorem?
9. Discuss the purpose of the coefficient of variation.
10. How is the coefficient of variation calculated?
11. Why is the coefficient of variation important?

Exercises

12. Find the missing age in the following set of four student ages.

Student Ages

Student	Age	Deviation from the Mean
A	19	–4
B	20	–3
C	?	+1
D	29	+6

13. Find the missing weight in the following data set.

Weights

Person	Weight	Deviation from the Mean
A	144	–20
B	156	–8
C	?	+1
D	176	+12

14. Consider the following time until failure for 10 randomly selected car batteries (measured in years).

Years Until Failure for Car Batteries

5	3	4	6	2	5	7	10	8	4

   a. Calculate the sample variance of the time until failure.
   b. Calculate the sample standard deviation of the time until failure.
   c. Calculate the range of the time until failure.
   d. What are some of the factors which might contribute to the variation in the observations?

15. Consider the following distances jumped (in feet) by 8 randomly selected long jumpers.

Jump Distances (Feet)

21	15	12	18	10	14	17	11

   a. Calculate the sample variance of the distances jumped.
   b. Calculate the sample standard deviation of the distances jumped.
   c. Calculate the range of the distances jumped.
   d. What are some of the factors which might contribute to the variation in the observations?

16. The interest rates on 30-year mortgages offered by seven randomly selected banks in a large metropolitan area are recorded in the following table.

Interest Rates (%)						
7.5	8.0	7.0	7.25	8.5	8.25	7.75

a. Calculate the sample variance of the interest rates.

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

c. Calculate the range of the interest rates.

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

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

Test Scores							
Males	65	100	75	45	85	73	95
Females	75	80	95	85	82	72	49

a. Calculate the average test score for male students and female students separately.

b. Calculate the variance of the test scores for male students and female students separately.

c. Calculate the standard deviation of the test scores for male students and female students separately.

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

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

18. Consider the following market values of two portfolios of stocks at five randomly selected times during a year.

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

b. Calculate the statistics you proposed in part **a.**

c. Which portfolio has the least amount of risk? Why?

19. Add 20 to each of the following data values.

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

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

b. Compare the mean and standard deviation of the adjusted data to the mean and standard deviation of the original data.

c. Describe the effect on the mean and standard deviation of adding a constant to a data set.

20. Adjust the following data values by subtracting 20 from each data value.

745	789	712	764	736
758	722	773	751	741

a. Calculate the mean and variance for the original and adjusted data.

b. Compare the mean and variance of the adjusted data to the mean and the variance of the original data.

c. Describe the effect of subtracting a constant value from each member of a data set on the mean and variance of the data.

21. 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.

22. 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.

23. There is an annual chowder eating contest in a small New England town. The average amount of chowder eaten at the contest was 32 ounces with a variance of 64 ounces. Given that one hundred people participated in the contest, find:

a. The approximate number of people who ate between 24 and 40 ounces of chowder.

b. The approximate number of people who ate between 16 and 48 ounces of chowder.

c. What assumptions did you make about the amount of chowder eaten by each contestant in answering parts **a.** and **b.**?

24. The manager of a local diner has calculated his average daily sales to be $4500 with a standard deviation of $750.

a. In what range can the manager expect his daily sales to be 68% of the time?

b. In what range can the manager expect his daily sales to be 95% of the time?

c. In what range can the manager expect his daily sales to be 99.7% of the time?

d. What assumption did you make about daily sales when answering parts **a.**, **b.**, and **c.**?

25. 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 following table for three of the job grades.

Salaries ($)			
Job Grade	**25**	**33**	**40**
$\bar{x}$	22,000	35,000	45,000
s	1500	2000	5000

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.**?

26. 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. Calculate the coefficient of variation for Brand A.

b. Calculate the coefficient of variation for Brand B.

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

27. A manufacturer of bolts has two different machines. One machine is used to produce $\frac{1}{4}$-inch bolts; the other machine is used to produce $\frac{1}{2}$-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 following table.

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. Calculate the coefficient of variation for Machine X.

b. Calculate the coefficient of variation for Machine Y.

c. Which machine more consistently produces bolts of the correct diameter? Explain.

4.3 Measures of Relative Position

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.

Definition

*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.

Percentiles

The most commonly used measure of relative position is the percentile. In fact, we have already discussed the 50th percentile; it is the median. For example, in data sets that do not contain significant quantities of identical data, the 30th percentile is a value such that about 30 percent of the values are below it, and about 70 percent are above it.

SOLUTION

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

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

On the marketing test you scored 1.20 standard deviations above the mean, compared to only 1.09 standard deviations above the mean for the management test. Even though the raw score on the management test is larger than the raw score on the marketing test, relative to the means of the data sets, the performance on the marketing test was slightly better. Once again, changing the scale of the data has beneficial effects. It enables the comparison of two measurements that are drawn from different populations.

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

4.3 Exercises

Basic Concepts

1. What are two methods for describing relative position?
2. If a data value is calculated to be the 72nd percentile, what does this mean?
3. Describe how to find the percentile of a particular data 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. The following test scores were recorded for an economics final examination.

Test Scores														
60	81	100	44	90	56	71	42	64	100	69	80	90	87	94
41	78	100	50	96	77	61	38	41	68	50	69	85	47	86

   a. Calculate the 20th percentile.
   b. Calculate the 95th percentile.
   c. Interpret the meaning of each of these percentiles.
   d. Determine the percentile rank for the student who scored 56.
   e. Determine the percentile rank for the student who scored 80.

11. Copiers Etc. collects data on the number of copiers sold each day by each salesperson. The number of copiers sold for each salesperson for a small office on a randomly selected day is listed below.

Numbers of Copiers Sold											
1	5	2	3	7	6	1	0	0	3	4	5

a. Calculate the 25th percentile.

b. Calculate the 90th percentile.

c. Interpret the meaning of each of these percentiles.

d. Determine the percentile rank for the salesperson who sold 5 copiers.

e. Determine the percentile rank for the salesperson who sold 1 copier.

12. Subjects in a marketing study were shown a film and at the end of the film were given a test to measure their recall. The scores are listed in the following table.

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

a. Calculate Q_1, the first quartile.

b. Calculate Q_2, the second quartile.

c. Calculate Q_3, the third quartile.

d. Explain the meaning of these quartiles in the context of the marketing study.

e. Calculate the interquartile range.

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

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

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

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

13. A baseball recruiter is interested in 20 perspective players. He goes to several games and determines the batting average for each player. The batting averages are displayed in the following table.

Batting Averages									
.330	.260	.180	.150	.200	.400	.020	.190	.290	.200
.170	.150	.250	.270	.320	.280	.270	.220	.270	.300

a. Calculate Q_1, the first quartile.

b. Calculate Q_2, the second quartile.

c. Calculate Q_3, the third quartile.

d. Explain the meaning of these quartiles in the context of the batting averages.

e. Calculate the interquartile range.

f. Construct a box plot for the batting averages. Are there any outliers? (Guess which player is the pitcher.)

g. Compute the z-score for a batting average of .020.

h. Compute the z-score for a batting average of .330.

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

j. Determine the percentile rank for a player who had a batting average of .270.

k. Determine the percentile rank for a player who had a batting average of .150.

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

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

15. A statistics student scored a 75 on the first exam of the semester and an 82 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		
	First Exam	Second Exam
μ	74	85
σ	10	7

16. A hospital measures babies' heights when they are born in both inches and centimeters. Eight baby girls are randomly selected and the following heights are recorded in both inches and centimeters.

Newborn Heights								
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. Calculate the mean height in inches and centimeters for the baby girls.

b. Calculate the standard deviation of the heights of baby girls in both inches and centimeters.

c. Calculate the z-score for the height of Baby Girl 3 measured in inches.

d. For Baby Girl 3, calculate the z-score for the height measured in centimeters.

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

4.4 Data Subsetting

Data subsetting is used to provide more clarity and structure to the data. Referring to the histogram in Example 4.2.6, we know that the data consists of tuition and fees of two-year and four-year institutions. The histogram below (Figure 4.4.1) depicts the same data in Example 4.2.6 but has more intervals to give us a better spread of the data. Due to the large number of observations between \$3,750 and \$5,000, the histogram appears to be right-skewed. Were the histogram bell-shaped, it would be easier to identify the center of the data. Thus, it is sometimes prudent to separate the tuition data (i.e., data subsetting) into two groups—tuition for two-year institutions and tuition for four-year institutions.

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?

Exercises

Data

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

3. 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. Calculate the coefficient of variation of the ABV values for both the Renegade and Wynkoop breweries. Which brewery has more consistent ABV values?

Data

stat.hawkeslearning.com
Discovering Business Statistics, Second Edition > Data Sets > Mount Pleasant Real Estate Data

4. 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 set 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 are newly built (2015–2017)? Which subdivision has the most new homes?
   f. What is the average price of new homes (2015–2017) in Carolina Park? Round your answer to the nearest whole dollar.
   g. For all new homes (2015–2017) 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?

5. You are asked to evaluate whether there are issues with how funds are distributed to individuals with developmental disabilities in California. There is a concern that expenditures may not be allocated equitably across various demographic groups. Using the California DDS Expenditures data set, answer the following questions:

Data

stat.hawkeslearning.com

Discovering Business Statistics, Second Edition > Data Sets > California DDS Expenditures

- **a.** What are the mean, mode, and median for the variable Expenditures of the entire data set?
- **b.** What are the variance, standard deviation, and range for the variable Expenditures of the entire data set?
- **c.** What are the 1st and 3rd quartile for the variable Expenditures of the entire data set?
- **d.** Which variables could be used to subset the data?
- **e.** Find the average Expenditure for each Age Group
- **f.** Which Age Group has the highest average Expenditure? Do you notice any trends by Age Group? What might account for differences that exist?
- **g.** Which age group has the highest level of dispersion in Expenditure as measured by the standard deviation and coefficient of variation? Do you notice any trends by Age Group? What might account for differences that exist?
- **h.** Find the average Expenditure for each Ethnicity.
- **i.** What proportion of Expenditures is allocated to each Ethnicity?
- **j.** Briefly discuss your findings based on the analysis in the previous sections.

6. You have been hired by a large company to investigate their employees' satisfaction level. There is some concern that there is high turnover with experienced employees. Using the Employee Satisfaction data set, answer the following questions:

Data

stat.hawkeslearning.com

Discovering Business Statistics, Second Edition > Data Sets > Employee Satisfaction

- **a.** What are the mean, mode, and median for the variable Satisfaction Level?
- **b.** What are the variance, standard deviation, and range for the variable Satisfaction Level?
- **c.** What are the 1st and 3rd quartile for the variable Satisfaction Level?
- **d.** Which variables could be used to subset the data?
- **e.** What is the correlation coefficient between employee satisfaction and the employee's last evaluation score? What does this correlation tell you about the relationship?
- **f.** Find the average Satisfaction Level for each Department?
- **g.** Which salary grouping (low, medium, high) has the highest level of dispersion for Satisfaction Level as measured by standard deviation and coefficient of variation?
- **h.** Find the average Satisfaction Level of each year of experience. Are there differences in Satisfaction Level based on years spent at the company?
- **i.** Briefly discuss your findings based on the analysis in the previous sections.

Data

stat.hawkeslearning.com
Discovering Business Statistics, Second Edition > Data Sets > San Francisco Salaries 2014

7. You have been applying for jobs in San Francisco. You want to research to understand what salary level you can expect to be offered. Using the San Francisco Salaries 2014 data set, answer the following questions:

a. What are the mean, mode, and median for the variable Total Pay and Benefits?

b. What are the variance, standard deviation, and range for the variable Total Pay and Benefits?

c. What are the 1st and 3rd quartile for the variable Total Pay and Benefits?

d. Which variables could be used to subset the data?

e. Construct a frequency distribution for Base Pay? Include the relative frequency of each class.

f. From part e. which pay group has the highest relative frequency? What trends do you notice? What might account for differences that exist?

g. Determine the percentage of jobs that have overtime.

h. Which group (overtime or no overtime) has the highest level of dispersion for total pay as measured by standard deviation and coefficient of variation?

i. Briefly discuss your findings based on the analysis in the previous sections.

Assuming the data in Table 4.5.1 are population data, the mean cash on hand for the 45 companies is calculated as follows.

$$\mu = \frac{\Sigma(f_i M_i)}{N} = \frac{1575}{45} = \$35 \text{ million}$$

The variance of the grouped data is calculated as follows.

$$\sigma^2 = \frac{\Sigma(f_i M_i^2) - \frac{(\Sigma(f_i M_i))^2}{N}}{N}$$

$$= \frac{90125 - \frac{1575^2}{45}}{45} \approx 777.7778$$

If the data are sample data, then the variance is:

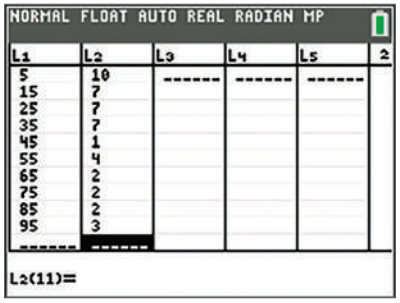
$$s^2 = \frac{\Sigma(f_i M_i^2) - \frac{(\Sigma(f_i M_i))^2}{n}}{n-1}$$

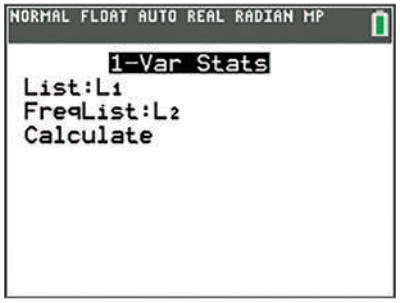
$$= \frac{90125 - \frac{1575^2}{45}}{44} \approx 795.4545.$$

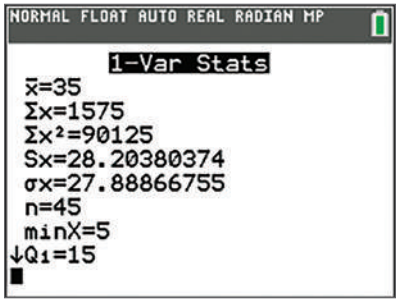
It is important to remember that the calculations of the mean and variance are approximate. That is, if the raw data are available, the actual mean and variance would differ from the measures calculated using the grouped data.

Technology

For technology instructions to calculate the sample statistics for grouped data, like the mean and standard deviation, visit stat.hawkeslearning.com and navigate to **Discovering Business Statistics, Second Edition > Technology Instructions > Descriptive Statistics > Two Variables**.







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. Calculate the mean and 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. 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. Calculate the average credit card rate charged by the top 100 issuers based on the frequency distribution.

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

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

5. A frequency distribution for the Beers and Breweries data set from the companion website is shown below. Use the frequency distribution to perform the following.

Data

stat.hawkeslearning.com
Discovering Business Statistics, Second 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. Calculate the average ABV of all beers based on the frequency distribution. Round your answer to three decimal places.

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

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

4.6 Proportions

Definition

Proportion

A **proportion** measures the fraction of a group that possesses some characteristic.

The **proportion** is one of the more common summary measures.

To calculate a proportion, simply count the number in the group that possess the characteristic and divide the count by the total number in the group. Let

x = number of observations that possess the characteristic,

N = number of observations in the population, and

n = number of observations in the sample, then

4.6 Exercises

Basic Concepts

1. What is a proportion?
2. What is the difference in notation between a population proportion 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. A survey of shoppers at a local mall was taken to study the shopping habits of consumers. The mall contains a variety of specialty stores, especially stores that specialize in electronics and gadgets. One question in particular asked, "Do you enjoy shopping for electronics?" Of the 300 men surveyed, 175 answered "Yes." Of the 200 women surveyed, 55 answered "Yes."
   a. What proportion of men enjoy shopping for electronics?
   b. What proportion of women enjoy shopping for electronics?
   c. What is the overall proportion of consumers at the Tech Mall that enjoy shopping for electronics?
5. A survey released in April 2011 conducted by the consulting firm Booz & Company regarding the automobile industry found that U.S. automotive executives were skeptical about the industry's economic recovery. Suppose that 118 original equipment manufacturers (OEM) and 82 supplier executives participated in the study. When asked whether the overall state of the industry is fairly similar to, or somewhat better than, its low point in January 2009, 59 OEMs and 28 supplier executives answered that the current state of the industry was "about the same," and 55 OEMs and 52 supplier executives answered that the current state of the industry was "somewhat better."

   Source: Booz & Company; 2011

   a. Calculate the sample proportion of original equipment manufacturers that believe the state of the automobile industry is "about the same" as in January 2009.
   b. Calculate the sample proportion of supplier executives that believe the state of the automobile industry is "about the same" as in January 2009.
   c. Calculate the sample proportion of original equipment manufacturers that believe the state of the automobile industry is "somewhat better" than in January 2009.
   d. Calculate the sample proportion of supplier executives that believe the state of the automobile industry is "somewhat better" than in January 2009.
   e. Do these responses seem to support Booz & Company's conclusion that automotive executives are skeptical about the industry's economic recovery? Discuss.

6. An experiment was conducted to study how investors selected mutual funds. Two groups of investors were selected. Group 1 consisted of investors that used online brokerages and Group 2 was made up of investors that used full-service brokerages. Of the 150 investors in Group 1, 120 indicated that they selected mutual funds on their own, while 30 stated that they selected mutual funds using recommendations of the brokerage, family, and friends. Of the 200 investors in Group 2, 25 indicated that they selected their funds on their own, while 175 indicated that they selected the mutual funds using recommendations of the brokerage, family, and friends.

a. What proportion of Group 1 investors selected mutual funds on their own?

b. What proportion of investors in Group 2 selected mutual funds on the recommendation of others?

c. What does this tell you about investors that use online brokerages versus those using full-service brokerages?

7. 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. 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.

Source: CNNMoney.com; 2011

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?

8. What college football conference has the right to brag about putting players in the NFL? A random sample of 100 current NFL players was surveyed to determine the conference in which they played college football. The following table displays the results of the survey.

NFL Players and College Football Conferences

Conference	Number of Players
SEC	20
Big 12	16
Big 10	12
Pac 10	6
ACC	6
MAC	2
Big East	2
Other	36

a. What proportion of players are from the SEC?

b. What proportion of players are from a conference other than the first 6 listed?

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

9. According to a survey administered by the market research group ChangeWave in February 2011, approximately 27% of respondents reported that they plan on buying a tablet device in the future. This result was 2 percentage points higher than in a similar survey administered in November of 2010. In the February study, 3091 customers were surveyed on tablet demand and future buying trends. Suppose that in the November study, 721 people said they planned on buying a tablet device in the future.

Source: InvestorPlace; 2011

a. In the February study, how many people said that they planned on buying a tablet device in the future?

b. In the November study, how many total customers were surveyed?

10. It is no secret that Wall Street firms compete aggressively to lure their clients. Having more high-end clients translates into fees and revenues that turn into profits. A survey of 150 high-end clients asked what lured them to their respective Wall Street firm. The following table shows the results.

High-End Client Response

Perk Received	Client Response
Pay a Kick-Back	25
Lucrative Golf Outings	12
Lavish Dinners	8
Free Private Jet Use	33
Prime Seats at Sports Events	20
Other	22
No Perk Received	30

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

b. What proportion of clients were lured 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.

4.7 Measures of Association Between Two Variables

Oftentimes a manager or decision maker is interested in the relationship between two variables. Such relationships could be the amount of sales and advertising expenditures, education level and income, number of contacts made and sales, sales and earnings, number of real estate foreclosures and home prices, and so on. In earlier sections, we discussed methods used to study each of these measurements individually. Now we want to study the measurements of two different variables at the same time to see if there is a relationship between them. The purpose of studying the relationship between two variables is three-fold: to describe and understand the relationship, to forecast and predict a new observation, and if one is working with a process, not understanding the relationship could be detrimental to making necessary adjustments. There are statistical tools that can aid in the discovery of relationships.

Thinking about relationships is something that everyone does. Why, for example, does an admissions counselor want to know the relationship between SAT scores and college performance? The admissions counselor's task is to select students who will be successful at that college. If college performance is related to SAT scores and the relationship can be specified explicitly, then SAT scores can be used to *predict* performance. Consequently, the

4.7 Exercises

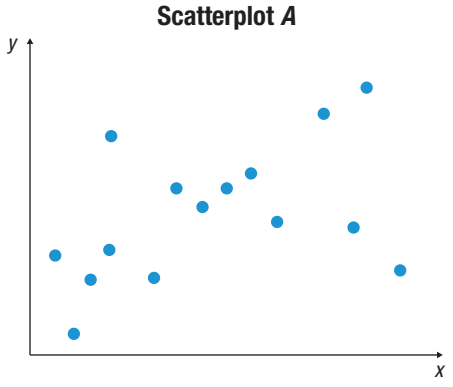
Basic Concepts

1. Give an example of a business 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 are bivariate data? How is bivariate data different from univariate data?
4. What graphical tool is often used in the discovery of relationships?
5. What are four common questions you should ask when studying a graphical representation of bivariate data?
6. If bivariate data exhibit an inverse relationship, what does that mean?
7. How do you construct exact relationships between two variables?
8. In what range is the value of r when bivariate data exhibit 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?

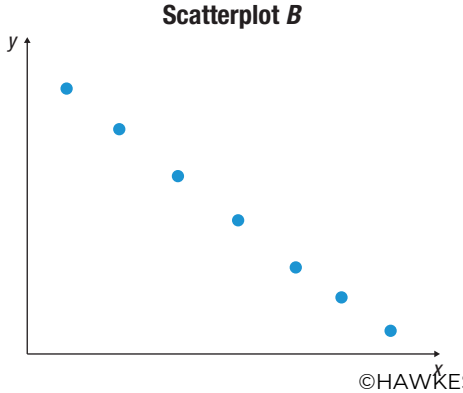
Exercises

11. Consider the following scatterplots and answer the following questions regarding the overall pattern of the data for each of the graphs.
    - Does the pattern roughly follow a straight line?
    - Is the pattern upward sloping or downward sloping?
    - Are the data values tightly clustered in the pattern or widely dispersed?
    - Are there significant deviations from the pattern?

a.



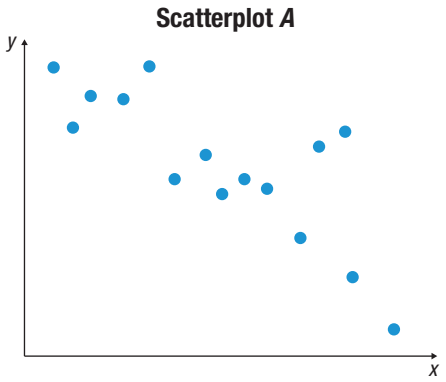
b.



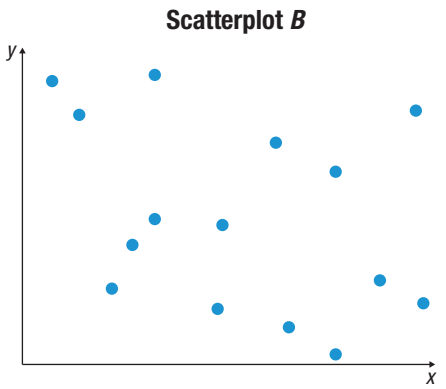
12. Consider the following scatterplots and answer the following questions regarding the overall pattern of the data for each of the graphs.

- Does the pattern roughly follow a straight line?
- Is the pattern upward sloping or downward sloping?
- Are the data values tightly clustered in the pattern or widely dispersed?
- Are there significant deviations from the pattern?

a.



b.



13. A manufacturing company which produces laminate for countertops is interested in studying the relationship between the number of hours of training an employee receives and the number of defects per countertop produced. Ten employees are randomly selected. The number of hours of training which each employee has received is recorded and the number of defects on the most recent countertop produced is determined. The results are as follows.

Employee Training	
Hours of Training	Defects per Countertop
1	1
4	4
7	0
3	3
2	5
2	4
5	3
5	2
1	5
6	1

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

i. Do the variables selected for measurement seem appropriate for answering the question that the manufacturing company is interested in?

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

iii. How are 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 straight line?

ii. Is the pattern upward sloping or downward sloping? Are the data values tightly clustered in the pattern or widely dispersed?

iii. Are there significant deviations from the pattern?

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

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

16. 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. $r = 0.9$ **c.** $r = -0.9$ **e.** $r = 0$

b. $r = 0.5$ **d.** $r = -0.5$

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. $r = 0.8$ **c.** $r = -0.8$ **e.** $r = 0.1$

b. $r = 0.4$ **d.** $r = -0.4$

18. A sample of 10 female swimmers, all 17 years old, is selected from a local swim league. Each swimmer's best time (in seconds) in the 50-yard freestyle and in the 100-yard individual medley are obtained. The 100-yard individual medley consists of swimming 25 yards with each of the four major strokes. The data are given in the following table.

Best Times

Freestyle	27.4	27.0	26.8	30.7	28.5	28.6	29.6	30.8	31.5	29.8
Medley	66.3	66.4	66.7	78.7	69.4	72.0	73.5	81.1	78.6	73.5

a. Construct a scatterplot of the data.

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

c. Compute the correlation coefficient.

19. A personnel director is interested in studying the relationship (if any) between age and salary. Sixteen employees are randomly selected and their ages and salaries are recorded.

Ages and Salaries			
Age	**Salary ($)**	**Age**	**Salary ($)**
25	22,000	49	39,000
55	45,000	37	45,000
27	43,000	62	60,000
30	30,000	40	35,000
22	24,000	35	34,000
33	53,000	29	30,000
19	18,000	58	73,000
45	38,000	52	42,000

a. Plot the data points on a scatterplot.

b. Determine the correlation coefficient.

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

20. The following variables have high positive linear correlations. Is it reasonable to conclude that an increase in one variable causes an increase in the other variable? Explain what could be causing this apparent relationship.

a. Height and vocabulary

b. Absenteeism from school and sale of cough syrup

c. Sale of turkey and sale of toys

21. The following variables have high positive linear correlations. Is it reasonable to 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. The number of wrecks on a local highway and absenteeism from work

Definition

Statistical Inference

The process of making judgments about population parameters is called **statistical inference**.

Lloyd's of London

This very modern looking building is the home of the world's second largest commercial insurer and the sixth largest reinsurance group, Lloyd's of London. At Lloyd's, like all other insurers, risk is measured in probabilities, which are usually 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. Lloyd's annual marine insurance premiums amount to more than $30 billion a year, and that represents little more than one-third of their aggregate income. In addition to sizable revenues, Lloyd's employs about 70,000 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. To become a Name you must have a net worth in excess of $550,000 (excluding the value of your home) and apply to Lloyd's committee for approval. 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.

Criticism of the Subjective View

If science is defined as finding out what is probably true, there should be a probability criterion on which all reasonable persons could agree. But if probability is subjective, how can it be used as a universally accepted criterion? Two reasonable persons might examine the same data and reach different conclusions about their degree of belief about some proposition.

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**. 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. All statistical conclusions must be endowed with a degree of uncertainty. Because probability is used to assess the reliability of sample inferences, it is the foundation 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 criterion in designing and evaluating product reliability, evaluating insurance, 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 of 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.

For example, the next time you watch a 30-second commercial during the Super Bowl, consider the fact that a company has just spent roughly $3 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 uses subjective probability to assess the risk and reward.

5.1 Exercises

Basic Concepts

1. Describe randomness.
2. What is probability?
3. What are the conditions of a random experiment?
4. Consider the random experiment of flipping a fair coin twice. What is the sample space for this experiment?
5. What is an event?
6. Consider the random experiment of rolling a fair die once. Give an example of an event for this experiment and list the outcomes associated with that particular event.
7. What are the two main branches of probability?

8. What are the two approaches to objective probability?

9. What are some of the problems associated with the relative frequency approach?

10. True or false: According to the mathematical law of probability, the observed relative frequency of heads when flipping a coin will eventually reach 0.5 since the probability of heads is 0.5.

11. What is statistical regularity?

12. Describe the classical approach to probability.

13. Using the classical approach, describe how you would determine the probability of event A.

14. What is the subjective approach to probability? Discuss the problems of applying the subjective interpretation.

15. What is statistical inference?

16. Discuss the relationship between probability and statistics.

17. Give three applications of probability in business.

18. Describe the importance of probability in the insurance industry.

19. What type of probability does the manager of a company use when purchasing a commercial spot during the Super Bowl? Explain why.

Exercises

20. Consider the following random experiment. A potato chip manufacturer is interested in determining if the brand of potato chip which it manufactures is preferred over three of its major competitors. Several customers are randomly selected and asked which brand of potato chip they prefer: Brand A, Brand B, Brand C, or Brand D.

a. Determine the sample space for the experiment described.

b. If the manufacturer makes Brand A, list the outcomes in the event M = {customer does not prefer the manufacturer's brand}.

21. 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}.

22. 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 upper-most surface. The results of the experiment are shown in the following table.

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

a. Using the relative frequency approach, what is the probability of observing each side?

b. Which side do you think the gambler will bet on when the die is tossed?

23. Assume there are two red, two yellow, and two blue buttons in a hat. A button is 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

Using the relative frequency approach, what is the probability of drawing each color?

24. Twenty-five insurance agents are randomly selected and asked if they own a handgun. Twenty-two of those surveyed said that they do own a handgun. If an insurance agent is randomly selected, estimate the probability that the agent will own a handgun.

25. Thirty elementary school teachers are randomly selected and asked if they favor standardized testing of elementary school children. Twenty of those surveyed said that they did favor standardized testing of elementary school children. If an elementary school teacher is randomly selected, estimate the probability that the teacher will favor standardized testing for elementary school children.

26. 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, find the probability that the CEO will have a salary in excess of $500,000.

27. Forty emergency calls to which a local police department responded were randomly selected. Of the forty emergency calls fifteen were categorized as domestic arguments. Estimate the probability that the next emergency call to which the local police department responds will be a domestic argument.

28. 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.

29. 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.

30. A couple plans to have two children.

a. List all possible outcomes for the sexes of the two children.

b. Find the probability that the couple will have 2 boys.

c. Find the probability that the couple will have at least 1 girl.

31. 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.

32. 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 color 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.

5.2 Laws of Probability

Interpreting probability using the classical approach is a good way of thinking about the basic probability principles. In this section we will discuss certain laws that probabilities must obey, regardless of how probability is defined.

Probability Law 1

A probability of zero means the event cannot happen.

For example, the probability of observing three heads in two tosses of a coin is zero.

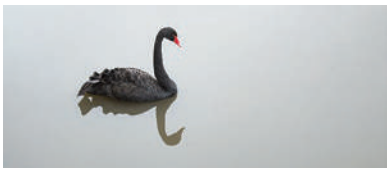
Probability Law 2

A probability of one means the event must happen.

For example, if we toss a coin, the probability of getting either a head or tail is one.

Probability Law 3

All probabilities must be between zero and one, inclusively. That is, $0 \le P(A) \le 1$.



Black Swan Events

Black Swan events are unexpected extreme events. The term stems from 16th century London: at this time, all known swans in the Euro-centric world were white. Subsequently, upon colonization of Western Australia, black swans were unexpectedly discovered. The term was popularized in Nassim Nicholas Taleb's book *The Black Swan: The Impact of the Highly Improbable* (2007). While the discovery of black swans did not adversely impact society, the black swan term today carries the connotations that the event is damaging, unexpected, and in hindsight, quite explainable. Two events occurring since 2000 that arguably qualify for black swan status are 9/11 and the disappearance of the MH-370 aircraft.

Statisticians quantify how rare events are via return periods. For example, if a 50-year earthquake at a fixed location has Richter magnitude 7.0, then the probability that a Richter magnitude 7.0 or greater earthquake occurs at the location over one year is roughly 1 / 50. Statisticians have a sub-discipline called extreme value theory that contains justifiable methods to estimate return periods (see Coles, 2001; *An Introduction to Statistical Modeling of Extreme Values*). This said, the field is often controversial and data void. Imagine trying to estimate a 200 year earthquake from only 50 years of data — a 200-year earthquake event is probably not contained in the data record!

While extreme value statisticians seldom refer to black swan events, the term is common in financial and insurance settings today. There, it often simply serves as a reminder that unexpected rare events do happen and are difficult to quantify.

Courtesy of Robert Lund

We want to know the probability of event A or event B or both. Thus, the desired probability is the union of the two events given by

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.2 + 0.3 - 0.08 = 0.42.$$

Therefore, the probability of finding someone who is earning over \$50,000 or subscribes to more than one sports magazine, or both, is 0.42.

5.2 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(receiving an A on the test) = 0

b. P(receiving an A on the test) = 1

c. P(receiving an A on the test) = 0.3

3. What is a compound event?

4. Define the following set operations: union, intersection, and complement.

5. If you know the probability of two events, what else must you know in order to calculate the probability of *one event or the other*?

6. If events A and B are mutually exclusive, what is $P(A \cap B)$?

Exercises

7. 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

8. 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

9. 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. $P(\text{event}) = -1.0$

10. 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 having a boy or a girl in a single birth.

d. The probability of rolling a two and a five in a single toss of a die.

11. 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	25	30	35
Number Alive	100,000	99,090	98,912	98,815	98,716	98,477	98,204	97,897	97,500
Year	40	45	50	55	60	65	70	75	80
Number Alive	96,958	96,097	94,766	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?

12. 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.

13. 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 ($)			
		0 – 50,000	50,000 – 100,000	100,000 – 150,000	More than 150,000
Amount of Life Insurance on Wife ($)	0 – 50,000	400	200	50	50
	50,000 – 100,000	50	50	30	30
	100,000 – 150,000	20	10	25	25
	More than 150,000	20	10	15	15

a. For a randomly selected policy, what is the probability that the husband will have between $50,000 and $100,000 of insurance?

b. For a randomly selected policy, what is the probability that the wife will have between $100,000 and $150,000 of insurance?

c. For a randomly selected policy, what is the probability that the wife will have more than $150,000 of insurance or the husband will have more than $150,000 of insurance?

d. For a randomly selected policy, what is the probability that the wife will have between $0 and $50,000 of insurance and the husband will have between $0 and $50,000 of insurance?

e. For a randomly selected policy, what is the probability that the wife will not have between $0 and $50,000 of insurance?

f. For a randomly selected policy, what is the probability that the husband will have more than $50,000 of insurance?

g. What approach to probability did you use to calculate your answers?

h. Are the events {the wife has more than $150,000 in insurance} and {the husband has between $50,000 and $100,000 of insurance} mutually exclusive? Explain.

5.3 Conditional Probability

Researchers often want to examine a limited portion of the sample space. For example, consider the question of whether cigarette smoking harms those that are indirectly exposed to the smoke. Suppose that 3 percent of women who do not smoke die of cancer. However, if a nonsmoking woman is married to a smoking husband (not to be confused with a husband who is on fire), the probability of dying of cancer is 0.08. This probability is a **conditional probability**, because the sample space is being limited by some condition—in this case, limited to only wives of smoking husbands. In this instance, the dramatic effect of a smoking husband on cancer rates is readily evident.

Definition

Conditional Probability

The probability that one event will occur given that some other event has occurred is a **conditional probability**.

$$P(\text{a nonsmoking woman dies of cancer}) \neq P(\text{a nonsmoking woman dies of cancer given that her husband smokes})$$

Similarly, the results from a market survey indicate that 39 percent of the customers surveyed believe a product is of high quality. However, if the analysis is limited to only women, 54

Then the desired probability can be formulated as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

The $P(A \cap B)$ is called a joint probability since it is the probability of the occurrence of more than one event. To compute $P(A \cap B)$ use the empirical approach.

$$P(A \cap B) = \frac{193}{1403} \approx 0.1376$$

Similarly, $P(B)$ can be computed as

$$P(B) = \frac{444}{1403} \approx 0.3165.$$

Consequently, $P(A \mid B)$ is

$$P(A \mid B) \approx \frac{0.1376}{0.3165} \approx 0.4348.$$

Note that this answer could have also been obtained by simply dividing 193 by 444.

Let's Make a Deal

A long time ago, back in the 70s, there was a television show called "Let's Make a Deal" starring Monty Hall as the host. This show produced an interesting problem in probability which someone submitted to Marilyn vos Savant which she answered in her column in Parade magazine. Incidentally, Ms. Savant is in the Guinness Book of World Records as having the highest recorded IQ (228). Here's the problem that was posed to Ms. Savant.

"Suppose you're on a game show, and you're given a choice of three doors: Behind one door is a car: behind the others, goats. You pick a door, say number 1, and the host, who knows what's behind the other doors, opens another door, say number 3, which has a goat. He then says to you, 'Do you want to pick door number 2?' Is it to your advantage to take the switch?"

Marilyn vos Savant answered the question in her column saying that it was to your advantage to switch. This set off a firestorm of mail telling Ms. Savant that she was incorrect. Much of this mail came from people with Ph.D.s behind their names. The New York Times printed a front page article in 1991 discussing the problem.

What do you think? To find the answer to this problem type "The Monty Hall problem" in your search engine and go to some of the web sites and try some of the simulations.

5.3 Exercises

Basic Concepts

1. Define conditional probability.
2. How do you calculate $P(A \mid B)$?

Exercises

3. The following table was given in Section 5.2, Exercise 12.

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?

4. The following table was given in Section 5.2, Exercise 13.

Life Insurance Coverage

		Amount of Life Insurance on Husband ($)			
		0 – 50,000	50,000 – 100,000	100,000 – 150,000	More than 150,000
Amount of Life Insurance on Wife ($)	0 – 50,000	400	200	50	50
	50,000 – 100,000	50	50	30	30
	100,000 – 150,000	20	10	25	25
	More than 150,000	20	10	15	15

a. Given the wife has between \$100,000 and \$150,000 of insurance, what is the probability that the husband has more than \$150,000 of insurance?

b. Given the wife has between \$0 and \$50,000 of insurance, what is the probability that the husband has between \$0 and \$150,000 of insurance?

c. Given that the husband has between \$0 and \$50,000 of insurance, what is the probability that the wife will have more than \$150,000 of insurance?

d. Given that the husband has more than \$150,000 of insurance, what is the probability that the wife will have more than \$150,000 of insurance?

5. 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?

6. 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.

7. 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?

5.4 Independence

An extremely important concept in statistical analysis is **independence**. It describes a special kind of relationship between two events. Two events are said to be independent if knowledge of one event does not provide information of the other event's occurrence. In other words, the occurrence of one event does not affect the occurrence of another event if the events are independent.

Example 5.4.1

Determining the Independence of Events

Experiment: roll a fair die two times. Consider the two events

A = {rolling a six on the first roll of a fair die} and

B = {rolling a four on the second roll of a fair die}.

Are these two events independent?

SOLUTION

Since knowledge of the outcome of the first roll does not help one make an inference of the outcome of the second roll, events A and B are independent.

5.4 Exercises

Basic Concepts

1. Explain the difference between dependent and independent events.
2. Are mutually exclusive events dependent or independent? Explain your answer.
3. If events A and B are independent, what is $P(A \mid B)$ equal to?
4. What is the product rule?
5. 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?

Exercises

6. The following table was given in Section 5.2, Exercise 12.

Health Care Consumers		
Have Health Insurance Coverage	Housing Situation	
	Rent	Own
Yes	196	298
No	92	173

Are the events {customer rents their home} and {customer owns their home} independent? Explain.

7. The following table was given in Section 5.2, Exercise 13.

Life Insurance Coverage					
		Amount of Life Insurance on Husband ($)			
		0 – 50,000	50,000 – 100,000	100,000 – 150,000	More than 150,000
Amount of Life Insurance on Wife ($)	0 – 50,000	400	200	50	50
	50,000 – 100,000	50	50	30	30
	100,000 – 150,000	20	10	25	25
	More than 150,000	20	10	15	15

Are the events {the husband has more than $150,000 in insurance} and {the wife has more than $50,000 in insurance} independent? Explain.

8. 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?
9. 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?
10. 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?

11. 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?

12. 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 simple events 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.

13. 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 he or she would fail both tests?

b. If a person were a drug user, what is the probability he or she would pass both tests?

5.5 Bayes' Theorem

We have completed the discussions about conditional probability and independent events in Section 5.3 and Section 5.4. **Bayes' theorem** (also referred to as **Bayes' rule** or **Bayes' law**) is somewhat of an extension of conditional probability in which we calculate probabilities based on new information. Please note and understand that the additional information is obtained for a subsequent event, and the new information is used to revise the initial probability. Recall the following formulas for conditional probability.

From the items calculated in parts **a.** and **b.**, we know

$$P(P \mid B) = \frac{P(P \cap B)}{P(B)} = \frac{P(B \mid P)P(P)}{P(B)}$$
$$= \frac{P(B \mid P)P(P)}{P(B \mid M)P(M) + P(B \mid P)P(P) + P(B \mid C)P(C)}$$
$$= \frac{(0.6)(0.3)}{0.57}$$
$$\approx 0.3158.$$

Thus, we know that if the passenger is traveling on business, there is about a 32% chance that he or she will be traveling by private plane.

Even though it was fairly subtle (given that we performed the calculations in parts **a.** and **b.**), please note the use of Bayes' theorem in the previous calculation.

5.5 Exercises

Basic Concepts

1. Briefly explain the relationship between conditional probability and Bayes' theorem.
2. Other than conditional probability, which other rule which you previously studied is used in the derivation of Bayes' theorem?
3. What is Bayes' theorem?
4. How is Bayes' theorem used to "revise" a probability based on additional information?

Exercises

5. 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?
6. Adults (18 years and older) and kids (under 13 years of age) are observed to react differently to sad, emotional movies. It has been observed that 70% of the kids say they cry at some point during those types of movies, whereas only 40% of the adults admit to crying during those types of movies. A group of 40 people, of whom 25 are kids, was shown a sad, emotional movie and the subjects were asked if they cried. A response picked at random from the 40 indicated that they cried. What is the probability that it was an adult?
7. As items come to the end of a production line, an inspector chooses which items are to go through a complete inspection. Eight percent of all items produced are defective. Sixty percent of all defective items go through a complete inspection, and 20% of all good items go through a complete inspection. Given that an item is completely inspected, what is the probability that it is defective?

8. Two teaching methods for a business statistics class, online and face-to-face, are available during the course of an academic year. The failure rate (students that receive below a C– and thus, will have to repeat the course) is 4% for the online class and 8% for the face-to-face class. However, the online class is more expensive and hence is offered only 25% of the time. (The face-to-face class is offered the other 75% of the time.) A student takes the statistics class via one of the methods of delivery but failed the course. What is the probability that the student took the online class?

9. A personnel director has two lists of applicants for jobs. List 1 contains names of 15 women and 5 men whereas List 2 contains the names of 5 women and 12 men. A name is randomly selected from List 1 and added to List 2. A name is then randomly selected from the augmented List 2. Given that the name selected is that of a man, what is the probability that a woman's name was originally selected from List 1?

5.6 Counting Techniques

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 our counting.

The Fundamental Counting Principle

Theorem

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.

Example 5.6.1

Using the Fundamental Counting Principle to Count Employees

A local bank has three branches. Each branch has four departments and each department has two employees. How many employees does the bank have?

SOLUTION

$$\underset{\text{(number of branches)}}{3} \cdot \underset{\text{(departments)}}{4} \cdot \underset{\text{(employees per department)}}{2} = \underset{\text{(total number of employees)}}{24}$$

Thus, the bank has 24 total employees.

Example 5.6.2

Using the Fundamental Counting Principle to Count License Plates

Nonpersonalized license plates in the state of Utah consist of three numbers followed by three letters (excluding I, O, and Q). How many license plates are possible?

SOLUTION

There are ten digits (0 – 9) possible for each of the first three characters. Likewise, there are 23 letters possible for the last three characters. Therefore, we have the following.

$$\underset{\text{(digit)}}{10} \cdot \underset{\text{(digit)}}{10} \cdot \underset{\text{(digit)}}{10} \cdot \underset{\text{(letter)}}{23} \cdot \underset{\text{(letter)}}{23} \cdot \underset{\text{(letter)}}{23} = \underset{\text{possible license plates}}{12,167,000}$$

Example 5.6.8

Finding the Number of Distinguishable Permutations

How many distinguishable permutations can be made from the word *Mississippi*?

SOLUTION

There are 11 letters in the word *Mississippi*, one M, four I's, four S's, and two P's. So, there are

$$\frac{11!}{(1!)(4!)(4!)(2!)} = 34,650$$

distinguishable permutations of the letters in *Mississippi*.

It is important to remember that combinations are used when order is not important, and permutations are used when order is important.

Concept		Formula
Fundamental Counting Principle	If one event has n_1 outcomes and another event has n_2 outcomes, the number of ways the event can occur in sequence is the product of n_1 and n_2.	$n_1 \cdot n_2$
Factorial	$n!$ is the product of each of the positive whole numbers from 1 to n.	$n! = (n-1)(n-2)\cdots(3)(2)(1)$
Combination	A collection or grouping of objects where order is not important.	The number of combinations of n unique objects taken k at a time is given by the following formula. $${}_nC_k = \frac{n!}{(n-k)!k!}$$
Permutation	A specific order or arrangement of objects.	The number of arrangements for n objects when order is important is given by n factorial. $n!$
	The number of permutations of n unique objects taken k at a time.	$${}_nP_k = \frac{n!}{(n-k)!}$$
	Given n objects with n_1 alike, n_2 alike, …, n_k alike, then the number of distinguishable permutations is given by the formula to the right.	$$\frac{n!}{(n_1!)(n_2!)(n_3!)\cdots(n_k!)}$$

5.6 Exercises

Basic Concepts

1. What is the fundamental counting principle?
2. What is a factorial?
3. Describe the difference between permutations and combinations.
4. Give an example of a situation in which you would need to calculate the number of distinguishable permutations.

Exercises

5. The blue plate lunch at a local cafeteria consists of an entrée, a side item, and a desert. 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. Compute each of the following.

a. 2!

b. 4!

c. 6!

d. 8!

9. A DJ needs to select 6 songs from a CD containing 12 songs to compose an event's musical lineup. How many different lineups are possible?

10. In how many ways can 11 kids be picked for the 9 positions on a baseball team?

11. How many distinguishable permutations can be made from the word STATISTICS?

12. How many distinguishable permutations can be made from the word SASSAFRAS?

13. A person tosses a coin 11 times. In how many ways can he get 9 heads?

14. How many 5 card hands can be dealt from a deck of 52 cards?

6.1 Exercises

Basic Concepts

1. What is a random variable?

2. What is a probability distribution?

3. Do all random variables have easily determined probability distributions? Explain.

4. What are the two types of random variables discussed in the chapter? What distinguishes the two types?

Exercises

5. 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.

6. 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 reasoning test.

c. The number of pizzas delivered on a college campus each day.

d. The daily takeoffs at Chicago's O'Hare Airport.

e. The high temperatures in Maine and Florida tomorrow.

7. 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.

8. 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.

Table 6.2.10 – Return on Truck Purchase (cont.)

Return (Dollars)	Truck 1	Truck 2
15,000	0.20	0.30
20,000	0.03	0.20
25,000	0.02	0.05

Calculate the expected return and standard deviation for each truck and recommend which truck to purchase.

Table 6.2.11 – Calculations

Return (Dollars)	Truck 1	Truck 2	Truck 1 $xp(x)$	Truck 2 $xp(x)$	Truck 1 $(x-\mu)^2 \cdot p(x)$	Truck 2 $(x-\mu)^2 \cdot p(x)$
−5000	0.02	0.15	−100	−750	4,500,000	36,037,500
0	0.03	0.10	0	0	3,000,000	11,025,000
5000	0.20	0.10	1000	500	5,000,000	3,025,000
10,000	0.50	0.10	5000	1000	0	25,000
15,000	0.20	0.30	3000	4500	5,000,000	6,075,000
20,000	0.03	0.20	600	4000	3,000,000	18,050,000
25,000	0.02	0.05	500	1250	4,500,000	10,512,500
		$E(X) =$	$10,000	$10,500		
				$V(X) = \sigma^2 =$	25,000,000	84,750,000
				$\sqrt{V(X)} = \sigma =$	$5000	$9206

The expected value and standard deviation of purchasing Truck 1 are $10,000 and $5000, respectively. Similarly, the expected value and standard deviation of purchasing Truck 2 are $10,500 and $9206. Given that the risk (standard deviation) of Truck 2 is nearly twice that of Truck 1, it appears that selecting Truck 1 is the best decision, even though the expected return is slightly less.

6.2 Exercises

Basic Concepts

1. Discrete probability distributions always have three characteristics. What are they?
2. What is the value of describing a random variable with a probability distribution?
3. What are three different ways to express possible values of a random variable along with their associated probabilities?
4. How is a probability distribution created?
5. Identify four discrete probability distributions.
6. What is a probability distribution function?
7. Why is the notion of expected value important in the analysis of random phenomena?
8. True or false: the expected value of a random variable is usually one of the possible outcomes of the random variable.
9. Suppose the expected value of a random variable was known to be 6.3. Interpret the meaning of the expected value.

10. Give an example of a situation in which expected value would be useful to compare alternatives.

11. How is the variance (or standard deviation) of a random variable related to risk?

Exercises

12. 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	$P(X = x)$
1	$\frac{1}{3}$
2	$\frac{2}{3}$
3	$\frac{1}{3}$

13. 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	$P(X = x)$
–2	0.25
2	0.50
3	0.25

14. 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	$P(X = x)$
2	0.30
3	–0.50
4	0.50
5	0.70

15. 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	$P(X = x)$
5	0.46
10	0.25
15	0.25

16. 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	$P(X = x)$
–10	0.18
–5	0.39
3	0.08
8	0.35

17. 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	$P(X = x)$
100	–0.10
200	0.50
300	0.50

18. 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.

$$P(X = x) = \frac{x}{16}, \text{ for } x = 1,\ 2,\ 3,\ 4,\ 5$$

19. 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.

$$P(X = x) = \frac{x^2}{30}, \text{ for } x = 1,\ 2,\ 3,\ 4$$

20. Find 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

21. Find 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

22. 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.

Number of Helicopter Transports							
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. Find the average number of times the helicopter is used to transport critical patients each month.

b. Find the variance of the number of times the helicopter is used to transport critical patients.

c. Find the standard deviation of the number of times the helicopter is used to transport critical patients.

d. Find the probability that the helicopter will not be used at all during a month to transport critical patients.

e. Find the probability that the helicopter will be used at least once to transport critical patients.

f. Find the probability that the helicopter will be used at most twice to transport critical patients.

g. Find the probability that the helicopter will be used more than three times to transport critical patients.

23. 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.

Number of Times Examined					
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.

24. 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.

Customers 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?

25. 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.

26. 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. Calculate the expected payoff for each of the investment options.

b. Calculate the standard deviation of the payoff for each of the investment options.

c. Which investment option would you choose? Explain.

27. 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		Cereal B	
Sales	***P*(Sales)**	**Sales**	***P*(Sales)**
$150,000	0.2	$10,000	0.40
$200,000	0.3	$300,000	0.40
$300,000	0.3	$600,000	0.10
$400,000	0.2	$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.

6.3 The Discrete Uniform Distribution

The **discrete uniform distribution** is one of the simplest probability distributions. Each value of the random variable is assigned an identical probability. There are many situations in which the discrete uniform distribution arises. Some common examples are rolling a fair die or flipping a fair coin.

Formula

Discrete Uniform Probability Distribution Function

Mathematically, the **discrete uniform probability distribution function** is given by

$$P(X = x) = \frac{1}{n}$$

where n = the number of values that the random variable may assume.

Example 6.3.1

Determining the Probability Distribution of Throwing a Die

What is the probability distribution for the outcome of the throw of a single six-sided die?

SOLUTION

If the die is fair, then each of the outcomes is equally likely, and thus we have a discrete uniform distribution in which all probabilities equal $\frac{1}{6}$. The probability distribution is given in Table 6.3.1.

distribution. Over time the purchasing agent will undoubtedly revise the distribution as more information is gathered about the company's delivery schedule.

b. The expected value is calculated as follows.

$$\begin{aligned} E(X) &= \Sigma\left[x_i P(x_i)\right] \\ &= (1)\frac{1}{4} + (2)\frac{1}{4} + (3)\frac{1}{4} + (4)\frac{1}{4} \\ &= 0.25 + 0.50 + 0.75 + 1 \\ &= 2.5 \text{ weeks} \end{aligned}$$

The variance is calculated as follows.

$$\begin{aligned} \sigma^2 = V(X) &= \Sigma\left[(x_i - \mu)^2 P(x_i)\right] \\ &= (1-2.5)^2 \frac{1}{4} + (2-2.5)^2 \frac{1}{4} + (3-2.5)^2 \frac{1}{4} + (4-2.5)^2 \frac{1}{4} \\ &= (2.25)\frac{1}{4} + (0.25)\frac{1}{4} + (0.25)\frac{1}{4} + (2.25)\frac{1}{4} \\ &= 1.25 \end{aligned}$$

Therefore, the standard deviation is $\sqrt{1.25} \approx 1.12$ weeks.

Example 6.3.3 illustrates an important principle in the application of the discrete uniform distribution. That is, when there is little or no information concerning the outcome of a random variable, the discrete uniform distribution may be a reasonable initial alternative.

6.3 Exercises

Basic Concepts

1. What is the most significant property of the uniform distribution?
2. What is the discrete uniform probability distribution function?
3. Explain why the uniform distribution is often used when there is little or no information concerning the outcome of a random variable.

Exercises

4. In the casino game of roulette, a wheel is spun and a ball is set in motion, ultimately coming to rest in one of the 38 slots on the wheel. Any slot is as likely as any other to capture the ball. Of the 38 slots, 18 are red, 18 are black, and 2 are green. Suppose the entry fee to play a single game is \$1 and the participant bets on red. If the ball comes to rest in one of the red slots, he wins \$1 in addition to getting back the original \$1 entry fee. If the ball does not end up in a red slot, the \$1 entry fee is lost. Let X denote the monetary gain when betting \$1 on red, in a single game of roulette. Gain is defined as the amount won minus the fee to play.
   - **a.** What are the possible values of X?
   - **b.** Is X a discrete or continuous random variable? Explain.
   - **c.** Construct the probability distribution of X.
   - **d.** Find the expected value of X and interpret this number.
   - **e.** Do you feel that in any casino games you would have a positive expected gain? Why?

5. An experiment consists of tossing two coins and a die simultaneously.

a. List the 24 equally likely outcomes.

b. Define the random variable X as the sum of the number of heads on the two coins and the number of dots on the die. What are the possible values of X?

c. Construct the probability distribution of X in the form of a table.

d. Find the expected value of X.

6. A classmate walks into class and states that he has an extra ticket to a rock concert on Friday night. He asks everyone in the class 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 there are 16 people in class that night, what is your chance of being chosen to attend the concert?

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. Sharlene's uncle is one of the builders contracted by the subdivision manager. What is the probability that Sharlene's uncle will be the builder that builds her house?

8. You order some clothing online and get an estimated delivery date of June 6–June 11. You know you will be out of town June 8th and 9th and are a little concerned about the package arriving when you are away. Assuming the delivery date follows a discrete uniform distribution, what is the likelihood your package will be delivered while you are out of town?

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
$P(X = x)$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

6.4 The Binomial Distribution

The binomial distribution arises from experiments with repeated two-outcome trials, where only one of the outcomes is counted. Experiments of this kind are rather common in the business world. In market research, a survey respondent (a trial) either will or will not recognize a company's brand. The number that recognize the brand is a count that may be modeled as a **binomial random variable**. When a customer (a trial) enters a bank for service, he or she may have to wait. If we are counting customers who have to wait, then the count may conform to the binomial model. Experiments are required to meet several conditions in order to qualify as a binomial experiment.

Example 6.4.4

Calculating the Expected Value and Variance of a Binomial Random Variable

Compute the expected value and the variance of the number of customers that will approve the change in return policy in Example 6.4.3.

SOLUTION

Since the random variable is binomial, we can use the shortcuts $E(X) = np$ and $V(X) = np(1 - p)$. Since $n = 10$ and $p = 0.7$, the expected value is given by the following expression.

$$E(X) = np = 10(0.7) = 7$$

The variance is

$$\sigma^2 = V(X) = np(1 - p) = 10(0.7)(0.3) = 2.1,$$

which implies that the standard deviation is $\sqrt{2.1} \approx 1.4491$.

Thus, if 10 randomly selected customers are polled, we would expect 7 of the 10 to approve the change in return policy, and the standard deviation would be 1.4491 customers.

6.4 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 in a business context.
4. What is the binomial probability distribution function?
5. Describe the shape of a binomial distribution. Does the shape change? What influences the shape of the distribution?
6. How do you calculate the expected value of a binomial random variable? The variance? The standard deviation?

Exercises

7. Calculate ${}_nC_x$ 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. Calculate ${}_nC_x$ 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 airlines decide 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 strike at any one plant and the risk of a strike at one plant is independent of the risk of a strike at another plant. Let X = number of plants of the garment manufacturer that strike.

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. Parents have always wondered about the sex of a child before it is born. Suppose that the probability of having a male child was 0.5, and that the sex of one child is independent of the sex of other children.

- **a.** Determine the probability of having exactly two girls out of four children.
- **b.** What is the probability of having four boys out of four children?

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. The Alvin Secretarial Service procures temporary office personnel for major corporations. They have found that 90% of their invoices 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 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 Gallup Poll taken in April 1996. 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.

6.5 The Poisson Distribution

The binomial random variable requires a fixed number of repetitions of the experiment, where the outcomes are either successes or failures. The **Poisson distribution** is similar to the binomial in that the random variable represents a count of the total number of successes. The major difference between the two distributions is that the Poisson does not have a fixed number of trials. Instead, the Poisson uses a fixed interval of time or space in which the number of successes are recorded. Thus, there is no theoretical upper limit on the number of successes, although large numbers of successes are not very likely. The word *success* in the Poisson context can sometimes take on rather unpleasant connotations. For example, the randomness exhibited by the number of airplane crashes, oil tanker spills, and car accidents in some fixed period of time seem to conform to the randomness described by a Poisson random variable.

In business environments many variables seem to follow a pattern of randomness similar to that described by the Poisson distribution. One of the Poisson's principal areas of use in business is the analysis of waiting lines. Other random phenomena, such as airplane arrivals at an airport, trucks arriving at a loading dock, users logging on to a computer system, or the number of defects in a given surface area, can be modeled with a Poisson distribution. These variables are often of interest in determining personnel requirements, inventories, and quality control.

Procedure

Poisson Random Variable

In order to qualify as a **Poisson random variable** an experiment must meet two conditions.

1. Successes occur one at a time. (That is, two or more successes cannot occur at exactly the same point in time or exactly at the same point in space.)
2. The occurrence of a success in any interval is independent of the occurrence of a success in any other interval.

If these two conditions are met, it can be proven that the random variable for the number of successes follows the Poisson probability distribution function.

Kicked by Horses

A real world example of the Poisson distribution involves the distribution of Prussian cavalry deaths from getting kicked by horses, in the period 1875–1894. The Prussian military kept meticulous records on horse-kick deaths in each of its army corps, and the data are neatly summarized in a 1963 book called Lady Luck, by the late Warren Weaver. There were a total of 196 kicking deaths—these being the successes. The trials were each army corps's observations on the number of kicking deaths sustained during the year. With 14 army corps and data for 20 years, there were 280 trials. The Poisson formula predicts, for example, that there will be 34.1 instances of having exactly two deaths in a year. In fact, there were 32 such cases. Pretty good, eh?

Example 6.5.2

Calculating 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 λ = 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

$$\lambda = \frac{100,000}{200,000} = \frac{1}{2} \text{ (average number of defects per 100,000 ft of cable).}$$

Using Table F in Appendix A or technology,

$$P(X = 0) = 0.6065.$$

6.5 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 business?
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. Suppose that, on average, 5 students enrolled in a small liberal arts college have their automobiles stolen during the semester. What is the probability that exactly 2 students will have their automobiles stolen during the current semester?
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. Calculate 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 twenty foot by ten foot 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 fewer 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 his 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. Calculate 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.

6.6 The Hypergeometric Distribution

The binomial and the hypergeometric random variables are very similar. Both random variables have only two outcomes in 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.

Because the binomial and hypergeometric distributions 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

Example 6.6.2

Calculating the Expected Value and Variance of a Hypergeometric Random Variable

Compute the expected value and variance for the random variable defined in Example 6.6.1.

SOLUTION

$$E(X) = 2\left(\frac{4}{15}\right) \approx 0.5333$$

$$\sigma^2 = V(X) = 2\left(\frac{4}{15}\right)\left(1-\frac{4}{15}\right)\frac{(15-2)}{(15-1)} \approx 0.3632$$

Thus, if the experiment were repeated many times, the average number of stocks in the mutual fund that had positive gains would be 0.5333. This leads us to believe that this particular fund is not very promising.

6.6 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 calculate 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.

b. Find the standard deviation of the number of homes the auditor will find with market values less than the outstanding balance of the mortgage.

c. What is the probability that all of the audited homes will have outstanding balances in excess of the mortgage?

d. What is the probability that none of the audited homes will have outstanding balances in excess of the mortgage?

8. A small liberal arts college in the Northeast has 200 freshmen. Eighty of the freshmen are female. Suppose thirty freshmen are randomly selected (without replacement).

a. Find the expected number of females in the sample.

b. Find the standard deviation of the number of females in the sample.

c. Find the probability that none of the selected students will be female.

d. Find the probability that all of the selected students will be female.

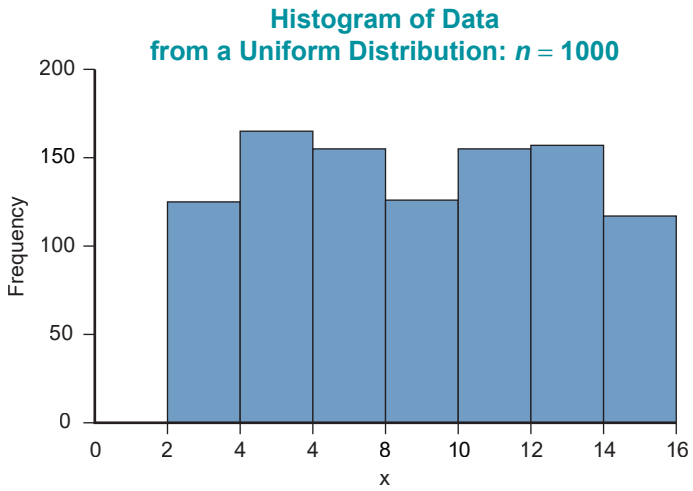


Figure 7.1.4

7.1 Exercises

Basic Concepts

1. Probability is defined differently for discrete and continuous random variables. Describe this difference.
2. What is a probability density function?
3. How is the continuous uniform distribution different from the discrete uniform distribution?
4. What is the uniform probability density function?
5. Describe the shape of the density function for a uniform distribution.

Exercises

6. 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.
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 increase in height of cedar trees is believed to be distributed uniformly between six and eleven 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.

7.2 The Normal Distribution

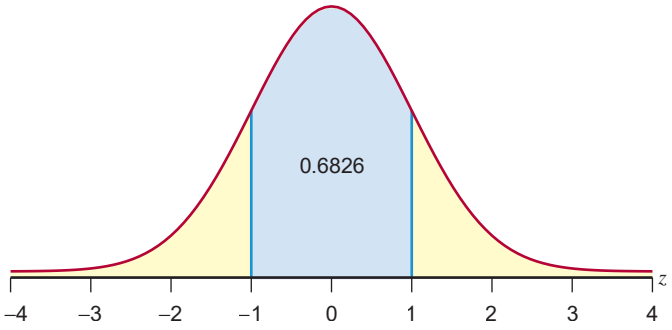


Figure 7.2.1

The **normal distribution**, originally called the Gaussian distribution, was named after Karl Gauss who published a work in 1833 describing the mathematical definition of the distribution. 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, median, and mode are all equal.

Although normally distributed random variables can range in value from negative infinity to positive infinity, values that are a great distance from the mean rarely occur. You may recall that when we discussed box plots, these values were called outliers.

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 a paper in 1733 de Moivre published 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 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.

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.

7.2 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 affect 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.

7.3 Assessing Normality Graphically

The normal distribution is the most important continuous probability distribution. The distribution follows a bell-shaped curve, centered about its mean, and usually, outliers do not have a large impact on the value of the mean. Numerous statistical methods used to analyze data make assumptions about normality including *t*-tests, regression analysis, and analysis of variance, which is the reason we first test the normality of the data before performing any analysis. If the data follow a normal distribution, then we can use parametric procedures to analyze the data; if not, then we will use nonparametric methods to analyze the data.

As stated above, an assessment of normality is the prerequisite for many statistical tests. Typically, a visual check is sufficient to assess normality. In this section, we will discuss graphical techniques that can be used to assess the normality of the data. Graphically assessing normality has the advantage of using the good judgment and expertise of the analyst. Of course, the more expertise the analyst has, the less likely it is that there will be an incorrect interpretation of the data.

In this section, we will discuss the histogram, box plot, and normal probability plot which are graphical techniques used to assess the normality of data. There are many statistical software programs that can be used to create the histogram, box plot, and normal probability plot such as JMP, Excel, and Minitab. We will use JMP to graphically assess normality in the example that follows.

Histogram

As discussed in Chapter 3, a histogram is a graphical method that shows the distribution of a continuous random variable. Even though histograms resemble bar graphs, we analyze them quite differently given that we are using quantitative data. When analyzing the histogram, we can visually see the structure of the data—the shape (symmetric or skewed), the modality of the data, and the existence of outliers.

Lastly, the normal probability plot is displayed in Figure 7.3.6. The solid line represents the theoretical normal data and we can see that the actual data do not stray too far from the line. Again, the normal probability plot supports that the data follow a normal distribution.

Normal Probability Plot of the Sales Revenue Data

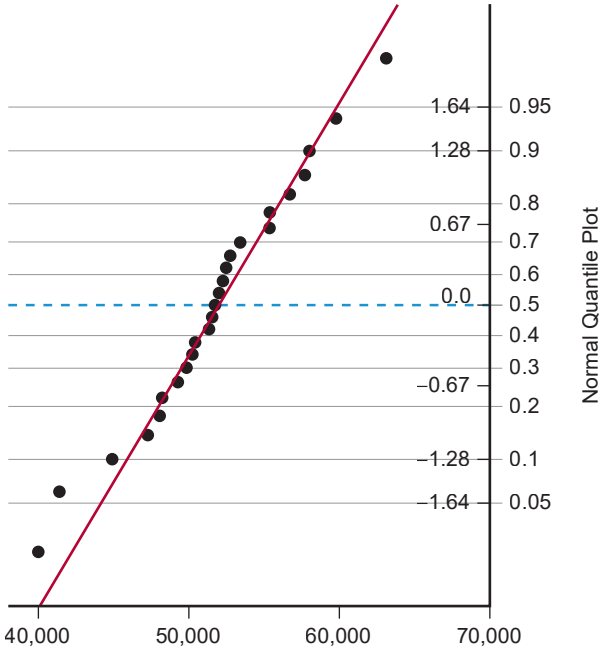


Figure 7.3.6

Given that all three graphical displays used in the example tend to support that the data follow a normal distribution, it is safe to continue with any analysis based on the assumption that the data are normally distributed.

More formal and precise tests are available which calculate the probability that a sample is collected from a normal population. To name a few, these tests are the Anderson-Darling Test, Kolmogorov-Smirnov Test, and Shapiro-Wilk W Test. These tests have the advantage of allowing the analyst to make an objective judgment of normality. However, the primary disadvantage is that these tests are sensitive to small sample sizes.

We have discussed using graphical techniques to determine if the data are collected from a normal distribution. What if the data do not follow a normal distribution? You can still analyze the data but first you must perform some type of transformation on the data or use nonparametric procedures. Keep in mind that if you use nonparametric procedures, they have less power than the classical parametric procedures. Lastly, parametric tests are robust to violations of normality when you have large sample sizes.

7.3 Exercises

Basic Concepts

1. List three ways to graphically assess the normality of a data set.
2. Describe the general procedure for creating a normal probability plot.
3. How should a normal probability plot look to indicate normality?

Exercises

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

The Moneyball data set can be found by visiting stat.hawkeslearning.com and navigating to **Discovering Business Statistics, Second Edition > Data Sets > Moneyball**.

Data

The Housefly Wing Lengths data set can be found by visiting stat.hawkeslearning.com and navigating to **Discovering Business Statistics, Second Edition > Data Sets > Housefly Wing Lengths**.

5. Create a normal probability plot of the housefly wing lengths data. What do you observe?

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 millimeters) in South Carolina was gathered by a weather station in Aiken, South Carolina from 2001-2015. Use a normal probability plot to determine if the data appear to come from a population that is normally distributed.

Total Annual Rainfall in South Carolina

Year	Total Precipitation (in millimeters)	Year	Total Precipitation (in millimeters)	Year	Total Precipitation (in millimeters)
2001	895.7	2006	1031.3	2011	991.8
2002	1106.9	2007	1002.7	2012	1089.5
2003	1681.3	2008	1321.6	2013	1584.0
2004	1003.6	2009	1434.0	2014	1070.2
2005	1166.1	2010	946.2	2015	1537.4

Source: The United States Historical Climatology Network.

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	86.7
86.7	87.5	87.5	90.3	90.4	90.8

Source: https://openmv.net/info/unlimited-time-test

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). 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

Source: Biometrics of Douglas firs. http://seattlecentral.edu/qelp/sets/076/076.html. White River Valley, Washington, 20-Apr-01.Students: Ingrid McNeely and Dylan Morgan, Seattle Central Community College.

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

11. The underwriters for a new type of auto insurance policy gathered monthly mileage data from 18 city drivers. What conclusions can be drawn about the distribution of the data?

a. Using the mileage data provided below, construct a box plot to assess the normality of the data set.

b. Using the mileage data provided below, construct a histogram with eight classes to assess the normality of the data set.

385	410	416.5	421	433.5	451.5
408.5	411	416.5	425.2	437.5	452
408.5	412.5	421	433.5	437.5	460

12. Billings Marketing is asked to develop a recruiting campaign for ABC University. Using the age data from recent college applications shown below, construct a box plot and a histogram with eight classes. Describe the normality of the data set. The results will aid in the development of a marketing strategy.

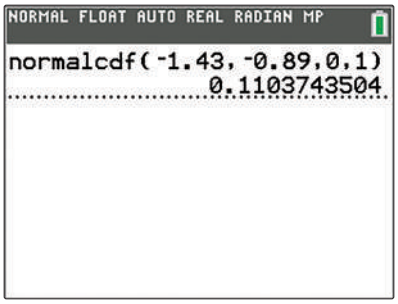
18	19	20	26	22	20
18	19	20	27	22	19
18	28	20	19	23	20
19	19	21	29	19	30
24	19	21	18	24	19
19	19	21	19	25	18

13. Using the data from Exercise 8, construct a box plot and a histogram with 8 columns. The results seem to suggest that the data are not normally distributed. A normal probability plot indicated that the data was normally distributed. Why do these graphical methods for accessing normality seem to contradict one another?

80.8	81.7	81.7	81.7	81.7	82.5
83.3	83.3	84.2	84.2	85	86.7
86.7	87.5	87.5	90.3	90.4	90.8

Technology

For instructions on how to compute this probability using technology, please visit stat.hawkeslearning.com and navigate to **Discovering Business Statistics, Second Edition > Technology Instructions > Normal Distribution > Normal Probability (cdf)**.



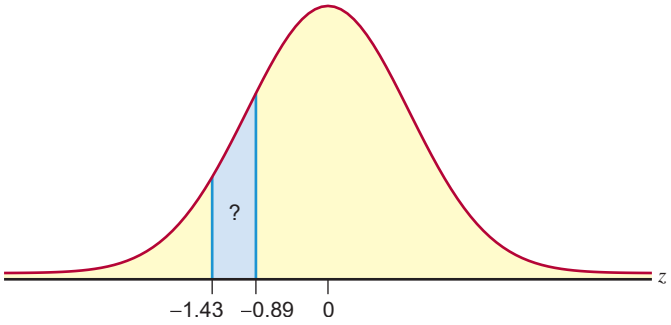


Figure 7.4.19

To find the probability that z is between −1.43 and −0.89, we will need to find the probability that z is less than −0.89 and subtract the probability that z is less than −1.43. Using Table A in Appendix A, we have the following.

$$P(-1.43 < z < -0.89) = P(z < -0.89) - P(z < -1.43) = 0.1867 - 0.0764 = 0.1103$$

Thus, there is approximately an 11% chance that the flight will last between 100 and 150 hours.

7.4 Exercises

Basic Concepts

1. What is the standard normal distribution? What are the parameters of the distribution?
2. Why is the standard normal distribution important?
3. Describe the connection between the z-transformation and the standard normal random variable.

Exercises

4. 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
5. 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
6. 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
7. 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

8. Using the standard normal tables in Appendix A, determine the following probabilities. Sketch the associated areas.

a. $z \le 0$ **c.** $z \le -1$ **e.** $z \ge -1$

b. $z \ge 0$ **d.** $z \le 1$ **f.** $z \ge 1$

9. Using the standard normal tables in Appendix A, determine the following probabilities. Sketch the associated areas.

a. $z \le -0.44$ **d.** $z \le -0.67$

b. $z \ge 0.44$ **e.** $z \ge 0.67$

c. $-0.44 \le z \le 0.44$ **f.** $-0.67 \le z \le 0.67$

10. Using the standard normal tables in Appendix A, determine the following probabilities. Sketch the associated areas.

a. $z \le -1.28$ **d.** $z \le -1.96$

b. $z \ge 1.28$ **e.** $z \ge 1.96$

c. $-1.28 \le z \le 1.28$ **f.** $-1.96 \le z \le 1.96$

11. Using the standard normal tables in Appendix A, determine the following probabilities. Sketch the associated areas.

a. $P(0 \le z \le 0.79)$ **c.** $P(z \ge 1.89)$

b. $P(-1.57 \le z \le 2.33)$ **d.** $P(z \le -2.77)$

12. Using the standard normal tables in Appendix A, determine the following probabilities. Sketch the associated areas.

a. $P(0 \le z \le 1.24)$ **c.** $P(z \ge 3.22)$

b. $P(-2.64 \le z \le 3.32)$ **d.** $P(z \le -3.39)$

13. Find the value of z such that 0.05 of the area under the curve lies to the right of z.

14. Find the value of z such that 0.01 of the area under the curve lies to the right of z.

15. Find the value of z such that 0.10 of the area under the curve lies to the right of z.

16. Find the value of z such that 0.05 of the area under the curve lies to the left of z.

17. Find the value of z such that 0.01 of the area under the curve lies to the left of z.

18. Find the value of z such that 0.10 of the area under the curve lies to the left of z.

19. Find the value of z such that 0.7458 of the area under the curve lies between $-z$ and z.

20. Find the value of z such that 0.9505 of the area under the curve lies between $-z$ and z.

21. Find the value of z such that 0.90 of the area under the curve lies between $-z$ and z.

22. The random variable X has a normal distribution with a mean of 30 and a standard deviation of 5.

a. Find the probability that X is between 25 and 35.

b. Find the probability that X is greater than 40.

c. Find the probability that X is less than 20.

23. The random variable X has a normal distribution with a mean of 200 and a standard deviation of 25.

a. Find the probability that X is between 160 and 220.

b. Find the probability that X is greater than 240.

c. Find the probability that X is less than 150.

24. The Arc Electronic Company had an income of $200,000 last year. Suppose the mean income of firms in the industry for the year is $1,000,000 with a standard deviation of $500,000. If incomes for the industry are normally distributed, what proportion of the firms in the industry earned less than Arc?

25. 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?

26. 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 fraction of the time will users attempt to operate the remote outside of the operating limit?

27. According to the Bureau of Labor Statistics, the mean weekly earnings for people working in a sales related profession in 2010 was $631. Assume that the weekly earnings are approximately normally distributed with a standard deviation of $90.

Source: Bureau of Labor Statistics

- **a.** What are the mean weekly earnings for people working in a sales related profession in 2010?
- **b.** If a salesperson was randomly selected, find the probability that his or her weekly earnings exceed $700.
- **c.** If a salesperson was randomly selected, find the probability that his or her weekly earnings are at most $525.
- **d.** If a salesperson was randomly selected, find the probability that his or her weekly earnings are between $400 and $615.
- **e.** Do you feel that it is reasonable to assume that the weekly earnings have a normal distribution? Why or why not?

28. 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 calculated 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 calculated in **a.**, what is the probability that the repair time for an air conditioning unit will be less than 25 minutes?

29. VGA monitors manufactured by TSI Electronics have life spans which have a normal distribution with an average life span of 15,000 hours and a standard deviation of 2000 hours. If a VGA monitor is selected at random, find the following probabilities.

- **a.** The probability that the life span of the monitor will be less than 12,000 hours.
- **b.** The probability that the life span of the monitor will be more than 18,000 hours.
- **c.** The probability that the life span of the monitor will be between 13,000 hours and 17,000 hours.

30. 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.

31. 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.

32. An investor believes that the yields of his mutual funds have a normal distribution with an average yield of 10% and a standard deviation of 2%. The investor would like to identify the stocks which yield the highest 5% to keep in his portfolio.

- **a.** Calculate the yield which separates the highest 5% of yields from the lowest 95% of yields.
- **b.** If a stock yielded 14% would it be kept? Explain.
- **c.** If a stock yielded 13% would it be kept? Explain.

33. In order for you to become a member of Mensa, a worldwide organization with approximately 100,000 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 50th birthday. American Mensa Ltd., which was founded in 1960 has almost 50,000 members. Marilyn vos Savant, who is reputed to have the highest recorded IQ, is a member. 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?

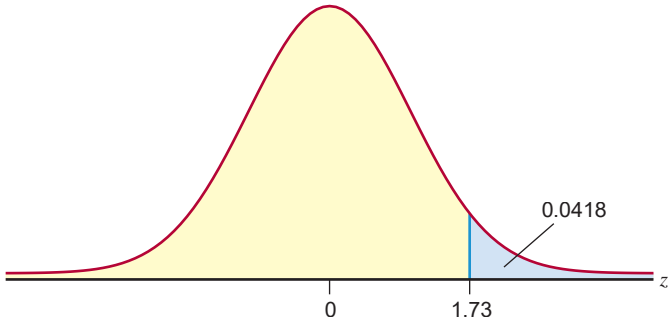


Figure 7.5.9

Thus, using the normal approximation to the Poisson, the probability that at least 40 customers arrive in a given hour is 0.0418.

7.5 Exercises

Basic Concepts

1. Why would you want to use the normal distribution to approximate a binomial distribution or a Poisson distribution?

2. What are the parameters of a normal distribution used to approximate a binomial distribution?

3. What are the parameters of a normal distribution used to approximate a Poisson distribution?

4. What is continuity correction? How does it improve the normal approximation to the binomial or Poisson?

Exercises

5. 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?

6. 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.

7. A local electronics store purchased a market research study which suggests that 60 percent of all homes have DVD recorders/players. 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 DVD recorders/players.

b. Find the standard deviation of the number of homes in the sample which will have video recorders/players.

c. What is the probability that at most 80 of the sampled homes will have DVD recorders/players?

d. What is the probability that between 100 and 120 (inclusive) homes sampled will have DVD recorders/players?

e. What is the probability that at least 130 of the sampled homes will have DVD recorders/players?

8. 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.

9. A company manufacturing metal sheets believes that the number of defects on a 10' by 10' sheet of metal follows a Poisson distribution with an average defect rate of 5 per sheet.

a. Find the standard deviation of the number of defects per sheet.

b. Using the Poisson table in Appendix A, Table F, find the probability of observing at least 10 defects per sheet.

c. Using the normal approximation to the Poisson, find the probability of observing at least 10 defects per sheet.

d. How do the answers in parts **b.** and **c.** compare?

10. Service calls arriving at an electric company follow a Poisson distribution with an average arrival rate of 60 per hour.

a. Find the average number of service calls in a 30-minute period.

b. Find the standard deviation of the number of service calls in a 30-minute period.

c. Using the normal approximation to the Poisson, find the probability that the electric company receives at least 40 service calls in a 30-minute period.

d. Using the normal approximation to the Poisson, find the probability that the electric company receives at most 20 service calls in a 30-minute period.

e. Using the normal approximation to the Poisson, find the probability that the electric company receives between 25 and 50 (inclusive) service calls in a 30-minute period.

11. Patients arriving at the emergency room of a local hospital follow a Poisson distribution with an average arrival rate of 15 per half hour.

a. Find the average number of patients that arrive at the emergency room in one hour.

b. Find the standard deviation of the number of patients that arrive at the emergency room in one hour.

c. Find the probability that at least 15 patients will arrive at the emergency room in one hour.

d. Find the probability that between 30 and 50 patients (inclusive) will arrive at the emergency room in one hour.

e. Find the probability that at most 35 patients will arrive at the emergency room in one hour

Definition

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.

close to their population counterparts. In other words, the sample mean ought to be close to the population mean, the sample proportion ought to be close to the population proportion, and the sample standard deviation should be close to the population standard deviation. Since sample statistics will be used as the basis of the statistical inference, we must know how those statistics vary from one sample to another. Once the variability of the sample statistic is understood, we will be able to make probability statements regarding our inferences.

Why Calculate the Sample Mean?

When analyzing ratio data, the first piece of summary information that an analyst wants to determine is the mean. For most populations, performing a census to determine the population mean is impractical. The only alternative is to use sample information. It seems reasonable that the sample mean, $\bar{x}$, would contain an enormous amount of information about the population mean, μ, and would thus be a sensible estimate of the population mean. Generally, if you wish to estimate a population value—be it the mean, standard deviation, or proportion—the corresponding sample value will be a good **point estimator**.

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. The next section will introduce the distribution of the sample mean and how this distribution can be used to make statistical inferences.

Lava Lamps, Randomness, and Your Bank Account

True randomness is something computers are not currently very good at creating. To generate "random" numbers, computers use an algorithm that produces pseudo-random numbers. A simple example of a pseudo-random number generator would be to take the current day of the month and add the current time in minutes to it, then find the digit of π that corresponds to that "seed" number. For example, if the date is 7/4/2020 and the time is 10:48 AM, then the seed number would be 52 and the 52nd digit of π is 5. Therefore, 5 would be the "random" number that this simple pseudo-random number generator returns.

Every digital transaction that occurs relies on encryption to protect your money. Encryption keys are a random string of characters. Encryption relies on non-predictable encryption keys. The problem with a pseudo-random number generator for encryption purposes is that if a hacker knows how the generator works, then they may be able to "predict" the encryption key and thus destroy the security that the encryption provides. If encryption keys are "predictable," all wire transfer and other online financial transactions would be vulnerable to fraud. For encryption purposes the security of the encryption key is directly associated with the true randomness of its key.

If true randomness is required, the best place to find randomness is in nature.

Cloudflare is an internet security company that uses a wall of lava lamps to generate truly random strings for encryption. The chaotic nature of the fluid dynamics (the bubbles) in the lava lamps makes it impossible to predict what they will look like at any given moment. This enables Cloudflare to take billions of pictures of their wall of lava lamps and generate billions of truly random encryption strings from these pictures.

8.1 Exercises

Basic Concepts

1. Why is the quality of sample data so important?
2. Why is randomness useful in sampling?
3. What is wrong with a voluntary survey?
4. What is a biased sample?
5. What is a sampling frame? Why is this concept important?
6. Discuss how you would draw a simple random sample of the students at your college.
7. What makes drawing a simple random sample from a geographic area a difficult task?
8. Is the sample mean always close to the population mean?
9. Under what conditions is the sample mean considered a random variable?
10. What is the sampling distribution of the sample mean?
11. Describe how statistics as random variables are crucial to statistical inference.
12. What is a point estimator? Give an example.

Exercises

13. 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.

14. 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.

15. 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?

16. Television news programs often conduct opinion surveys by announcing some question on the air and advising viewers to call different numbers for a *yes* or *no* response. National television programs do the same thing except they use 900 numbers and the respondent must pay for the call. Suppose that a national news program asks its viewers to phone in a response to the following: *Women should be permitted to assume combat roles in the military*. The results of the particular survey were 34% *yes* and 66% *no*. Is it reasonable to believe that the results of the survey reflect the attitudes of the nation on this issue? What biases exist in this sampling method?

17. A local politician wants to know what the residents of his community think about an increase in the local property tax to pay for improvements to the highway. He decides to conduct a survey.

a. What is the population of interest to the politician?

b. Can you think of any good sources for a sampling frame?

c. What are the shortcomings (if any) of the sources you picked for the sampling frame?

8.2 The Distribution of the Sample Mean and the Central Limit Theorem

Sample means vary because sample data vary from sample to sample. As an illustration, suppose that an automobile manufacturer wished to determine the average miles per gallon (mpg) of a specific vehicle model that it manufactures. Since determining the mpg of each vehicle is very time consuming, the manufacturer has decided to select two vehicles from a batch of six. Suppose that the actual mpg of the six vehicles are given in Table 8.2.1.

Table 8.2.1 – Miles per Gallon

Car	MPG
A	25
B	27
C	40
D	29
E	28
F	30
Mean	**29.8**
Variance	**23.14**
Standard Deviation	**4.81**

It is important to realize that we are assuming the above set of data constitutes a population. The mean mpg rating of the population in Table 8.2.1 is approximately 29.8 and the population standard deviation is approximately 4.81.

$$\mu \approx 29.8$$
$$\sigma \approx 4.81$$

Both of these measures are considered population parameters. The mpg ratings given in Table 8.2.1 are not known by the manufacturer when the shipment arrives. The manufacturer's job is to estimate the population mean using a sample estimate, in this case using the sample mean from a sample of size two.

How many different samples of size two can be drawn? Assuming no replacement, there would be 15 possible samples of size two if order does not matter. A list of all possible samples and the resulting sample means is given in Table 8.2.2.

8.2 Exercises

Basic Concepts

1. What key three questions should be asked when considering a random variable?
2. Explain the difference between a biased estimator and an unbiased estimator.
3. Give three examples of estimators that are unbiased.
4. Is an unbiased estimator always closer to the parameter being estimated than a biased estimator? Explain.
5. What is the standard error of the mean? What does it indicate?
6. What are two desirable characteristics of the sample mean?
7. Explain the Central Limit Theorem.
8. What effect does increasing the sample size have on the accuracy of an estimate?
9. What is the error of estimation?

Exercises

10. Suppose the random variable X has a mean of 20 and a standard deviation of 5. 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 the random variable X has a mean of 50 and a standard deviation of 10. 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. If there is a normally distributed random variable with a mean of 75 and a standard deviation of 22, what is the probability that the mean of a sample of size 19 will be greater than 80?
13. If a sample of size 40 is drawn from a population that has a mean of 276 and a variance of 81, what is the probability that the mean of the sample will be less than 273?
14. Suppose there is a normally distributed population with a mean of 250 and a standard deviation of 50. If $\bar{x}$ is the average of a sample of 36, find the following probabilities.
    - **a.** $P(\bar{x} \le 240)$
    - **b.** $P(\bar{x} \ge 255)$
    - **c.** $P(246 \le \bar{x} \le 260)$
    - **d.** $P(234 \le \bar{x} \le 245)$
15. Suppose there is a normally distributed population with a mean of 100 and a standard deviation of 10. If $\bar{x}$ is the average of a sample of 50, find the following probabilities.
    - **a.** $P(\bar{x} \le 110)$
    - **b.** $P(\bar{x} \ge 90)$
    - **c.** $P(95 \le \bar{x} \le 115)$
    - **d.** $P(85 \le \bar{x} \le 98)$

16. 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.

17. 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 $110 per day 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 $99 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 $99 per day?

18. 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.

19. 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.

20. 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.

21. The AQI, or the Air Quality Index, is an index used to determine the ozone level in a city. Depending upon the AQI reading, it may not be safe to jog or even to go outside. Readings in the 0–50 range mean that the air quality conditions are considered "good," 51–100 are "moderate," 101–150 means "unhealthy for sensitive groups," 151–200 means "unhealthy," 201–300 means "very unhealthy," and 301–500 means "hazardous." Suppose that an industrial region has an average AQI reading of 102 with a standard deviation of 40. Find the probability that for a random sample of 50 days, the average AQI reading is:

Source: airnow.gov

- **a.** at least 105.
- **b.** at most 90.
- **c.** between 100 and 115.

($np \geq 5$ and $n(1-p) \geq 5$), the distribution of $\hat{p}$ will be approximately normal with

$$\mu_{\hat{p}} = p = 0.6$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.6(1-0.6)}{150}} = 0.04.$$

To find the probability that $\hat{p}$ is within 0.05 of the true proportion, we must find $P\left(p-0.05 < \hat{p} < p+0.05\right) = P\left(0.55 < \hat{p} < 0.65\right)$.

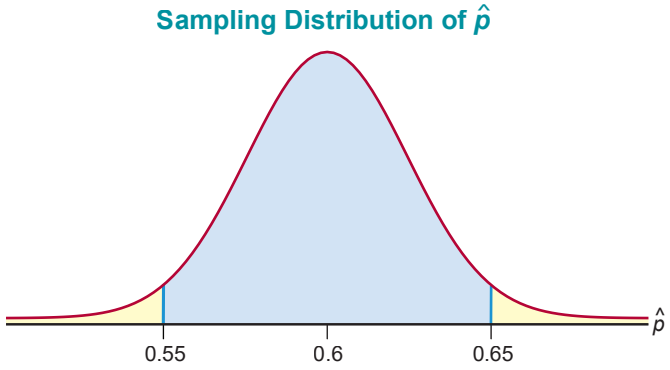


Figure 8.3.6

Using the z-transformation,

$$= P\left(\frac{0.55-0.6}{0.04} < z < \frac{0.65-0.6}{0.04}\right)$$
$$= P\left(-1.25 < z < 1.25\right)$$
$$= P\left(z < 1.25\right) - P\left(z < -1.25\right)$$
$$= 0.8944 - 0.1056$$
$$= 0.7888.$$

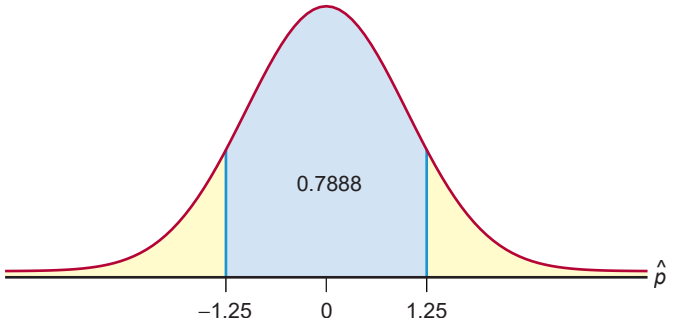
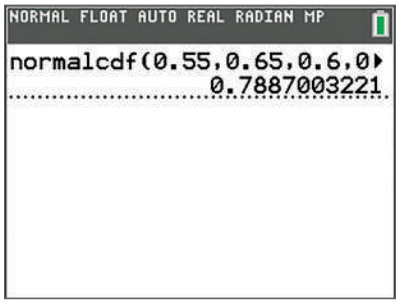


Figure 8.3.7

Therefore, for a sample of 150 U.S. citizens over 18 years of age, it is likely (0.7888) that the error of estimation will be less than five percentage points.

Technology

For instructions on calculating this probability using technology, please visit stat.hawkeslearning.com and navigate to **Discovering Business Statistics, Second Edition > Technology Instructions > Normal Distributions > Normal Probability (cdf)**.



8.3 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.

Exercises

6. A random sample of 40 electronic components has 5 defective components.
   a. Find the sample proportion of components that are defective.
   b. Find the sample proportion of components that are not defective.
7. A random sample of 100 employees of a large steel company has 30 females and 70 males.
   a. Find the sample proportion of female employees.
   b. Find the sample proportion of male employees.
8. Suppose that the true proportion of registered voters who favor the Republican presidential 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?
9. 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 sample proportion for samples of the following 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?
10. Suppose the true population proportion is $p = 0.50$. What is the probability that the sample proportion of a sample of size 20 will be greater than 0.60?
11. Suppose the true population proportion is $p = 0.30$. What is the probability that the sample proportion of a sample of size 30 will be less than 0.20?
12. 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 38, 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. $P\left(0.18 < \hat{p} < 0.25\right)$
13. Suppose that the true proportion of airline pilots between the ages of 35 and 45 is 0.60. If $\hat{p}$ is the sample proportion of airline pilots between the ages of 35 and 45 from a sample of size 100, find the following probabilities.
    a. $P\left(\hat{p} > 0.55\right)$
    b. $P\left(\hat{p} < 0.45\right)$
    c. $P\left(0.50 < \hat{p} < 0.60\right)$
    d. $P\left(0.60 < \hat{p} < 0.75\right)$

14. 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 rock 'n' roll music has increased from 25% to 35%. The director randomly decides to select 50 young professionals and ask them if they prefer rock 'n' roll to any other type of music. If the sample proportion is greater than 0.35, he will switch to a new format emphasizing rock 'n' roll.

a. If the true proportion of young professionals who prefer rock 'n' roll 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 rock 'n' roll has changed as the director suspects, find the probability that the radio director will switch to the new format.

15. The property manager of a large office building would like to make the building smoke free; however, he does not want to upset too many of his customers. 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 property manager will make the building smoke free.

a. Find the probability that the property manager will make the building smoke free when the true proportion of smokers is 0.5.

b. Find the probability that the property manager will not make the building smoke free when the true proportion of smokers is 0.2.

16. Eighty percent of the flights arriving in Atlanta for a large U.S. 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.

17. Approximately 7% of the nation's public school children in grades 2 through 5 took medication in 2003 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. Assume that a suburban elementary school had an enrollment of 286 students in 2003.

a. Find the probability that at least 4% of the school children took medication for ADHD.

b. Find the probability that between 5% and 8% of the school children took medication for ADHD.

8.4 Sampling Methods

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 consideration is a requirement for simple random sampling, a requirement which can be time-consuming and expensive. In addition, it is sometimes not even possible to list all the members of a population. There are other sampling strategies that are designed to reduce the cost of sampling or add control to the sampling procedure. These techniques can be categorized as probability samples or non-probability samples.

Probability samples enable an analyst to determine the probable errors that an estimator might generate. Essentially, they allow the analyst a known degree of confidence in his or her estimation. All of statistical inference relies on probability sampling. **Non-probability samples**

Definition

Probability Sample

A **probability sample** is a sample used to estimate a population parameter that has known errors, allowing a statement about the reliability of the estimate to be made.

8.4 Exercises

Basic Concepts

1. What are the 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?

Exercises

8. An employee-owned company has 6000 female employees and 2000 male 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 sexes, it has decided to randomly select 500 females and 500 males. 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.
9. Explain why a systematic sample is not a random sample.
10. Suppose you were instructed to draw a simple random sample from a metropolitan area in order to gain information on the citizens' view on a proposed amendment to the state constitution. To create a simple random sample, you must create a sampling frame. You have decided to use the telephone directory as your frame for the metropolitan area.
    - **a.** Identify the population under consideration.
    - **b.** What kinds of people will be omitted from your frame?
    - **c.** What kinds of biases will be introduced in your sample as a consequence of the omission you described in part **b.**? Can you think of ways of compensating for the bias?

11. 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.

12. 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

13. 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

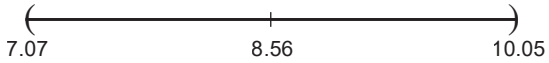
- **a.** What sampling technique was used for this survey?
- **b.** What biases may be present in the responses?
- **c.** Is 50% a reasonable point estimate of the proportion of Orlando residents who favor the state income tax? Explain.

A 98% confidence interval is then calculated as follows.

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$8.56 \pm 2.33\frac{7.85}{\sqrt{150}}$$

$$7.07 \text{ to } 10.05$$



Thus, we are 98% confident that the true mean number of new products introduced in the last 12 months will be contained in the above interval.

Technology

For instructions on calculating this confidence interval using technology, please visit stat.hawkeslearning.com and navigate to **Discovering Business Statistics, Second Edition** > **Technology Instructions** > **Confidence Intervals** > ***z*-Interval**.

So far, the confidence interval has been discussed as a way of placing bounds on the location of a parameter with a specific degree of confidence. But we can also think about the confidence interval as a means of describing the quality of a point estimate. Let's look at the expression for the confidence interval for the population mean.

Confidence Interval for μ

$$\underbrace{\bar{x}}_{\text{point estimate}} \quad \pm \quad \underbrace{z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}_{\substack{\text{margin of error with a} \\ \text{specific level of} \\ \text{confidence}}}$$

Another interpretation of the confidence interval is given below the expression of the confidence interval for μ. The part of the expression that is added and subtracted to the point estimate, $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, can be thought of as the **margin of error** (also known as the **maximum error of estimation**) using the point estimate $\bar{x}$ with a specified level of confidence. For example, the 95% confidence interval in Example 9.1.1 was given as

$$425 \pm 1.96 \cdot \frac{900}{\sqrt{100}}$$

$$425 \pm 176.4.$$

We could say that we are 95% confident that the point estimate of μ, $\bar{x} = 425$, has a margin of error of 176.4 or an error of estimation no larger than 176.4. Being able to assess the error of an estimate is one of the most useful applications of statistical methods.

Definition

Margin of Error

The **margin of error**, or **maximum error of estimation** (often denoted as E), is the largest possible distance from the point estimate that a confidence interval will cover.

9.1 Exercises

Basic Concepts

1. What is statistical inference?
2. What is an estimator?
3. What is a judgment estimate? What are some drawbacks of judgment estimates?
4. Explain, in your own words, the difference between the terms *estimator* and *estimate*.
5. What is the difference between a point estimate and an interval estimate?
6. Give three examples of point estimators. Identify the parameters being estimated by these estimators.
7. Describe the primary advantages of *random* sampling procedures.
8. What are two important questions to consider when estimating a population mean?

9. What is mean squared error?

10. What is an unbiased estimator? Give an example.

11. Why is the sample mean considered the best point estimate of the population mean?

12. Are all estimators unbiased? Explain.

13. Generally, we expect most sample statistics to be good estimators of their population counterparts. Which statistic is the exception to this idea?

14. What are two characteristics of the best available estimate for a parameter?

15. What is an interval estimator?

16. What is the distinction between probability and confidence?

17. What is the role of the z-value in the confidence interval expression?

18. Describe in words the ideas behind the construction of a confidence interval.

19. Consider the following statement: *If the sample size is greater than or equal to 30, then by the Central Limit Theorem there is a 0.95 probability that the sample mean will be within 1.96 standard deviations of the population mean before a particular sample is selected.* Explain why the phrase "before a particular sample is selected" is important here.

20. Explain what is wrong with the following expression: $P(111 < \mu < 189) = 0.95$.

21. Define the following terms: confidence level, confidence coefficient, confidence interval.

22. What are the conditions required in order to construct a $100(1-\alpha)\%$ confidence interval using the expression $\bar{x} \pm z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$?

23. Describe the effect on the width of a confidence interval as each of the following increases:

a. n **b.** $1-\alpha$ **c.** α **d.** $\bar{x}$

24. What expression indicates the margin of error? Is this the same as the maximum error of estimation?

Exercises

25. Find $z_{\alpha/2}$ for the following levels of α.

a. $\alpha = 0.05$ **b.** $\alpha = 0.01$ **c.** $\alpha = 0.10$

26. Find $z_{\alpha/2}$ for the following levels of α.

a. $\alpha = 0.04$ **b.** $\alpha = 0.02$ **c.** $\alpha = 0.08$

27. Find $z_{\alpha/2}$ for the following confidence levels.

a. 98% **b.** 94% **c.** 92%

28. Find $z_{\alpha/2}$ for the following confidence levels.

a. 96% **b.** 88% **c.** 85%

29. Consider a normally distributed population with a standard deviation of 64. If a random sample of size 90 from the population produces a sample mean of 250, construct a 95% confidence interval for the true mean of the population.

30. 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.

31. A psychologist is studying learning in rats. The psychologist wants to determine the average time required for rats to learn to traverse a maze. She randomly selects 40 rats and records the time it takes for the rats to traverse the maze in minutes. The sample average time required for the rats to traverse the maze is 5 minutes with a population standard deviation of 1 minute. Estimate the average time required for rats to learn to traverse the maze with a 90% confidence interval.

32. 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.

33. 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 breaking strength for a sample of 50 pieces of chain is 5000 pounds. Based on past experience, the breaking strength of metal chains is known to be normally distributed with a standard deviation of 100 pounds. Estimate the actual mean breaking strength of the metal link chain with 99% confidence.

34. Tomatoes are grown in Florida for shipment to other parts of the country by the Anderson Produce Company. A random sample of 40 boxes is selected at one warehouse for weighing. The average weight for the sample is 33.5 pounds per box with a population standard deviation of 2.1 pounds. Find a 90% confidence interval for the true average weight of the boxes of tomatoes.

35. Thirty-five strands of piano wire were selected at random from a recent shipment by the quality control department at Elkins Piano Company. The strands of piano wire were tested to failure in tests of tensile strength. The mean tensile strength of the sample was 30,000 pounds per square inch (psi) with a population standard deviation of 1950 psi. Find the 98% confidence interval for the true mean tensile strength.

36. When preparing a standardized test to be given to all the sixth graders, the Standard Test Company gave a version of the test to a random sample of 45 sixth graders and timed how long it took them to finish the test. The average time required to finish the test for the sample was 2 hours and 15 minutes with a population standard deviation of 30 minutes. Estimate the true average time required to finish the test with 95% confidence.

37. According to the 2009 College Senior Survey administered by the Higher Education Research Institute at UCLA, 56.4% of college seniors spend 10 hours or less studying or doing homework in a typical week. Suppose a random sample of 50 college seniors was selected from all the college seniors in the Southeast region to determine the homework habits of college seniors in the Southeast region of the United States. Each student in the sample is asked approximately how many hours per week he or she spends studying or doing homework. If the mean is 9.6 hours and the population standard deviation is 3.1 hours, construct a 99% confidence interval for the mean number of hours a week that a college senior in the region spends studying or doing homework per week.

Source: Cooperative Institutional Research Program at the Higher Education Research Institute at UCLA

9.2 Estimating the Population Mean, σ Unknown

In the previous section, we assumed that the population standard deviation was 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?

To be assured of finding the desired level of confidence, always round up. Thus, we are 90% confident that a sample of $n = 98$ observations would produce an estimate of the mean amount of cleaning fluid in a 12-ounce bottle to within 0.05 ounce. Being able to know the accuracy of your estimate is one of the significant benefits of inferential statistics. In this case, if 98 bottles are measured, we will be 90% confident that the resulting sample mean is within five one-hundredths of an ounce of the true mean. That's close.

Determining the Sample Size: σ Unknown

In the previous discussion of determining the sample size necessary to estimate a population mean with a desired accuracy, σ was assumed to be known. This assumption is usually unreasonable in most problem-solving environments.

The most obvious method for obtaining an estimate of σ is to take a small sample and use the sample standard deviation as an estimate of the population standard deviation. Replacing σ with s in the sample size determination relationship will provide an initial estimate of the required sample size. Another alternative is to use the value of the sample standard deviation obtained in a previous study, sometimes called a **pilot study**. Keep in mind that in order to construct a confidence interval for a population mean when the population standard deviation is unknown, the distribution of the population is assumed to be normal.

Example 9.2.5

Calculating the Sample Size Needed for 99% Confidence of the Mean Time Required to Replace a Jet Engine

An airline's maintenance manager desires to estimate the average time (in hours) required to replace a jet engine in a Boeing 767. How large a sample would be necessary if the manager wishes to be 99% confident of estimating the population mean to within one-quarter of an hour ($E = 0.25$)? Assume a preliminary sample of size $n = 30$ has a mean replacement time of 16.7 hours with a standard deviation of 4.3 hours.

SOLUTION

Using the results from the initial sample,

$$n = \left(\frac{z_{\alpha/2}s}{E}\right)^2 = \left(\frac{2.575 \cdot 4.3}{0.25}\right)^2 = 1961.6041$$

$n = 1962$ (Always round up to assure 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 1000 observations was 4.1. Using this estimate of s instead of 4.3 results in a sample size of 1784 compared to the original specification of 1962. The notion of modifying the sample size estimate as additional data are observed can be applied at regular intervals during the sampling process until the estimate of the standard deviation stabilizes.

Six Degrees of Separation: A Law of Small Worlds

What is the number of people a randomly chosen person in Omaha, Nebraska needs to contact before she can find a connection with a randomly chosen housewife in New England? How many intervening people do you think separates you from the President of the United States? Unsuspecting readers might guess very large numbers but the actual numbers are quite small. The answer to both of these questions may very well be less than 6! Psychologists have done ingenious experiments and have actually calculated this degree of separation, on average, to be six. What is amazing about this degree of separation is that it is equally true for the President of the United States and a sweet vendor in Bangladesh.

continued on next page...

9.2 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 are the conditions in which the t-distribution is used in interval estimation of the population mean?
5. What is the parameter of the t-distribution? How is it calculated?
6. What is the value of having a confidence interval with a small width?
7. Can a confidence interval be constructed with a width of your choice? Explain.
8. What are the three components that affect the width of the confidence interval for the population mean? Describe how changes in these three components affect the width of the confidence interval.
9. 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?
10. What is the rounding rule regarding the determination of the sample size?
11. What is the difference between the method of determining the sample size when σ is known versus when σ is unknown?
12. What is a pilot study?
13. Note that E and s can be viewed as measures of variation. Compare and contrast the meanings of E and s in layman's terms.

continued...

The degree of separation for the number of clicks that you will need to make to get to a website that interests you, as well as the analysis of terrorist networks, turn out to be of similar nature. The new science of networks can shed useful light and help to derive general laws applicable to many of these types of questions.

Exercises

14. 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.
15. 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.
16. 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. $\alpha = 0.10, n = 8$
17. Find $t_{\alpha/2,df}$ for the following combinations of α and n.
    a. $\alpha = 0.05, n = 12$
    b. $\alpha = 0.01, n = 18$
    c. $\alpha = 0.10, n = 22$
18. 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

19. 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

20. 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.

Fat Contents			
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?

21. 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.

22. 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 12.3% 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.

23. 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 $115 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.

24. In 2010 the median home price in all regions of the United States was $221,800. It is commonly thought that better schools are found in wealthier areas. In *Forbes* magazine's list of the "Best Schools for your Real Estate Buck," the top 10 cities in America were identified where your housing dollar will go the furthest in getting your children a great education. 17,589 towns and cities were analyzed using results from the most recent National Assessment for Educational Progress data, and the top 10 school districts were identified. The list counteracted the idea that more money equals better schools, as Falmouth, Maine topped the list beating out high-dollar school districts like Manhattan Beach, California. The top 10 cities are given below, along with the median home price for each city.

Best Schools for Your Real Estate Buck		
Education Rank	**City**	**Median Home Price ($)**
1	Falmouth, Maine	351,550
2	Mercer Island, Washington	708,740
3	Pella, Iowa	148,200

Best Schools for Your Real Estate Buck (cont.)

Education Rank	City	Median Home Price ($)
4	Barrington, Rhode Island	296,010
5	Bedford, New Hampshire	293,730
6	Manhattan Beach, California	1,278,980
7	Moraga, California	722,010
8	Parkland, Florida	426,390
9	St. Johns, Florida	181,700
10	Southlake, Texas	476,880

Source: Forbes magazine

a. Construct a 90% confidence interval for the median home price of cities on the top 10 list.

b. Is the average median price for these cities higher than the median price for the U.S. as a whole?

c. What population assumption needs to be made here?

d. How would your solutions to **a.** and **b.** change if these were mean rather than median values?

25. 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?

26. 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.

27. A hot-dog 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?

9.3 Estimating the Population Proportion

An attribute is a characteristic that members of a population either possess or do not possess. Attributes are almost always measured as the **proportion** of the population that possesses the characteristic.

Many decisions require a measure of a population attribute. Television and radio stations base their advertising charges on ratings reflecting the *percentage of television viewers who are watching a particular program*. A political analyst wants to know the *fraction of voters who favor a particular candidate*. A social researcher needs the *fraction of teachers who believe group learning is a beneficial instructional method*. An insurance company is interested in estimating the *fraction of their policies that will result in claims*. A quality control engineer requires the *percentage defective in a lot of goods*. A marketing researcher demands the

9.3 Exercises

Basic Concepts

1. What is a proportion? What type of information does it give us about the population?
2. How is the sample proportion found?
3. Describe, in layman's terms, how a confidence interval is constructed for a population proportion.
4. It seems that estimating proportions produces estimates which are much more precise than those for means. Explain why this is the case.
5. The population proportion is often unknown. How is this issue dealt with when determining sample size?
6. What is the guideline to follow when there is no estimate available for the population proportion? Why is this done?
7. How do the resulting required sample sizes differ when there is an estimate available versus when there is no estimate available for the population proportion?

Exercises

8. 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. He randomly selects 200 lakes and determines that 45 of the selected lakes 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?
9. *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.
10. 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?
11. The Peacock Cable Television Company thinks that 40% of their customers have more outlets wired than they are paying for. A random sample of 400 houses reveals that 110 of the houses have excessive outlets.
    a. Construct a 99% confidence interval for the true proportion of houses having too many outlets.

b. Do you feel the company is accurate in its belief about the proportion of customers who have more outlets wired than they are paying for? Justify your answer.

12. Running continues to be a very popular sport in America. At a major race, like the Peachtree Road Race in Atlanta, 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.

13. In the fourth quarter of 2010 the home ownership rate was 66.5%. This rate is 2.7 percentage points lower than the 2004 peak of 69.2%, and the lowest rate since 1998. Home ownership fell at an alarming pace in the fourth quarter of the year, despite the fact that home prices fell, affordability was much improved, and inventories of new and existing homes were running quite high. Suppose that a random sample of 120 households was selected from an area in the Midwest that is particularly economically depressed. Suppose that 57 of the households sampled were owned by the residents of the homes.
Source: U.S. Census Bureau

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?

14. 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. Assuming that a random sample was chosen, construct a 90% confidence interval for the percentage of all Americans who say that vegetables are their least favorite food.

15. The Federal Trade Commission (FTC) conducted a study investigating the accuracy of bar-code scanners. It was concluded that these computer scanners, used mainly at grocery, department store, and drugstore checkout counters, ring up the wrong price about 5% of the time. In most instances, however, the error was in favor of the shopper, according to the FTC. Suppose that your local grocery store conducts a study to determine the accuracy of its scanners. Assume 13 shoppers are randomly chosen and their bills, as indicated by the scanner, are checked against the correct bill computed by conventional means. Suppose that of the 200 items scanned, 21 of the items were charged incorrectly by the scanner.

a. Construct a 95% confidence interval for the proportion of items that were rung up incorrectly by the scanner.

b. Does it appear that the local grocery store has a larger error rate than 5%?

16. The Big Green Poster Company wants to estimate the fraction of poster sites controlled by their competition, Bird's Billboard Service. What sample size would be necessary to estimate this fraction to within 3% with 95% confidence? (They think Bird's controls about 33 percent of the boards.)

17. 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?

18. 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?

19. 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.

9.4 Estimating the Population Standard Deviation or Variance

Recall that the sample variance is

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

and it serves as the point estimate of the population variance, σ^2. As with the other tests developed for the population mean and the population proportion, we first need to develop a sampling distribution for

$$\frac{(n-1)s^2}{\sigma^2}$$

that will allow us to calculate a confidence interval for the population variance.

Formula

χ^2 Test Statistic

If we have a random sample of size n taken from a normal population, then the sampling distribution of the test statistic is given by

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

which has a **chi-square distribution** with $n - 1$ degrees of freedom.

The chi-square distribution is a positively skewed (or skewed to the right) distribution. Like the *t*-distribution, the shape of the distribution is a function of its degrees of freedom. See Figure 9.4.1 which illustrates the chi-square distributions with 4 and 10 degrees of freedom.

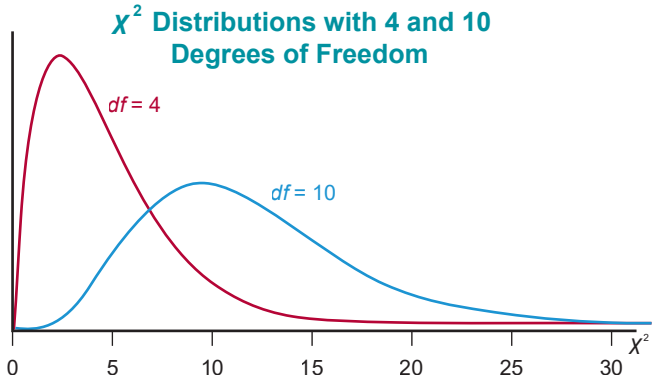


Figure 9.4.1

Example 9.4.1

Calculating a Confidence Interval for the Population Standard Deviation

The quality control supervisor of a bottling plant is concerned about the variance of fill per bottle. Regulatory agencies specify that the standard deviation of the amount of fill should be less than 0.1 ounce. To determine whether the process is meeting this specification, the supervisor randomly selects ten bottles, weighs the contents of each, and finds that the sample standard deviation of these measurements is 0.04. Assume that the data are collected from a normal population and compute a 95% confidence interval for the standard deviation of ounces of fill for the bottling plant.

SOLUTION

We want to find a 95% confidence interval for the variance. We are given that

$$n = 10,\ s = 0.04, \text{ and } \alpha = 0.05.$$

To calculate a 95% confidence interval, we use the following formula.

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

Thus, we need to find the values of $\chi^2_{0.025}$ and $\chi^2_{0.975}$ for $n - 1 = 10 - 1 = 9$ degrees of freedom.

Using Table G in Appendix A, at 9 degrees of freedom,

$$\chi^2_{0.025} = 19.023$$
$$\chi^2_{0.975} = 2.700.$$

Substituting the values in the formula above, we have

$$\frac{(10-1)(0.04)^2}{19.023} < \sigma^2 < \frac{(10-1)(0.04)^2}{2.700}$$
$$0.000757 < \sigma^2 < 0.00533$$

So, a 95% confidence interval for the variance of fill of the bottles is between 0.000757 and 0.00533 ounce. However, the problem mentions the tolerance for the standard deviation of fill. So, to ensure that we make our interpretation in terms of the problem, to find a 95% confidence interval for the standard deviation, we take the square root 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 and 0.0730 ounce, indicating that the process is meeting the specifications of being less than 0.1 ounce.

Technology

The confidence interval for the population variance can be obtained using Minitab. For detailed instructions, visit stat.hawkeslearning.com and navigate to **Discovering Business Statistics, Second Edition** > **Technology Instructions** > **Confidence Intervals** > **Variance**.

9.4 Exercises

Basic Concepts

1. What is the sampling distribution for $\frac{(n-1)s^2}{\sigma^2}$?
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. Give an example where we would want to calculate a confidence interval for σ^2

Exercises

5. A bolt manufacturer is very concerned about the consistency with which his machines produce bolts that are ¾ inch in diameter. When the manufacturing process is working normally the standard deviation of the bolt diameter is 0.05 inch. A random sample of 30 bolts has an average diameter of 0.25 inch with a standard deviation of 0.07 inch.

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 diameters of the bolts in constructing the confidence interval in part **a.**?

6. 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 amounts of drug contained in the tablets in constructing the confidence interval in part **a.**?

7. 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.**?

8. 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?

9. 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.

10. Photographers are always concerned about the number of shutter actuations that they will get from their cameras before they need to be serviced or the shutter needs to be replaced. To get an idea of the variability associated with the number of actuations, a photographer took a random sample of 20 cameras and found that the average number of actuations before failure was 200,000 with a standard deviation of 50,000.

a. Construct a 95% confidence interval for the standard deviation of the shutter actuations. Interpret the interval.

b. What assumptions did you make about the number of shutter actuations for the cameras?

A Procedure for Testing a Hypothesis

A procedure for testing a hypothesis is given below. In examining this procedure, you will notice we have already discussed the first two steps. The next four steps assist us in defining the **decision rule**, which is a criterion used to determine whether the null or the alternative hypothesis will be chosen. As you look over these steps, do not be overly concerned if you do not understand everything. You will learn by following the examples.

Steps in the Test of a Hypothesis

Step 1: Determine the null hypothesis. In this process, select the appropriate statistical measure, such as the population mean, proportion, or variance.

Step 2: Determine the alternative hypothesis and whether it should be one-sided or two-sided.

Step 3: Select the appropriate test statistic based on the information at hand and the assumptions you are willing to make.

Step 4: Determine the critical value of the test statistic. Two factors must be considered.

1. The type of alternative hypothesis: two-sided, one-sided left, one-sided right. If the alternative hypothesis is two-sided, the hypothesis test will be **two-tailed**. If the alternative hypothesis is one-sided left, the hypothesis test will be a **left-tailed** or **lower-tailed** test. If the alternative hypothesis is one-sided right, the hypothesis test will be a **right-tailed** or **upper-tailed** test.
2. The specification of α, the significance level of the test.

Step 5: Collect the sample data and compute the value of the test statistic.

Step 6: Make the decision and state the conclusion in terms of the original question.

- If the value of the test statistic is in the rejection region, reject the null hypothesis in favor of the alternative.
- If the value of the test statistic is not in the rejection region, fail to reject the null hypothesis.

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.

Type I and Type II Errors in the Trial of the Pyx

If the coins did in fact weigh less than they were intended to, the currency would become debased, and the Mint would be making a profit because they would be pocketing some of the metals they should be turning into coins. If the coins weighed more than they were supposed to, someone could collect these overweight coins, and sell them back to the Mint for a profit. Either way, the king is not happy that someone besides him is able to profit. And in those days, if the king is not happy, there is a high likelihood of important body parts being involuntarily cut off. So, if the coins are found to be off from the standard value, it could mean serious consequences for the head of the Mint.

The hypotheses are set up such that a Type I error implies that the coins were believed to be off from the standard value, when in fact they were meeting the standard. A Type II error would mean that the panel has believed the coins to be matching the weight standard, when in fact they are overweight or underweight. This is the preferred formulation for the head of the Mint, as the Type I error is the one he would like to control so that he does not lose his extremities for no reason! A Type II error would serve the Mint well, as it means the coins were in error, but it went undetected.

10.1 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. What is the connection between one and two-sided alternatives and one and two-tailed tests?
7. Is there a way to be absolutely certain your decision is correct when performing a hypothesis test? Explain.
8. What are the three important things you must be able to do in order to be successful at formulating hypothesis testing problems?
9. Describe a Type I error.

10. Describe a Type II error.

11. Explain how Type I and Type II errors influence the construction of a hypothesis.

12. Can both Type I and Type II errors be controlled in the hypothesis testing procedure? Explain.

13. What is the level of the test?

14. Why is a Type II error difficult to express numerically?

Exercises

15. 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?

16. 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?

17. 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 alternate hypotheses for this test.

18. 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.

19. 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.

20. 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.

21. An automobile manufacturer claims that their van has a 56.8 miles/gallon (MPG) rating. An independent testing firm has been contracted to test the MPG for this van. After testing 99 vans they found a mean MPG of 56.4 with a standard deviation of 1.2 MPG. Is there sufficient evidence at the 0.025 level that the vans underperform the manufacturer's MPG rating? State the null and alternative hypotheses for this test.

22. 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?

23. 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 30,000 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. Mrs. Russell, head product tester for Hathaway Tool Corporation, is testing a newly designed series of bar hooks. The hooks have been designed to give way if they get too hot. The previous design gave way at 240 degrees. Develop a test to determine if the newly designed hooks give way at a higher temperature than the previous design.

24. 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 believes that a new sports coupe gets significantly fewer miles to the gallon than advertised on the sales sticker. To confirm this belief, they randomly select several of the new coupes and measure the miles per gallon.

10.2 Testing a Hypothesis about a Population Mean, σ Known

The following example will be used to illustrate the hypothesis testing procedure. The example goes through all the steps and contains detailed explanations of several new concepts.

Example 10.2.1

Performing a Hypothesis Test for the Mean Cost of Textbooks

Suppose that the average amount of money a student spends on textbooks per semester at college campuses in the U.S. is \$500 with a standard deviation of \$100. A local university wants to know if its students are actually spending that amount on textbooks. The level of the test is to be set at 0.05. A random sample of 75 college students has been selected and the resulting average is \$540 with a sample standard deviation of \$95 spent per semester on textbooks.

SOLUTION

Step 1: Determine the null hypothesis. In this process, select the appropriate statistical measure, such as the population mean, proportion, or variance.

The null hypothesis is fairly straightforward. The local university's students are spending an average of \$500 per semester on textbooks. Thus, the statistical measure is the population mean which is

μ = the average amount spent on textbooks per semester by college students at the local university.

The null hypothesis should be written as H_0: $\mu = \$500$.

Step 2: Determine the alternative hypothesis and whether it should be one-sided or two-sided.

The alternative hypothesis is that the students are not spending an average of \$500 per semester on textbooks. Nothing is mentioned in the problem to indicate that the university is interested in learning if its students are spending more than the national average per semester on textbooks, nor is there anything mentioned that the university is interested in learning if its students are spending less than the national average per semester on textbooks. Thus, it is assumed that the university is interested in determining if there is any departure from the national average. Consequently, the problem needs to be formulated as a two-sided alternative which should be written as H_a: $\mu \neq \$500$.

Steps in the Test of a Hypothesis Using the *P*-Value Approach (cont.)

Step 5: Calculate the *P*-value using the test statistic. For the sake of this setup, suppose we are performing a hypothesis test on the mean when σ is known. Thus, the observed value of the test statistic will be

$$z_0 = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}.$$

The *P*-value will be found as follows:

a. If the alternative hypothesis is H_a: $\mu < \mu_0$, then the *P*-value is calculated as $P(z \le z_0)$.

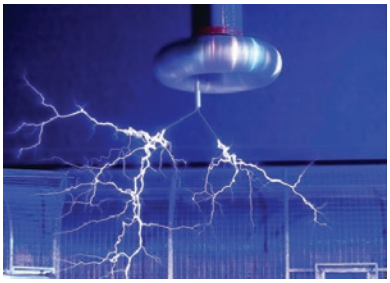
b. If the alternative hypothesis is H_a: $\mu > \mu_0$, then the *P*-value is calculated as $P(z \ge z_0)$.

c. If the alternative hypothesis is H_a: $\mu \ne \mu_0$, then the *P*-value is calculated as $2P(z \ge |z_0|)$.

Note that in this example, we are performing a test on the population mean with σ known. If the parameter that is being tested changes so that the test statistic changes, then we would use the appropriate test statistic in the above probability statements.

Step 6: Make the decision and state the conclusion in terms of the original question. The decision-making process is as follows.

- If the *P*-value is less than or equal to α, reject the null hypothesis in favor of the alternative hypothesis.
- If the *P*-value is greater than α, fail to reject the null hypothesis.



Different *P*-Values for Different Folks

In particle physics, the standard for "discovery" is a *P*-value less than 0.0000003. That is the probability which corresponds to observing a value that is at least 5 standard deviations from the mean for a one-tailed test. Particle physicists consider a *P*-value less than 0.003, which is the probability of observing a value at least 2.75 standard deviations from the mean, "evidence of a particle"—an encouraging result, but not "discovery".

The common significance levels we have used in this book are $\alpha = 0.05$ and $\alpha = 0.01$ which correspond to a value at least 1.645 and 2.33 standard deviations from the mean, respectively. This goes to show you that there is not any one significance level that everyone agrees on. Difference disciplines have different comfort levels with the idea of significance.

10.2 Exercises

Basic Concepts

1. What is the rationale for the *z*-statistic?

2. What are the three key questions to be asked in the hypothesis testing procedure in order to determine which test statistic is appropriate?

3. Describe the distribution of the *z*-test statistic.

4. What are critical values? How do critical values influence the decision rule in the hypothesis testing procedure?

Exercises

5. Determine the critical value(s) of the test statistic for each of the following tests for the population mean when the population standard deviation is known.

a. Left-tailed test, $\alpha = 0.01$

b. Right-tailed test, $\alpha = 0.10$

c. Two-tailed test, $\alpha = 0.05$

6. Determine the critical value(s) of the test statistic for each of the following tests for the population mean when the population standard deviation is known.

a. Left-tailed test, $\alpha = 0.05$

b. Right-tailed test, $\alpha = 0.02$

c. Two-tailed test, $\alpha = 0.08$

7. A random sample of 1000 observations produces a sample mean of 53.5 with a population standard deviation of 5.3. Test the hypothesis that the mean is not equal to 55 at $\alpha = 0.05$.

8. A random sample of 200 observations indicate a sample mean of 4117 with a population standard deviation of 300. Test the hypothesis that the mean is greater than 4100 at $\alpha = 0.01$.

9. The head of the Veterans Administration has been receiving complaints from a Vietnam veterans' organization concerning disability checks. The organization claims that checks are continually late. The checks are supposed to arrive no later than the tenth of each month. The administrator randomly selects 100 disabled veterans and measures the arrival time in relation to the tenth of the month for each check. If the check arrives early, it receives a negative value. For example, if the check arrives on the eighth of the month, it is measured as –2. If the check arrives on the twelfth of the month, it is measured as +2.

- **a.** What statistical measure should you use in your statement of hypothesis?
- **b.** Formulate hypotheses to test the veterans' organization's claim.
- **c.** Suppose in the sample of 100 disabled veterans receiving checks, the average number of days late was 1.2 with a population standard deviation of 1.4. Calculate the test statistic for your hypothesis.
- **d.** If the test is conducted at the 0.05 level, construct the decision rule for the test statistic.
- **e.** Is there overwhelming evidence at the 0.05 level that the checks arrive late?
- **f.** If you are the head of the Veterans Administration, what is your conclusion?

10. Hurricane Andrew swept through southern 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 of $24,000. 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 $27,500 with a population standard deviation of $2400.

- **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.

11. A retail computer store is considering offering a two-year service warranty, instead of its current one-year plan. In order to do this, they must determine the average service costs for their systems in the second year of operation. A committee of technicians, sales, and management staff believe that the average repair cost in the second year should be approximately $50. Seventy-five customers who purchased machines between two and four years earlier are randomly selected. Tracking the service needs of these customers reveals an average service cost of $38 with a population standard deviation of $10.

- **a.** What is the population being studied?
- **b.** What variable is being measured in this problem?
- **c.** What level of measurement does the variable possess?
- **d.** Test the committee's claim at the 0.10 level.
- **e.** What concerns might you have about the data that were collected?

12. In preparation for upcoming wage negotiations with the union, the managers for the Bevel Hardware Company want to establish the time required to assemble a kitchen cabinet. A first line supervisor believes that the job should take 45 minutes on average to complete. A random sample of 125 cabinets has an average assembly time of 47 minutes with a population standard deviation of 10 minutes.

a. Is there overwhelming evidence to contradict the first line supervisor's belief at a 0.05 significance level? Make your conclusion using the *P*-value approach.

b. What is the lowest average assembly time that would allow the union to conclude that the supervisor is incorrect?

13. The Better Business Bureau has received several complaints that a flour company is underfilling its five pound bags of flour. The Bureau randomly selects 750 bags of flour and determines the weight of each bag. The sample average weight of the bags is 4.80 pounds with a population standard deviation of 0.15 pounds.

a. Is there overwhelming evidence at the 0.01 level that the bags are underfilled?

b. What is the lowest average bag weight that would allow the Bureau to conclude that the bags are underfilled?

14. 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 1 cm per day. A random sample of 45 shrubs has an average growth of 0.90 cm per day with a population 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 day?

15. 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. 200 ears of corn are selected at random from a delivery. The sample has a mean of 5.01 ounces and a population standard deviation of 0.30 ounce. Will a test of hypothesis at $\alpha = 0.10$ support South Valley Farms' claim?

16. 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 with a population standard deviation of 0.30 parts per billion. Do the data support the claim that the average pollutants produced by REM Industries are below the mandated level at a 0.01 significance level?

17. 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 population standard deviation of 15 minutes.

a. What is the population being studied?

b. Are the complaints substantiated by the data at $\alpha = 0.10$?

18. The manufacturer of Brand X floor polish is developing a new polish that it hopes will dry faster than the competition's polish. The competition's polish is advertised to have an average drying time of 10 minutes. A random sample of 1000 Brand X polishes has an average drying time of 9.3 minutes with a population standard deviation of 0.5 minute. Based on the data, can the manufacturer conclude that the drying time for Brand X is faster than the competition's brand at a 0.05 significance level?

19. 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 = 0.0935, $\alpha = 0.10$

b. P-value = 0.0311, $\alpha = 0.05$

c. P-value = 0.0545, $\alpha = 0.01$

d. P-value = 0.0489, $\alpha = 0.05$

20. Consider the following 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: $\mu = 15$, 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: $\mu \neq 100$, $z = 1.90$

21. Consider the following 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: $\mu = 10$, 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: $\mu \neq 100$, $z = 2.70$

10.3 Testing a Hypothesis about a Population Mean, σ Unknown

The hypothesis testing strategy from the previous section assumes that the population standard deviation, σ, is known. However, in most instances, the population standard deviation is just as *unknown* as the population mean. Despite the added uncertainty, the general approach to testing a hypothesis is the same provided that it is reasonable to *assume* that the population from which you are sampling is normal. Some of the technical details concerning the distribution of the test statistic change, since the sample standard deviation, s, will be used in place of the population standard deviation, σ. This modification will cause a change in the distribution of the test statistic.

Formula

t-Test Statistic

If the standard deviation of the population is unknown, but the distribution of the population is assumed to be normal, then the test statistic is given by

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

where

n is the sample size,

$\bar{x}$ is the sample mean,

μ_0 is the hypothesized value of the population mean, and

s is the sample standard deviation.

It should be noted that this formula is only valid if $\bar{x}$ is normally distributed. The test statistic has a t-distribution with $n - 1$ degrees of freedom.

SOLUTION

To test Benny's claim, the hypothesis test is carried out with the following hypotheses.

$$H_0: \mu = 25 \text{ mph}$$

$$H_a: \mu > 25 \text{ mph}$$

The resulting test statistic is

$$z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{25.01 - 25}{\frac{0.10}{\sqrt{2000}}} = 4.47.$$

The *P*-value for this test is 0.000004 and suggests that the test statistic is extremely rare, if the null hypothesis is true. Considering that a test statistic this large would result from ordinary sampling variation in only about 4 in a million samples, we should reject the null hypothesis (H_0: $\mu = 25$ mph) in favor of the alternative. With Benny being a "strictly by the book" officer, he would conclude that the speeds are significantly higher than 25 mph in the school zone and the department should allocate extra resources in that area to reduce speeds. However, from a practical perspective, there isn't much difference (other than random variation) between 25 mph and 25.01 mph, and it isn't likely to be detected by radar. So, despite the "statistical significance" of the test, the practical significance is negligible.

It is important to keep the practical significance of the hypothesis test in mind when making conclusions. This is one of the reasons why in the 6-step procedure for hypothesis testing, **Step 6** is to state the conclusion in terms of the original problem. This step may help shed light on the practical significance of the hypothesis test.

10.3 Exercises

Basic Concepts

1. Suppose a null hypothesis were rejected at $\alpha = 0.05$. Would it be rejected at 0.10? Explain.
2. Suppose a null hypothesis were rejected at $\alpha = 0.05$. Would it be rejected at 0.01? Explain.
3. What is a *P*-value?
4. Discuss how *P*-values are used in the test of a hypothesis.
5. Describe the difference between statistical significance and practical significance.
6. Give an example of a situation in which results could be statistically significant but not practically significant.

Exercises

7. Determine the critical value(s) of the test statistic for each of the following tests for the population mean where the population standard deviation is unknown and the assumption of normality is satisfied.
   - **a.** Left-tailed test, $\alpha = 0.01$, $n = 15$
   - **b.** Right-tailed test, $\alpha = 0.10$, $n = 20$
   - **c.** Two-tailed test, $\alpha = 0.05$, $n = 8$

8. Determine the critical value(s) of the test statistic for each of the following tests for the population mean where the population standard deviation is unknown and the assumption of normality is satisfied.

a. Left-tailed test, $\alpha = 0.005$, $n = 12$

b. Right-tailed test, $\alpha = 0.025$, $n = 5$

c. Two-tailed test, $\alpha = 0.10$, $n = 25$

9. Consider the following random sample of size six from a normal population. Based on the sample, perform a hypothesis test to test the claim that the mean of the population is not equal to 10 at $\alpha = 0.05$.

10 15 12 9 11 10

10. Consider the following random sample of size eight from a normal population. Based on the sample, perform a hypothesis test to test the claim that the mean of the population is greater than 100 at $\alpha = 0.05$. Calculate the *P*-value for this hypothesis test.

100 150 120 90 95 110 100 80

11. Consider the following random sample of size seven from a normal population. Based on the sample, perform a hypothesis test to test the claim that the mean of the population is less than 0.5 at $\alpha = 0.10$. Calculate the *P*-value for this hypothesis test.

0.3 0.5 0.4 0.6 0.5 0.4 0.4

12. 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.**?

13. The admitting office at Sisters of Mercy Hospital wants to be able to inform patients of the average level of expenses they can expect per day. Historically, the average has been approximately $1240. The office would like to know if there is evidence of an increase in the average daily billing. Twenty randomly selected patients have an average daily charge of $1491 with a standard deviation of $342.

a. What is the population being studied?

b. Conduct a hypothesis test to determine whether there is evidence that average daily charges have increased at $\alpha = 0.10$.

c. What assumption did you make in performing the test in part **b.**?

14. A supplier has agreed to provide the manager of a large hospital with light bulbs that he claims will last more than 1000 hours. Twenty-five bulbs are randomly selected and tested by the hospital's maintenance department. The sample has an average life of 1099 hours with a standard deviation of 99 hours.

a. Perform a hypothesis test to determine whether the data support the supplier's claim at $\alpha = 0.05$.

b. What assumption did you make in performing the test in part **a.**?

c. What is the *P*-value for the hypothesis test performed in part **a.**?

15. The managers of a large department store wish to test reactions of shoppers to a new in-store video screen which will broadcast continuous information about the store and the items currently on sale. In past promotions, the video production company has indicated that the average shopper watched for five minutes. The managers randomly select 17 shoppers and determine how long they watch the video. The average time is 4.5 minutes with a standard deviation of 2.5 minutes.

a. Perform a hypothesis test to determine whether there is overwhelming evidence to indicate that the shoppers will watch for less than five minutes. Use $\alpha = 0.01$.

b. What assumption did you make in performing the test in part **a.**?

16. A group of local businessmen is thinking about developing land into a shopping mall. To evaluate the desirability of the location, they count the number of shoppers who visit the neighboring shopping center each day. A random sample of 25 days reveals a daily average of 107 shoppers with a standard deviation of 23 shoppers. The businessmen will develop the land if the average number of shoppers per day is more than 100.

a. Based on the sample data, should the businessmen develop the land? Perform a hypothesis test and use $\alpha = 0.10$.

b. What assumption did you make in performing the test in part **a.**?

17. The Dodge Reports are used by many companies in the construction field to estimate the time required to complete various jobs. The company has received several complaints that the time required to install 130 square feet of bathroom tile is greater than the eight hours reported in the current manual. A researcher for Dodge randomly selects 10 construction workers and determines the time required to install 130 square feet of bath tile. The average time required to install the tile for the sample is 8.5 hours with a standard deviation of 1 hour.

a. Use a hypothesis test to determine whether the customers' complaints are substantiated by the data. Use $\alpha = 0.05$.

b. What assumption did you make in performing the test in part **a.**?

18. 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. Video tapes 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?

19. Buckshot Heaven is developing a new shotgun shell that they hope will have a significantly tighter pellet pattern than their competition. Twenty-five shells are tested at fifty yards. The average pellet pattern of the sample was 8.7 inches in diameter with a standard deviation of 2.0 inches. Their competitor advertises that the average pellet pattern of their shells is nine inches.

a. Does the test completed by Buckshot Heaven support the claim that their shell pattern is tighter than the competition at a level of significance of 0.10?

b. What assumption did you make in performing the test in part **a.**?

20. 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, $\alpha = 0.01$

21. Consider the following 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 $\alpha = 0.01$. See the Discovering Technology section at the end of this chapter for instructions on finding exact *P*-values for *t*-statistics.

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: $\mu \neq 1000$, $t = 2.0$, $n = 15$

22. Consider the following small sample 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 $\alpha = 0.05$. See the Discovering Technology section at the end of this chapter for instructions on finding exact *P*-values for *t*-statistics.

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: $\mu \neq 50$, $t = 2.4$, $n = 5$

23. A.C. Bone has developed a duck hunting boot which it claims can remain immersed for more than 12 hours without leaking. Five hundred pairs of the boots are tested and the time until first leakage is measured. The average time until first leakage for the sample is 12.25 hours with a standard deviation of 3.0 hours.

a. Find the *P*-value to test the claim that the average time until first leakage for the hunting boot is more than 12 hours.

b. Does this sample support A.C. Bone's claim at $\alpha = 0.10$?

24. In preparation for upcoming wage negotiations with the union, the managers for the Bevel Hardware Company want to establish the time required to assemble a kitchen cabinet. A first line supervisor believes that the job should take 45 minutes on average to complete. A random sample of 125 cabinets has an average assembly time of 47 minutes with a standard deviation of 10 minutes. Is there overwhelming evidence to contradict the first line supervisor's belief at a 0.05 significance level?

Discuss the statistical and practical significance for this problem.

25. 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 1 cm per day. A random sample of 45 shrubs has an average growth of 0.90 cm per day 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 day?

Discuss the statistical and practical significance for this problem.

26. 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.

27. The managers of a large department store wish to test reactions of shoppers to a new in-store video screen which will broadcast continuous information about the store and the items currently on sale. In past promotions, the video production company has indicated that the average shopper watched for five minutes. The managers randomly select 17 shoppers and determine how long they watch the video. The average time is 4.5 minutes with a standard deviation of 2.5 minutes. Perform a hypothesis test to determine whether there is overwhelming evidence to indicate that the shoppers watch for less than five minutes. Use $\alpha = 0.01$.

Discuss the statistical and practical significance for this problem.

10.4 The Relationship Between Confidence Interval Estimation and Hypothesis Testing

Previously, we discussed interval estimation for the population mean and the population proportion. We know that when estimating the population mean, a $100(1 - \alpha)\%$ confidence interval for μ is given by

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} .$$

In this chapter, we have shown that the two-sided hypothesis test about the population mean μ is

$$H_0: \mu = \mu_0$$

$$H_a: \mu \neq \mu_0$$

where μ_0 is some specific value of the population mean.

From the previous chapter, we know that $100(1 - \alpha)\%$ of the confidence intervals will contain μ. Therefore, if we reject the null hypothesis when the confidence interval does not contain the value of μ_0, we will reject the null hypothesis when it is actually true with probability of α. You may recall that α represents the probability of committing a Type I error (i.e., we reject the null hypothesis when the null hypothesis is true). So, when we construct a $100(1 - \alpha)\%$ confidence interval and reject the null hypothesis when the interval does not contain μ_0, this is equivalent to performing a two-tailed hypothesis test using α as the level of the test.

$$H_0: \mu = \$74{,}231$$
$$H_a: \mu \neq \$74{,}231.$$

At the 1% level, a 99% confidence interval for the mean New Yorker salary is obtained by

$$\bar{x} \pm t_{df,\alpha/2} \frac{s}{\sqrt{n}}.$$

We know that the sample size is 3,351, the sample mean, $\bar{x} = \$73,707$, and the sample standard deviation is \$2,583. We also know that $\alpha = 0.01$. Thus, $t_{3350,0.005} = 2.577$. Using these data, we find that the 99% confidence interval is given by $73,707 \pm (2.577)(2,583) / \sqrt{3,351}$. Performing the calculations, you will get a confidence interval of (\$73,592, \$73,822). Note that the hypothesized mean of \$74,231 does not fall in the confidence interval above. Thus, we reject the null hypothesis and conclude that the average salary of residents of New York is significantly different than the average salary of residents of Virginia.

10.4 Exercises

Basic Concepts

1. How can a confidence interval be used to test a hypothesis?

Exercises

2. 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 with a sample standard deviation of 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.

3. 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 with a standard deviation of 3.8 inches. The average height for all six-year-old children is 45.75 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?

4. 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 suggest 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?

5. 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?

6. 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 at five minutes over the last two years. Video tapes 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. Assume the population standard deviation is 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.

7. In example 9.2.2 of Chapter 9, we found the 95% confidence interval for mean μ, the population average completion time for the stage in the production process, in minutes, to be (20.36, 26.54).

$$n = 10,\ \bar{x} = 23.45,\ s = 4.32$$

Conduct a hypothesis test, using a 5% level of significance, to see if the population average completion time for the stage in the production process, in minutes, differs from the following values.

a. The null and the alternate hypotheses are H_0: $\mu = 18.27$ versus H_1: $\mu \neq 18.27$.

b. The null and the alternate hypotheses are H_0: $\mu = 24.96$ versus H_1: $\mu \neq 24.96$.

c. The null and the alternate hypotheses are H_0: $\mu = 29.53$ versus H_1: $\mu \neq 29.53$.

8. The owner of an upscale restaurant in Atlanta, Georgia wanted to study the dining characteristics of her customers. She found that in a random sample of 290 customers, 60 purchased dessert. Find a 98% confidence interval for the proportion of customers who purchased dessert. Use this confidence interval to test if the proportion of customers who purchase dessert differs from 25%. Use a 2% level of significance.

9. 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 breaking strength for a sample of 50 pieces of chain is 5000 pounds. Based on past experience, the breaking strength of metal chains is known to be normally distributed with a standard deviation of 100 pounds. Estimate the actual mean breaking strength of the metal link chain with 99% confidence. Use this confidence interval to test if the mean breaking strength of the metal link chain is different from 5020 pounds. Use a 1% level of significance.

10. 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 rating measurements are given below.

Fat Contents							
13%	12%	14%	17%	15%	16%	18%	15%

a. Assuming that the population distribution of the fat content is approximately normal, construct a 90% confidence interval for the true mean fat content of all the packages of ground beef.

b. Use the confidence interval in part (a) to test if the true mean fat content of all the packages of ground beef differs from 17.24%. Use a 10% level of significance.

11. 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. Assume that the lengths of stay are approximately normally distributed.

a. Construct a 95% confidence interval for the mean length of stay for patients with abdominal surgery.

b. Use the confidence interval in part (a) to test if the mean length of stay for patients having abdominal surgery differs from 5.4 days. Use a 5% level of significance.

10.5 Testing a Hypothesis about a Population Proportion

The topic that we will develop in this section will be a hypothesis testing approach for categorical values (nominal data). The inferences that we will make with these data will concern one population. We will use the information in the sample proportion ($\hat{p}$) to test hypotheses about the population proportion, p.

Testing hypotheses about a population proportion could involve a variety of problems.

- What fraction of a student's grades will be A's?
- What fraction of graduating seniors obtain jobs with starting salaries in excess of $38,000?
- What fraction of products that a company produces are defective?
- What fraction of the voters favor the incumbent in the next election?
- What fraction of the customers who purchase a Ford Focus are extremely satisfied?
- What fraction of the time will a baseball player get a hit?
- What fraction of the time will a drug be successful in treating a specific disorder?

Developing the Test

Testing a hypothesis concerning a population proportion is nearly identical to testing a hypothesis about a population mean. The major changes in the procedure include the use of the population proportion (p) in the formulation of the hypotheses rather than the population mean (μ), and the calculation of the test statistic. Let's try an example.

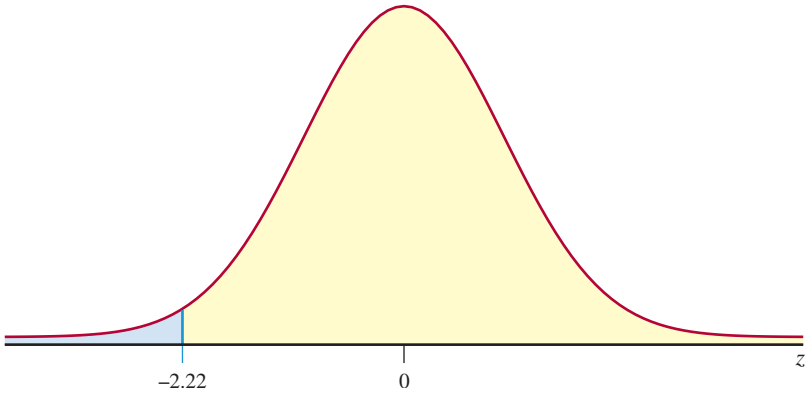


Figure 10.5.3

Assuming the null is true, $\hat{p}$ has a normal distribution that is centered around 0.39. If the null is really true, a z-value of –2.22 is uncommon. The probability of observing a value as small or smaller than –2.22 is the P-value. Using Table A in the Appendix,

$$P\text{-value} = P(z \le -2.22) = 0.0132.$$

Table 10.5.1 – If P-Value = 0.0132

Level of the Test	Reject or Fail to Reject H_0
0.10	Reject
0.05	Reject
0.01	Fail to Reject
0.005	Fail to Reject

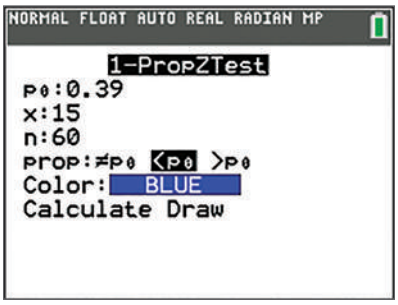
If the null is true, the P-value measures the rareness of the test statistic under ordinary sampling variation. In other words, how often would we see a test statistic as small or smaller than the test statistic we have observed. Presuming the null is true, ordinary sampling variation produces a test statistic less than or equal to –2.22 about 1 time out of every 100. Should H_0 be rejected? Have we observed a test statistic that is too "rare" for H_0 to be true? The level of the test defines an unacceptable level of rareness for the test statistic. In the previous example $\alpha = 0.01$. Setting $\alpha = 0.01$ implies we are only willing to make a Type I error (reject the null hypothesis when it is true) once in every 100 trials of the experiment. Since the P-value of our test statistic, 0.0132, is greater than 0.01, the null hypothesis was not rejected.

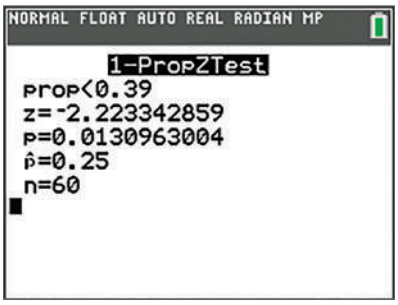
If the level of the test had been 0.05, then to reject H_0 in favor of H_a requires a test statistic whose rareness under ordinary sampling variation is less than or equal to 0.05. For $\alpha = 0.05$, the null hypothesis is rejected in favor of the alternative, since the test statistic has a P-value (0.0132) less than α. In general, if the P-value is less than or equal to the level of the test, α, then H_0 is rejected in favor of H_a. If the level of the test is less than the P-value, then H_0 is not rejected.

Note that if H_a had required a two-tailed test we would double the single tail area. Thus, for a test statistic of –2.22 and a two-tailed H_a the resulting P-value would be $2(0.0132) = 0.0264$.

Technology

P-values for a z-test can be found using the standard normal tables or from the output from a hypothesis test using the z-test statistic. For instructions on how to conduct a hypothesis test for a proportion, please visit stat.hawkeslearning.com and navigate to **Discovering Business Statistics, Second Edition > Technology Instructions > Hypothesis Testing > One Proportion z-Test.**





10.5 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 38%. The airline recently selected 300 passengers at random and determined that 148 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?
9. The National Center for Drug Abuse is conducting a study to determine if heroin usage among teenagers has changed. Historically, about 1.3 percent of teenagers between the ages of 15 and 19 have used heroin one or more times. In a recent survey of 1824 teenagers, 37 indicated they had used heroin one or more times.
   a. Is there overwhelming evidence of a change in heroin usage among teenagers? Test 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 U.S. 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. Do 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.

12. Grain elevators store hundreds of thousands of bushels of grain each year that are waiting to be processed. It is critical to control the amount of moisture in the grain so that it does not spoil. A large storage facility is deemed to be "in control" if 1% of the grain elevators have a moisture content of 10%. One-hundred fifty grain elevators are randomly selected and the moisture content is measured. Two of the grain elevators sampled have a moisture content in excess of 10%.

a. Is there sufficient evidence for the manager to conclude that the storage facility has a moisture content significantly greater than 1%? Use $\alpha = 0.05$.

b. Do you have any concerns about the sample size? Explain.

13. Electronic circuit boards are randomly selected each day to determine if any of the boards are defective. A random sample of 100 boards from one day's production has four 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.

14. Loch Ness Fish Farm breeds fish for commercial sale. The fish are kept in breeder tanks until more than 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 $\alpha = 0.05$?

15. 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 rocks contain more than trace amounts of the mineral. Eighty rocks are randomly selected and the amount of gypsum is measured. Thirteen rocks in the sample 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 $\alpha = 0.01$.

16. 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.

17. 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 $\alpha = 0.01$?

18. 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. Do the data suggest that more than 80% of the customers prefer the curry with the new spice at $\alpha = 0.05$?

19. 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 more than 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?

20. 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 $\alpha = 0.05$?

21. In order to discourage soldiers from smoking, the Pentagon raised the price of cigarettes by $4 a carton in October of 1996. This increased the average price of a carton of brand-name cigarettes to $17.50, an increase of about 30%. Prior to the price increase, about 32% of military personnel smoked, as opposed to 25% of all adult Americans. Suppose that following the price increase, a random sample of military personnel is selected to determine smoking habits. With $\alpha = 0.05$, can we conclude that the price increase was effective in decreasing the percentage of smokers if 50 of the 200 military personnel sampled smoke?

22. Wearing bright or fluorescent orange colored clothing clearly reduces the risk of being shot or killed by hunting. According to an October 1996 article appearing in *The Augusta Chronicle* (Georgia), about two-thirds of the hunters shot in Georgia and South Carolina during the preceding five years were not wearing bright clothes. Of the 52 that were killed, only 19 wore orange. Suppose that a random sample of 100 hunters in Georgia are surveyed and it is determined that of the 100, 62 routinely wear fluorescent orange colored clothing while hunting. With $\alpha = 0.10$, can it be concluded that over half the hunters in Georgia routinely wear fluorescent orange colored clothing while hunting?

23. According to the Federal Communications Commission, about 49% of the households in the United States had cable television in 1985. Suppose that a sample of 200 households is selected in 2003 and it is determined that 125 of them have cable television.

a. With $\alpha = 0.05$, can it be concluded that a higher proportion of households in 2003 have cable television as compared with 1985?

b. In the sample of 200, what is the fewest number of people who have cable television that would allow the conclusion that a higher proportion of households in 2003 have cable television as compared to 1985?

24. Selling autographed sports memorabilia has become a multimillion dollar industry in the United States. But just how does the purchaser of an autographed football jersey or an autographed baseball know that the autograph is indeed authentic? Unfortunately, the sports memorabilia market is teeming with con artists who prey upon the trusting nature of sports fans. In 1996, the FBI said that 70% of all autographed sports memorabilia is fraudulent. Assume that 50 pieces of autographed sports memorabilia are sampled at a large memorabilia show and that 40 of them are determined to be fraudulent.

a. With $\alpha = 0.05$, can it be concluded that the proportion of fraudulent autographed sports memorabilia at the show differs from the FBI claim?

b. What is the greatest number of fraudulent items in the sample of 50 that would not allow a conclusion that sports memorabilia at the show differs from the FBI claim?

Friedrich Robert Helmert

Friedrich Robert Helmert was born in Germany in 1843. His interests were in geodesy, which is a discipline concerned with measuring the earth on a global scale. He studied engineering science at the Polytechnische Schule and while still a student had the opportunity to work on some important geodesy projects with one of his teachers, August Nagel. He later studied mathematics and astronomy to earn his doctorate. Geodesy led him into statistics, first writing a book on least squares. In 1876 he discovered the chi-square as the distribution of the sample variance for a normal distribution. His work was in German and was not translated to English, so later in 1900 English statisticians rediscovered the chi-square distribution (Karl Pearson) and its application to the sample variance (William Gosset, Ronald Fisher).

10.6 Testing a Hypothesis about a Population Variance

In this section we want to adapt the hypothesis testing procedure to test a hypothesis concerning a population variance. Before we can perform the test about the population variance, we need to review a few topics that were discussed earlier in the text.

Recall that the sample variance is

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

and it serves as the point estimate of the population variance, σ^2. In Section 9.4 we determined that the sampling distribution of

$$\frac{(n-1)s^2}{\sigma^2}$$

is a chi-square distribution with $n - 1$ degrees of freedom.

Formula

χ^2-Test Statistic

If we have a random sample of size n taken from a normal population, then the sampling distribution of the test statistic is given by

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

which has a chi-square distribution with $n - 1$ degrees of freedom.

To adapt the hypothesis testing procedure to test a hypothesis concerning a population variance, let's look at the following example.

P-values for χ^2 Test Statistics

Similar to the *t*-Test in Section 10.2, we have numerous chi-square distributions, one for each degree of freedom. The chi-square table is constructed such that it provides us with chi-square values for frequently used tail probabilities. The chi-square table in Appendix A, Table G, has only ten tail areas. Thus, in most instances, we will not be able to determine the exact *P*-value. Instead, we will find the closest chi-square values with the appropriate degrees of freedom surrounding the test statistic. For example, in Example 10.6.1, the value of the test statistic was

$$\chi^2 = 56.84$$

which has 29 degrees of freedom. Since the *P*-value is a function of the alternative hypothesis, we write

$$P\text{-value} = P(\chi^2 > 56.84)$$

As stated earlier, we are unable to find the exact probability for the *P*-value so we will need to put bounds on the *P*-value. The excerpt from the chi-square table above Figure 10.6.3 highlights the chi-square values with 29 degrees of freedom. Note that the chi-square table provides the value of the chi-square with the area of α to the right. Also, note that the values of α across the top row decrease from left to right and the chi-square values in the body of the table increase from left to right. At 29 degrees of freedom, we see that the value of the test statistic falls to the right (i.e., it is greater than) of 52.336 at 29 degrees of freedom corresponding to $\alpha = 0.005$. In this case, we report that the *P*-value is less than 0.005. In Example 10.6.1, we conducted the test using $\alpha = 0.01$. Therefore, since the *P*-value is less than 0.005, it is obviously less than 0.01 which would lead us to reject the null hypothesis and conclude that the variance is significantly more than 0.01 mg.

The exact *P*-value can be found using technology. In Example 10.6.1, using technology we obtain a *P*-value of 0.0015, which is less than $\alpha = 0.01$. Again, this leads us to reject the null hypothesis.

10.6 Exercises

Basic Concepts

1. How does testing a hypothesis about a variance differ from testing a hypothesis about a mean?

2. What is the symbol for a critical value for the chi-square distribution? Describe the meaning of this critical value.

Exercises

3. 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, $\alpha = 0.005$, $n = 5$

4. 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, $\alpha = 0.05$, $n = 41$

5. A bolt manufacturer is very concerned about the consistency with which his machines produce bolts that are $\frac{3}{4}$ inch in diameter. When the manufacturing process is working normally the standard deviation of the bolt diameter is 0.05 inch. A random sample of 30 bolts has an average diameter of 0.25 inch with a standard deviation of 0.07 inch.

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.**?

6. 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. Do 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.**?

7. 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. Can the investor conclude that the standard deviation of share price of the bond fund is greater than 0.25? Test at the 0.01 level.

b. What assumption did you make about the share price of the bond fund in your test in part **a.**?

11.1 Exercises

Basic Concepts

1. What questions are we interested in answering when comparing two population means?
2. What is an independent experimental design?
3. Which sampling distribution do we use in the formulation of the test statistic when comparing two population means with population variances known? What are the properties of this distribution?
4. Does the determination of the critical value(s) for two-sample hypothesis tests differ from one-sample hypothesis tests?
5. What conditions are necessary to perform a test for the difference between two population means?

Exercises

6. Determine the critical value(s) of the test statistic for each of the following tests for the comparison of two population means where the population standard deviations are known.
   a. Left-tailed test, $\alpha = 0.05$
   b. Right-tailed test, $\alpha = 0.10$
   c. Two-tailed test, $\alpha = 0.01$
7. Determine the critical value(s) of the test statistic for each of the following tests for the comparison of two population means where the population standard deviations are known.
   a. Left-tailed test, $\alpha = 0.04$
   b. Right-tailed test, $\alpha = 0.08$
   c. Two-tailed test, $\alpha = 0.02$
8. 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.The results of the study are as follows.

Income (Thousands of Dollars)			
	n	$\bar{x}$	σ
North Side	35	50	10
South Side	40	43	5

   a. Calculate a 90% confidence interval for the difference in average income between the north and south sides of town. 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 $\alpha = 0.05$.

9. 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 salesmen, Mr. Ellis and Mr. Ford. The auditor randomly selects 45 days and determines the daily expenses for each of the salesmen.

Expenses (Dollars)			
	n	$\bar{x}$	σ
Mr. 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 Mr. Ellis and Mr. Ford. 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 salesmen? 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.**

10. 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

Military Training Programs			
	n	$\bar{x}$	σ
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. 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 $\alpha = 0.01$.

11. Tom Sealack, a supply clerk 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.

Battery Life (Hours)			
	n	$\bar{x}$	σ
New Battery	35	700	30
Old Battery	35	710	35

a. Do 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 $\alpha = 0.05$?

12. 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.

Checking Account Balances			
Age Group	n	$\bar{x}$	σ
30 – 39	200	\$2500	\$550
40 – 49	150	\$3500	\$950

a. Do 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 $\alpha = 0.10$?

11.2 Comparing Two Population Means, σ_1 and σ_2 Unknown

It is still possible to make comparisons between two population means if the population standard deviations are unknown.

There are several assumptions that must be met which are outlined below.

Assumptions

Assumptions for Inferences about $\mu_1 - \mu_2$ when the Population Standard Deviations are Unknown

1. An independent experimental design is used.
2. Both populations of interest are approximately normal.
3. Both of the populations have approximately equal (but unknown) variances, $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

In order to determine if both populations of interest are approximately normal, it is helpful to draw histograms of the sample observations from each population. If these histograms appear to be approximately normal, then it is reasonable to infer this assumption is satisfied. With limited data, it is sometimes difficult to determine if the sample data are from a normal population. In these situations, you may have to assume normality and recognize that your inferences are predicated on the validity of the assumption. Figure 11.2.1 shows three histograms of sample data drawn from normal populations.

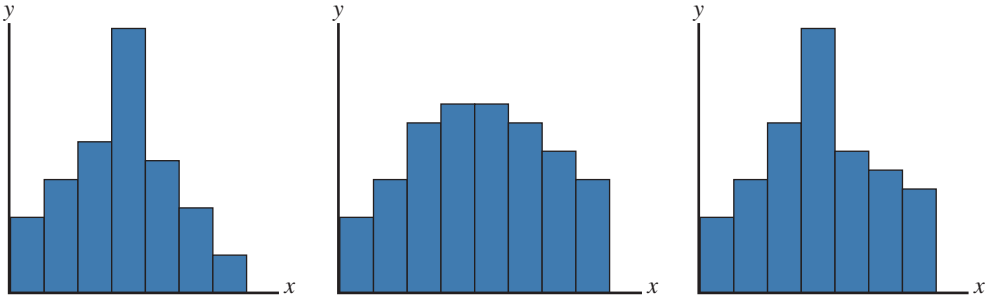


Figure 11.2.1

Step 4: Determine the critical value of the test statistic.

Since we have a two-tailed alternative hypothesis and a significance level of 5% (i.e., $\alpha = 0.05$), the critical value will be $t_{\alpha/2,df}$ where the degrees of freedom is calculated as follows.

$$df = \frac{\left(\frac{15.57}{10} + \frac{3.07}{10}\right)^2}{\frac{1}{10-1}\left(\frac{15.57}{10}\right)^2 + \frac{1}{10-1}\left(\frac{3.07}{10}\right)^2} \approx 12.42 .$$

We will use 12 degrees of freedom for this t-test which yields a critical value of $t_{0.05/2,12} = t_{0.025,12} = 2.179$. We will reject the null hypothesis if the test statistic is greater than or equal to 2.179 or if the test statistic is less than or equal to –2.179.

Step 5: Collect the sample data and compute the value of the test statistic.

Using the values in the table above, the test statistic is

$$t = \frac{(27.7 - 22.8) - 0}{\sqrt{\frac{15.57}{10} + \frac{3.07}{10}}} \approx 3.59 .$$

Step 6: Make the decision and state the conclusion in terms of the original question.

The critical values of the test statistic are ±2.179. Thus, since the test statistic, t, is greater than 2.179 we reject the null hypothesis in favor of the alternative.

Using the Technology Instructions for conducting the Two-Sample t-Test, we get P-value = 0.0035. Since we are performing the test using a significance level of 0.05, the P-value approach still leads us to reject the null hypothesis in favor of the alternative.

Conclusion and Interpretation: This indicates that there is evidence to conclude that the average miles per gallon between the Porsche Cayenne Hybrid model and the Porsche Cayenne Gas model are significantly different.

Note that we did not test whether the average miles per gallon for one model was more or less than the other, thus, the conclusion remains consistent with the stated hypotheses.

Technology

For technology instructions to do a two-sample hypothesis test of the means using the t-distribution, visit stat.hawkeslearning.com and navigate to **Discovering Business Statistics, Second Edition > Technology Instructions > Hypothesis Testing > Two-Sample *t*-Test**.

11.2 Exercises

Basic Concepts

1. Why might large samples not be available when attempting to make inferences about two population means?
2. What assumptions are necessary to perform a test for the difference between two population means when the population variances are unknown?
3. What is the test statistic for an hypothesis test about two population means when the population variances are unknown? How does this statistic differ from the test statistic used in Section 11.1?
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 where the assumptions of normality and equal variance have been satisfied.

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, $\alpha = 0.01$, $n_1 = 5$, $n_2 = 7$

6. 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 and equal variance have been satisfied.

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, $\alpha = 0.10$, $n_1 = 15$, $n_2 = 15$

7. *Popular Science* (Vol. 242, No. 3) reported the results of a comparison of several popular minivans. 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 15 minivans of each type were tested and that the sample standard deviation of the times required to accelerate from 0 to 60 for each minivan was 4 seconds. Assume that the population variances are approximately equal.

a. Calculate a 95% confidence interval for the difference in average acceleration time between the two types of minivans. Interpret the interval.

b. Do 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.**?

8. 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.

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.**?

9. 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.

Hourly Wages (Dollars)			
	n	$\bar{x}$	s
City A	20	\$7	\$3
City B	20	\$8	\$2

a. Calculate a 99% confidence interval for the difference in average hourly wage between City A and City B. Interpret the interval.

b. Do 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.**?

10. 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 approximately equal.

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. Do 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.**?

11. *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.

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. Do the data substantiate the claim that the Brand A amplifier has a higher average maximum power output than Brand B at $\alpha = 0.05$?

12. 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.

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.**

13. In 2009 U.S. charitable giving fell 3.6 percent to $303.75 billion for the year. Total charitable contributions from American individuals, corporations, and foundations fell to $303.75 billion from $315.08 billion for 2008. The largest share of contributions went to religious organizations, representing 33 percent of total giving. The next largest shares went to educational organizations, receiving an estimated 13 percent of the total, and foundations, which received 10 percent of the total. Suppose a sample of 6 employees is randomly chosen from a large corporation and their charitable contributions in 2008 and 2009 are determined. The following table gives these amounts (in dollars). Assume that the population variances are approximately equal.

Charitable Contributions ($)	
Giving in 2008	Giving in 2009
232	215
150	125
50	50
400	350
325	210
175	150

Source: Giving USA Foundation, the Center on Philanthropy at Indiana University.

a. Can we conclude with $\alpha = 0.01$ that the average contribution to charity has decreased in this corporation from 2008 to 2009?

b. Give the assumptions for your test.

c. Repeat part **a.**, assuming that the population variances are not equal.

d. Compare the results of part **a.** and part **c.**

11.3 Paired Difference Test

Suppose we are interested in comparing the durability of the soles of two brands of tennis shoes, Spikes and Kickers. One approach to making this comparison is the independent experimental design discussed in Section 11.1. Using this design one may randomly select 10 people to wear the Spikes brand of shoes for six months and then randomly select 10 other people to wear the Kickers brand of shoes for six months. After the six-month period,

Step 6: Make the decision and state the conclusion in terms of the original question.

As shown in Figure 11.3.3, the value of the test statistic falls in the rejection region to the left. The test statistic indicates that the observed average daily sales are more than 9 standard deviations below the hypothesized value of 0. It is highly unlikely that the difference between the observed value and the hypothesized value is due to ordinary sampling variation. Thus, the null hypothesis is rejected at $\alpha = 0.01$.

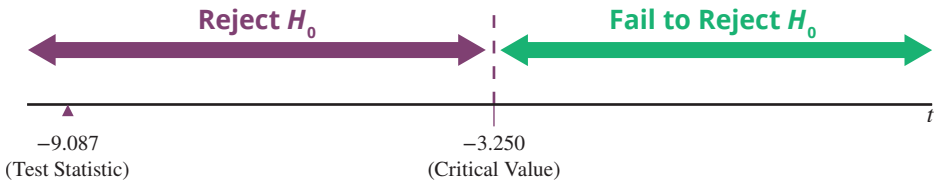


Figure 11.3.3

Technology

For technology instructions on finding the *P*-value for a given *t*-statistic, please visit stat.hawkeslearning.com and navigate to **Discovering Business Statistics, Second Edition > Technology Instructions > *t*-Distribution > *t*-Probability(cdf)**.

To find the *P*-value for this test, we use the same methodology as in Example 11.2.1. The *P*-value is given by

$$P\text{-value} = 2P(t < |-9.087|) = 2P(t > 9.087).$$

We cannot use the *t*-table to find $P(t > 9.087)$; we can only put bounds on the probability. Thus, at 9 degrees of freedom, the largest value in the respective row is 3.250. The best information that we can gain from the *t*-table is that $P(t > 9.087) < 0.005$. Since we need to multiply the probability by 2, we have $P\text{-value} < 0.01$.

Since we are performing the test at $\alpha = 0.01$, we reject the null hypothesis because the *P*-value is less than α—the same decision we made using the rejection region approach.

Conclusion and Interpretation: There is sufficient evidence for the owner to conclude at the $\alpha = 0.01$ level that the average daily sales between the two restaurants are significantly different.

11.3 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, $\alpha = 0.01$, $n_d = 15$

b. Right-tailed test, $\alpha = 0.10$, $n_d = 20$

c. Two-tailed test, $\alpha = 0.05$, $n_d = 8$

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, $\alpha = 0.005$, $n_d = 12$

b. Right-tailed test, $\alpha = 0.025$, $n_d = 5$

c. Two-tailed test, $\alpha = 0.10$, $n_d = 25$

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 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			

Data

This data set can be found on stat.hawkeslearning.com under **Discovering Business Statistics, Second Edition > Data Sets > Textbook Prices**.

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. Do 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.

9. The management for a large grocery store chain would like to determine if a new cash register will enable cashiers to process a larger number of items on average than the cash register 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 cash register and the new cash register. 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 Cash Register	60	70	55	75	62	52	58
New Cash Register	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. Do 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 cash register and the new cash register. Interpret this interval.

e. Can the management conclude that the new cash register will allow cashiers to process a significantly larger number of items on average than the old cash register 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.

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. Do 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 $\alpha = 0.10$.

As displayed in Figure 11.4.2, the value of the test statistic does not fall in the rejection region because $-1.645 < -0.48 < 1.645$. Thus, the difference between the observed value and the hypothesized value is likely due to ordinary sampling variation. We fail to reject the null hypothesis at $\alpha = 0.10$.

Using the *P*-value approach, we want to find $2P(Z < -0.48) = 2(0.3156) = 0.6312$. You may recall that the decision rule when using the *P*-value approach is that we reject the null hypothesis if the *P*-value is less than α. Thus, since the *P*-value is 0.6312 which is greater than 0.10, we fail to reject the null hypothesis.

Conclusion and Interpretation: There is insufficient evidence at $\alpha = 0.10$ for the cell phone executive to conclude that the proportion of defective phones produced differs between the two plants.

11.4 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 large sample tests for the comparison of two population proportions.
   - **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 large sample tests for the comparison of two population proportions.
   - **a.** Left-tailed test, $\alpha = 0.025$
   - **b.** Right-tailed test, $\alpha = 0.02$
   - **c.** Two-tailed test, $\alpha = 0.04$
9. A fund-raiser believes that women are more likely to say "Yes" when asked to donate to a worthy cause than men. To test this theory, she randomly selects 100 men and 95 women and asks for donations to the same cause. The results of the survey are as follows.

Fund-Raiser Survey

	Number Surveyed	# of "Yes" Responses
Men	100	6
Women	95	9

a. Are the sample sizes large enough such that a hypothesis test for the difference between two population proportions may be performed? If so, do 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 men and women who would most likely donate to a worthy cause. Interpret the interval.

10. A poll is conducted to determine if U.S. citizens think that there should be a national health care system in the U.S. 69% of the 300 women surveyed and 63% 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?

11. Major television networks have never seemed to have issues showing commercials for beer and other alcoholic beverages. Even though adult viewers tend to enjoy the commercials, most adults seem to think that the commercials target teenagers and young adults (those under 21 years old). To study this belief, the networks conducted a joint poll of viewers and asked them if they felt that beer and other alcoholic beverage commercials targeted teenagers and young adults. The results of the survey are as follows.

Network Advertising Survey

Age Group	Number Surveyed	Number of "Yes" Responses
30 or Younger	1000	450
Older than 30	1000	655

a. Are the sample sizes large enough such that inferences about the difference between two population proportions can be made? If so, calculate a 99% confidence interval for the difference in the proportions of those older than 30 and those 30 or younger that believe alcoholic beverage commercials targeted teenagers and young adults. Interpret the interval.

b. Based on the data, can the networks conclude that the percentage of viewers who believe beer and alcoholic beverage commercials target teenagers and young adults is significantly higher in the over 30 age group than in the 30 or younger age group at $\alpha = 0.01$?

12. 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.

Number of Defective Parts

	Number Surveyed	Number of Defective Parts
Supplier A	400	8
Supplier B	300	5

Step 5: Collect sample data and compute the value of the test statistic.

The test statistic is given by

$$F = \frac{s_1^2}{s_2^2} = \frac{2500}{900} = 2.7778.$$

This statistic indicates that the variance of revenue for dramas is close to three times that of the variance of the revenue of comedies. Does that indicate that the ratio of the variances is significantly different at the 5% level? We will discuss in **Step 6**.

Step 6: Make the decision and state the conclusion in terms of the original question.

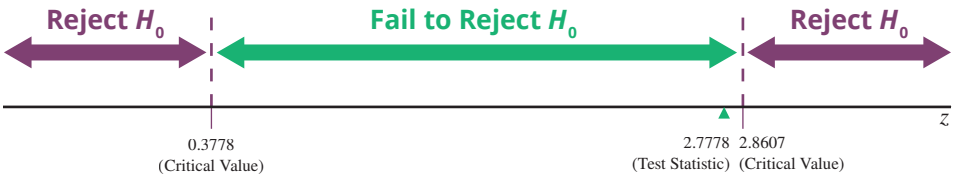


Figure 11.5.7

Since the test statistic does not fall in the rejection region, we fail to reject the null hypothesis.

Using the P-value approach, we want to find $2P(F > 2.7778)$ since this is a two-tailed test. This results in a P-value of 0.0564. Since the P-value is greater than the significance level of 0.05, we fail to reject the null hypothesis.

Conclusion and Interpretation: We conclude that there is insufficient evidence 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.

Technology

To find the P-value for an F-distribution using technology, please visit stat.hawkeslearning.com and navigate to **Discovering Business Statistics, Second Edition > Technology Instructions > *F*-Distribution > *F*-Probability (cdf)**.

11.5 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. $\alpha = 0.100$
   b. $\alpha = 0.050$
   c. $\alpha = 0.025$
   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. $\alpha = 0.100$

b. $\alpha = 0.050$

c. $\alpha = 0.025$

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?

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?

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?

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?

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?

Technology

For instructions on performing an ANOVA test visit stat.hawkeslearning.com and navigate to **Discovering Business Statistics, Second Edition > Technology Instructions > ANOVA > One-Way**.

One way Anova

Summary of Fit

RSquare	0.338598
Adj Rsquare	0.3296
Root Mean Square Error	22.38679
Mean of Response	121.8667
Observations (or Sum Wgts)	150

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-Ratio	Prob > F
Age Group	2	37715.57	18857.8	37.6276	< 0001*
Error	147	73671.76	501.2		
C. Total	149	111387.33			

Means for One way Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
8-12 Years Old	50	128.000	3.1660	121.74	134.26
13-18 Years Old	50	137.480	3.1660	131.22	143.74
Over 18 Years Old	50	100.120	3.1660	93.86	106.38

Std Error uses a pooled estimate of error variance

Figure 12.1.3

Note that the *P*-value is less than 0.0001 which indicates that we would reject the null hypothesis and conclude that the average viewing time between age groups is significantly different. However, from the results of the one-way ANOVA, we only know that the mean viewing time between age groups is different. We will discuss in Section 12.4 some popular multiple comparison procedures that will let us know specifically which of the group means is significantly different.

12.1 Exercises

Basic Concepts

1. What is the price that is paid when pairing experimental units in a paired difference experiment?
2. Give two examples of business situations in which a manager would be interested in comparing several population means.
3. Give an example of a business situation in which a manager would be interested in the average response of a variable that depends on more than one factor.
4. What are experimental units?
5. What is a treatment?
6. Why does simply comparing the sample means for multiple populations not suffice when determining if there is a significant difference in the population means?
7. Explain how box plots can be useful in analyzing data when comparing population means.
8. How is the total variation in the dependent variable broken down in analysis of variance?
9. What does the total sum of squares describe? What are its degrees of freedom?
10. What is the mathematical expression for the sum of squares for treatments?

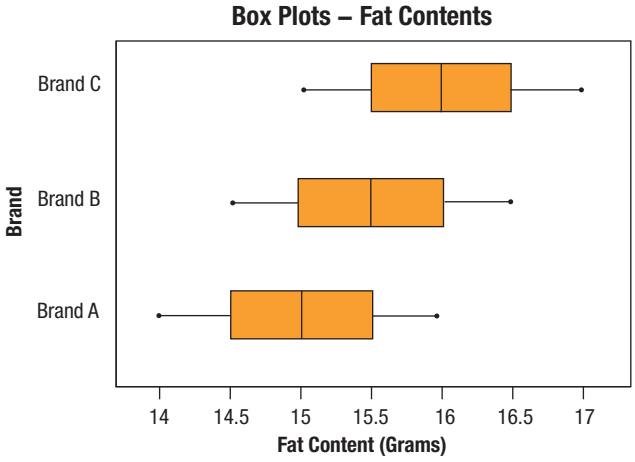
11. What is the grand mean? How is it calculated?

12. What is the mean square for treatments?

13. What is the relationship between TSS, SST, and SSE? Explain why this relationship makes sense.

Exercises

14. Consider the following box plots for data collected to compare the average fat contents (in grams) per serving (2 tablespoons) of three popular brands of peanut butter.



a. Based on the box plots, do you think that there may be a significant difference in the average fat contents per serving of Brand A and Brand B? Explain.

b. Based on the box plots, do you think that there may be a significant difference in the average fat contents per serving of Brand B and Brand C? Explain.

c. Based on the box plots, do you think that there may be a significant difference in the average fat content per serving of Brand A and Brand C? Explain.

15. Consider the following box plots for data collected to compare the average scores achieved on a standardized aptitude test by freshmen, sophomores, juniors, and seniors at a large university.

a. Based on the box plots, do you think that there may be a significant difference in the average scores achieved by freshmen and sophomores on the standardized test? Explain.

b. Based on the box plots, do you think that there may be a significant difference in the average scores achieved by freshmen and juniors on the standardized test? Explain.

c. Based on the box plots, do you think that there may be a significant difference in the average scores achieved by juniors and seniors on the standardized test? Explain.

16. Consider the following table containing yields for mutual funds in different asset classes (small, mid, and large cap).

Fund Yields by Asset Class

Small Cap		Mid Cap		Large Cap	
Fund	**Yield (%)**	**Fund**	**Yield (%)**	**Fund**	**Yield (%)**
Explorer Value	1.13	Capital Value	0.96	Equity Income	3.24
Small-Cap Value Index Admiral	2.46	Mid-Cap Value Index Admiral	2.47	High Dividend Yield Index	3.50
Small-Cap Index Admiral Shares	1.49	Extended Market Index Admiral Shares	1.22	500 Index Admiral Shares	2.35
Strategic Small-Cap Equity	1.10	Mid-Cap Index Admiral Shares	1.52	Diversified Equity	1.23
Explorer	0.17	Mid-Cap Growth	0.10	FTSE Social Index	1.42
Small-Cap Growth Index Admiral	0.21	Mid-Cap Growth Index Admiral	0.32	Growth Equity	0.60

Source: The Vanguard Group, Inc.

a. Identify the experimental units and the treatment in the context of this problem.

b. Compute the mean and median yields for each asset class.

c. Compute the values of the minimum, maximum, first, and third quartiles for each asset class.

d. Construct side-by-side box plots for the three asset classes.

e. Based on the box plots, do you think that there may be a significant difference in the average yields of small-cap and mid-cap funds? Explain.

f. Based on the box plots, do you think that there may be a significant difference in the average yields of mid-cap and large-cap funds? Explain.

g. Based on the box plots, do you think that there may be a significant difference in the average yields of small-cap and large-cap funds? Explain.

h. Based on your analysis, which asset class contains mutual funds with the largest yields, on average? Explain your answer.

17. 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.

18. The sales by strategy data given in Table 12.1.1 yield the following statistics.

Sales by Strategy (Millions of Dollars)		
Strategy 1	Strategy 2	Strategy 3
3	2	4
6	5	2
7	5	5
4	3	6
6	7	6
7	8	7
10	6	9
6	4	8
15	10	14
8	6	8
9	9	7
16	12	16

$$SST \approx 18.0556$$

$$SSE = 438.5$$

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?

c. Find the mean square for treatments, MST.

d. Find the mean square for error, MSE.

19. The fund yield data given in Exercise 16 give the following summary statistics.

Fund Yields by Asset Class

Small Cap		Mid Cap		Large Cap	
Fund	**Yield (%)**	**Fund**	**Yield (%)**	**Fund**	**Yield (%)**
Explorer Value	1.13	Capital Value	0.96	Equity Income	3.24
Small-Cap Value Index Admiral	2.46	Mid-Cap Value Index Admiral	2.47	High Dividend Yield Index	3.50
Small-Cap Index Admiral Shares	1.49	Extended Market Index Admiral Shares	1.22	500 Index Admiral Shares	2.35
Strategic Small-Cap Equity	1.10	Mid-Cap Index Admiral Shares	1.52	Diversified Equity	1.23
Explorer	0.17	Mid-Cap Growth	0.10	FTSE Social Index	1.42
Small-Cap Growth Index Admiral	0.21	Mid-Cap Growth Index Admiral	0.32	Growth Equity	0.60

Source: The Vanguard Group, Inc.

$$\text{MST} \approx 1.8464$$
$$\text{MSE} \approx 0.9423$$

a. Interpret the value of MST.

b. What are the degrees of freedom associated with the sum of squares for treatments?

c. Find the sum of squares for treatments.

d. What are the degrees of freedom for the sum of squares for error?

e. Find the sum of squares for error.

20. Consider the production data given in Exercise 17.

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. What is the value of the grand mean, $\bar{\bar{x}}$?

b. What is the value of n_j?

c. What is the value of k?

d. What is the value of n_T?

e. For these 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.

$$H_0: \frac{\sigma_{Max}^2}{\sigma_{Min}^2} = 1$$

against an alternative of

$$H_a: \frac{\sigma_{Max}^2}{\sigma_{Min}^2} \neq 1.$$

where σ_{Max}^2 represents the largest variance of the three age groups and σ_{Min}^2 represents the smallest variance of the three age groups. The rationale is that if there is not a significant difference between the largest and smallest variances, then there won't be a significant difference among all group variances.

Of course, if there is a significant difference between the largest and smallest variances, then our assumption is violated, and we may need to do one of the following: transform the data by taking the natural log or square root of the responses; use nonparametric procedures; or use some alternative statistics such as Welch's or Brown-Forsythe procedures which use an alternative F-statistic to determine if you have statistical significance.

Understanding that the data are collected from three independent normally distributed populations and that we are testing the ratio of two variances, the test statistic is given by

$$F = \frac{s_{Max}^2}{s_{Min}^2}.$$

Suppose we are testing at the 5% level of significance (i.e., $\alpha = 0.05$). Additionally, we know that we have a two-sided test based on the alternative hypothesis. Since the alternative hypothesis is two-sided, two tails of the F-distribution must be determined as rejection regions. That is, we want to determine if $F \leq F_{1-\alpha/2,df_{\text{num}},df_{\text{den}}}$ or if $F \leq F_{\alpha/2,df_{\text{num}},df_{\text{den}}}$. These critical values are

$$F_{1-\alpha/2,df_{\text{num}},df_{\text{den}}} = F_{0.975,49,49} = 0.5675$$
$$F_{\alpha/2,df_{\text{num}},df_{\text{den}}} = F_{0.025,49,49} = 1.7622$$

The rejection region is that we will reject the null hypothesis if the F-test statistic is less than or equal to 0.5675 or if the F-test statistic is greater than or equal to 1.7622.

As can be seen in the JMP output, the standard deviation (and thus, the variance) is largest for the group of teens (13–18 years old) and the standard deviation is smallest for the adults (more than 18 years old).

The test statistic is given by

$$F = \frac{s_{Max}^2}{s_{Min}^2} = \frac{(25.5704)^2}{(20.3235)^2} \approx 1.5830.$$

Since the test statistic does not fall in the rejection region, we fail to reject the null hypothesis and conclude that there is no evidence to indicate that the variances of screen time among tweens, teens, and adults are significantly different.

Technology

Using the Excel function, F.INV.RT, we can find the critical value $F_{0.975,49,49}$ by typing the following into a cell of the spreadsheet "=F.INV.RT(0.975, 49, 49)" which equals 0.5675.

For instructions on finding F critical values using technology visit stat.hawkeslearning.com and navigate to **Discovering Business Statistics, Second Edition > Technology Instructions > *F*-Distribution > Critical Value**.

12.2 Exercises

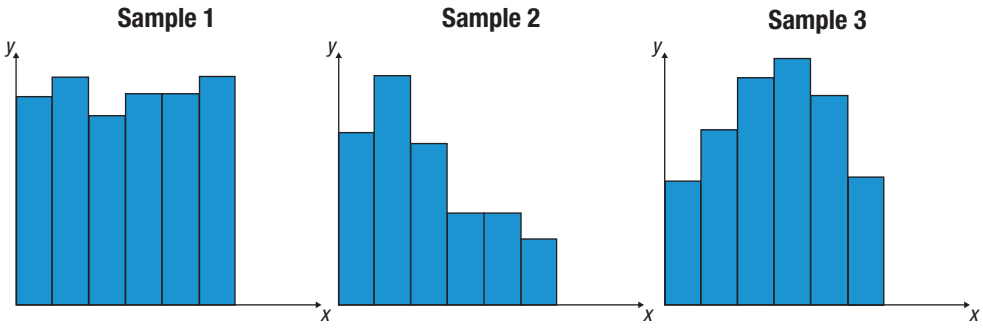
Basic Concepts

1. Why is it important to validate the assumptions upon which a hypothesis test is based?
2. What is the first assumption on which ANOVA is based?
3. How can we test to see if the data reasonably satisfy the first assumption?
4. What is the second assumption on which ANOVA is based?

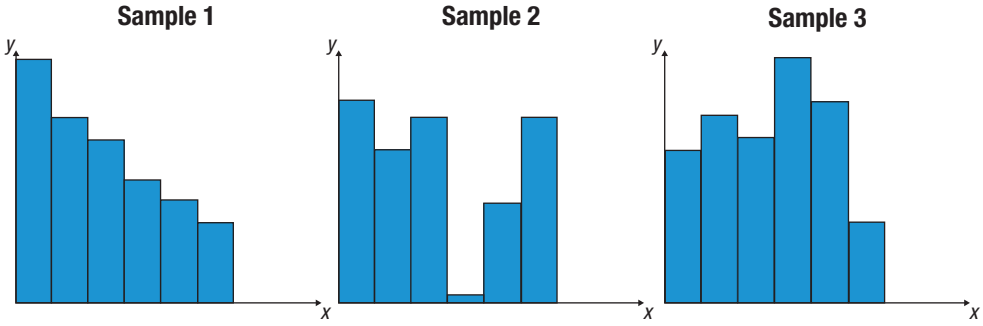
5. How can we determine if the second assumption is reasonable for the data we are interested in?
6. What is a simple "rule of thumb" that may be used to check the second assumption?
7. What is the third assumption that must be met before performing ANOVA?

Exercises

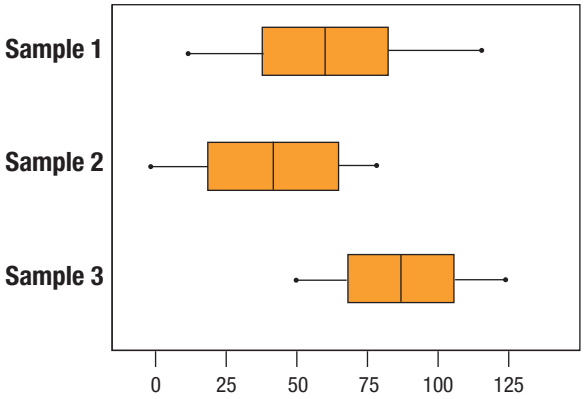
8. For each of the following histograms of sample data, decide whether or not you think it is reasonable to assume that the data were drawn from a population that has an approximately normal distribution.



9. For each of the following histograms of sample data, decide whether or not you think it is reasonable to assume that the data were drawn from a population that has an approximately normal distribution.

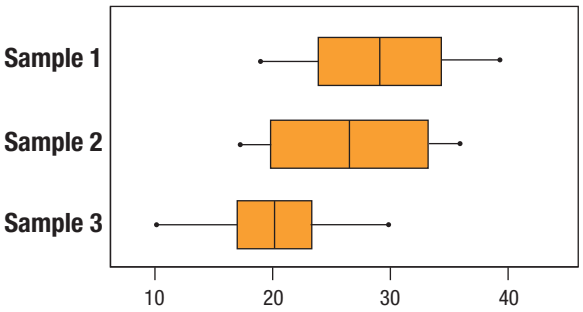


10. Consider the following box plots.



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.

11. 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.

12. Consider the following data on the diameter measurements (in inches) of soft drink bottles of three different brands.

Pepsi	Coca-Cola	Dr. Pepper
1.17	0.14	2.17
1.20	0.07	2.20
1.15	0.11	2.16
1.21	0.08	2.20
1.07	0.21	2.31
1.18	0.22	2.09
1.12	0.15	2.24
1.09	0.08	2.07
1.30	0.12	2.05
1.11	0.28	2.18

Test the assumption of homogeneity required to conduct the ANOVA test, that is, the three samples come from populations with equal variances at a 0.05 significance level.

13. Consider the given data on the foot lengths (in cms) of adult males obtained from three different states of the U.S.

Iowa	Hawaii	California
177.23	172.24	170.28
175.79	172.20	177.46
176.76	176.46	174.67
180.52	180.23	186.57
185.28	171.68	189.29
179.62	177.98	178.63
188.47	179.68	179.91
179.63	177.00	184.72
175.24	179.25	171.52
180.81	181.97	179.08
178.50	182.45	185.49
190.01	175.97	180.24
188.36	176.84	177.64
182.49	169.96	176.85
180.73	167.56	178.72

Do the data satisfy the assumption that the samples come from normally distributed populations? (Use the Shapiro-Wilk test or the Anderson-Darling test.)

14. Four different samples of 10 students each from the same age group are given four types of riddles to solve every week for a period of two months.

The standard deviations of the time to study and solve the riddles for each of the four samples of students are given below.

	Sample 1	Sample 2	Sample 3	Sample 4
Standard deviation (in hours/day)	2.24	1.10	5.49	2.72

Using the "rule of thumb" for equal variances, examine whether the four samples of study hours can be considered to come from populations with the same variance.

15. A survey is conducted among three different age groups of 20 people each at a shopping center located near the researcher's residence. The three samples are from the following age groups: Teen (15-19 years), Young Adult (20-30 years), and Adult (31 years and above). The researcher asks the participants about the amount of money they spend monthly on shopping.

A one-way ANOVA is conducted to test the difference between the mean amounts spent by the respondents for the three different age groups. Consider the assumptions necessary to perform an ANOVA and identify which assumption, if any, is violated in this scenario.

16. An assumption of the ANOVA test is that the k samples considered should come from populations with the same variance. If a boxplot is created for each sample on the same scale, what characteristic should be examined to conclude that the populations have the same variance?

17. Three samples of ten college students were randomly selected from the Japanese club, the Computer Science club, and the soccer team. Each sample was surveyed on their political views of the president. Can the samples in this scenario be considered independent? Explain your answer.

12.3 The *F*-Distribution and the *F*-Test

In Section 12.1, we introduced most of the formulas that are used to analyze data from more than two samples in order to determine whether or not there is a significant difference among population means. We developed MST, a measure for summarizing the variability among the sample means, and MSE, a measure for summarizing the variability within the samples themselves. We determined that if the variability among the sample means is much larger than the variability within the sample observations, we will doubt the hypothesis that the population means are the same. Alternatively, if the variability among the sample means is small when compared to the variability within the sample observations, it is not likely that the population means are significantly different.

Consider the ratio of the MST (mean square for treatments), the summary measure of the variability among the sample means, to the MSE (mean square for error), the summary measure of the variability within the samples.

$$\frac{\text{MST}}{\text{MSE}}$$

The ***F*-distribution**, named after the English statistician Sir Ronald Fisher, is a continuous distribution. It will be used in this and subsequent chapters to analyze variation in test statistics formed as ratios of two random variables. The *F*-distribution is not symmetrical; rather, it is skewed to the right. Like the *t*-distribution, its parameters are degrees of freedom. *F*-distributions are associated with test statistics that are quotients. What distinguishes the *F*-distribution is that it has a pair of values for its degrees of freedom. The number of degrees

12.3 Exercises

Basic Concepts

1. 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.
2. What kind of distribution does the ratio $\frac{\text{MST}}{\text{MSE}}$ have?
3. What are the degrees of freedom associated with $\frac{\text{MST}}{\text{MSE}}$?
4. 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.
5. Is the null hypothesis generally rejected for large or small values of the F-statistic? Explain why this is the case.
6. What are the null and alternative hypotheses for the F-test?
7. What are the assumptions of the F-test?
8. What is the rejection region for the F-test?
9. Can P-values be used to make a decision for the F-test? What is the decision rule?

Exercises

10. 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.**? Do the data appear to satisfy these assumptions? Explain.

11. 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
10	15	13	20
12	16	14	16
11	18	15	18
13	17	15	17
10	14	12	16

a. Do 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.**? Do the data appear to satisfy these assumptions? Explain.

12. A director of training at a large temporary services company has learned of three different methods for teaching a person to type. He is interested in determining if there is a difference in the average typing speeds for employees who are taught to type using each of the three methods. He randomly selects 15 new employees and then randomly assigns five employees to learn to type by each of the training methods. At the end of the course, he measures the number of correct words per minute for each employee. The results are as follows.

Typing Speeds (Correct Words per Minute)		
Method 1	Method 2	Method 3
45	50	60
50	55	63
40	49	55
43	52	52
47	53	58

a. Can the director of training conclude that there is a difference among the average typing speeds of the employees for the three methods at $\alpha = 0.10$?

b. What assumptions did the director of training make in performing the test in part **a.**? Do the data appear to satisfy these assumptions? Explain.

13. 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.**? Do the data appear to satisfy these assumptions? Explain.

14. 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. Do 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.**? Do 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.

15. 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. Do 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.**? Do 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.

12.4 Multiple Comparison Procedures

In the previous sections, 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

12.4 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

LSMeans Differences Tukey HSD

α = 0.050 Q = 2.38063

LSMean[i]	Mean[i]-Mean[j] Std Err Dif Lower CL Dif Upper CL Dif	LSMean[j] 8-12 Years Old	13-18 Years Old	Over 18 Years Old
8-12 Years Old		0	-4.5429	29.4
		0	5.11862	5.11862
		0	-16.728	17.2145
		0	7.64267	41.5855
13-18 Years Old		4.54286	0	33.9429
		5.11862	0	5.11862
		-7.6427	0	21.7573
		16.7284	0	46.1284
Over 18 Years Old		-29.4	-33.943	0
		5.11862	5.11862	0
		-41.586	-46.128	0
		-17.214	-21.757	0

Level		Least Sq Mean
13-18 Years Old	A	136.08571
8-12 Years Old	A	131.54286
Over 18 Years Old	B	102.14286

Levels not connected by same letter are significantly different.

Figure 12.5.5

12.5 Exercises

Basic Concepts

1. What is a completely randomized design? Give an example.
2. Identify a shortcoming that could arise from using a completely randomized design for the example you gave in Exercise 1.
3. What are blocks? What is their purpose?
4. What is a randomized block design? How is it different from a completely randomized design?
5. What are the null and alternative hypotheses when comparing means using a randomized block design?
6. What is the test statistic for the hypothesis test described in Exercise 5?
7. What is the breakdown of the sum of squares for a randomized block design? Does this breakdown make sense? Explain.
8. How are the corresponding degrees of freedom for TSS, SST, SSBL, and SSE related? Verify that this relationship is true.
9. If blocking is successful, how does the value of SSE change?
10. What are the assumptions when performing a two-way ANOVA for a randomized block design?

11. What is the rejection region for the test when performing a two-way ANOVA for a randomized block design?

12. What are the degrees of freedom associated with the test statistic for the two-way ANOVA described in Exercise 11?

Exercises

13. 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. Do you think a randomized block design is appropriate for the car dealer's study? Explain.

b. 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.

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	

Can the dealer conclude that there is a significant difference in average gas mileages of the four car models? Use $\alpha = 0.05$.

c. Was the dealer able to significantly reduce variation among the observed gas mileages by blocking? Use $\alpha = 0.05$.

14. 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, the farmer 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. Do you think a randomized block design is appropriate for the banana grower's study? Explain.

b. The results of the two-way ANOVA for the banana grower's study are given in the following table.

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	

Can the banana grower conclude that there is a significant difference among the average yields of the banana trees for the three fertilizers? Use $\alpha = 0.10$.

c. Was the banana grower able to significantly reduce variation among the observed yields by blocking? Use $\alpha = 0.10$.

15. 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. Do you think a randomized block design is appropriate for the FAA's study? Explain.

b. The results of the two-way ANOVA for the FAA's study are given in the following table.

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	

Can the FAA conclude that there is a significant difference among the average number of on-time arrivals for the four major airlines? Use $\alpha = 0.01$.

c. Was the FAA able to significantly reduce variation among the observed number of on-time arrivals by blocking? Use $\alpha = 0.01$.

16. A psychologist is interested in determining if there is a difference in the average numbers of suicides for several age groups. The psychologist believes that there may be some variation in the numbers of suicides depending on the region of the country (Northeast, Northwest, Southeast, or Southwest). The psychologist randomly selects 100,000 deaths from each region of the country for each of the age groups of interest and determines the number of suicides. The results of the study are as follows.

Suicides							
	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. Do you think a randomized block design is appropriate for the psychologist's study? Explain.

b. The results of the two-way ANOVA for the psychologist's study are given in the following table.

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	

Can the psychologist conclude that there is a significant difference among the average number of suicides for the different age groups? Use $\alpha = 0.10$.

c. Was the psychologist able to significantly reduce variation among the observed number of suicides by blocking? Use $\alpha = 0.05$.

17. In an experiment designed to compare automated blood pressure devices with those of the standard cuff method, each man in a sample of six patients has his systolic blood pressure determined by three different automated devices and by the standard cuff method. The data are given in the following table.

Blood Pressure (mmHg)				
	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. Why was a randomized block design used in this experiment?

b. 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?

c. 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?

12.6 Exercises

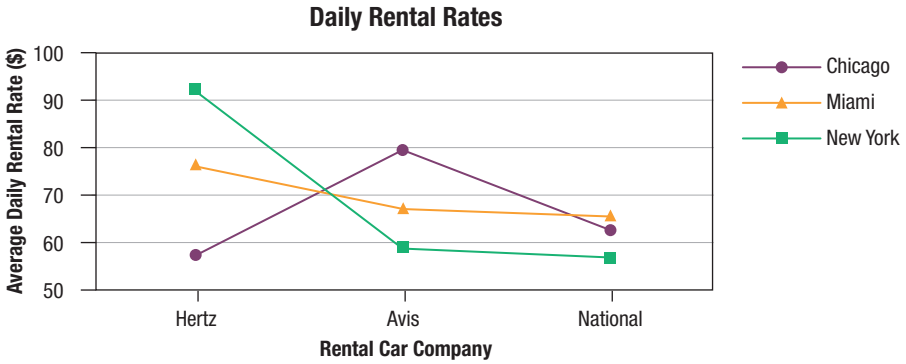
Basic Concepts

1. What is the difference between a randomized block design and a factorial design?
2. What is a complete factorial experiment?
3. What is a profile plot? What kind of information does this plot give us?
4. Why is it so 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. Also, give the acronym associated with each component.
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
Hertz	93.99	54.99	71.99
	90.99	63.99	87.99
	96.99	57.99	68.99
Avis	58.86	81.99	61.99
	52.10	85.99	70.99
	68.98	71.99	66.99
National	56.00	64.99	66.00
	63.00	67.00	58.99
	52.00	52.99	71.99



a. 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?

b. The results of the two-way ANOVA for the study are given in the following table.

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	

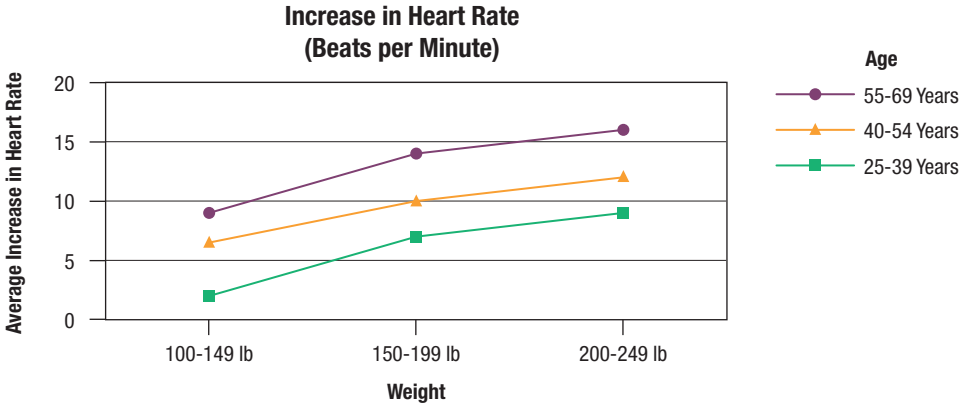
Perform a hypothesis test to determine if there is any interaction between the variables major rental car company and airport location at $\alpha = 0.05$. Does this agree with your observation in part **a.**?

c. If there is no interaction found in part **b.**, 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. 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.



b. The results of the two-way ANOVA for the study are given in the following table.

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	

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 **a.**?

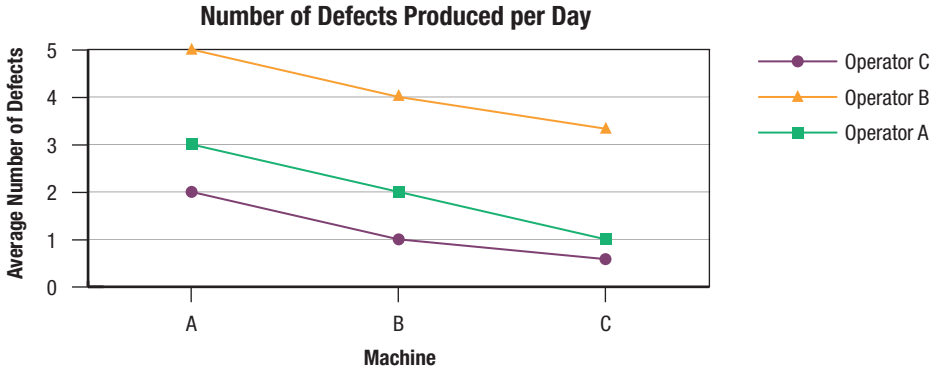
c. Is there sufficient evidence to conclude that there is a significant difference among the average increases in heart rate for the different weight categories? Use $\alpha = 0.01$.

d. 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
Machine A	3	7	3
	3	5	2
	3	3	1
Machine B	2	6	2
	2	4	1
	2	2	0
Machine C	1	5	1
	1	3	0
	1	2	1

a. 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?



b. 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.

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	

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 **a.**?

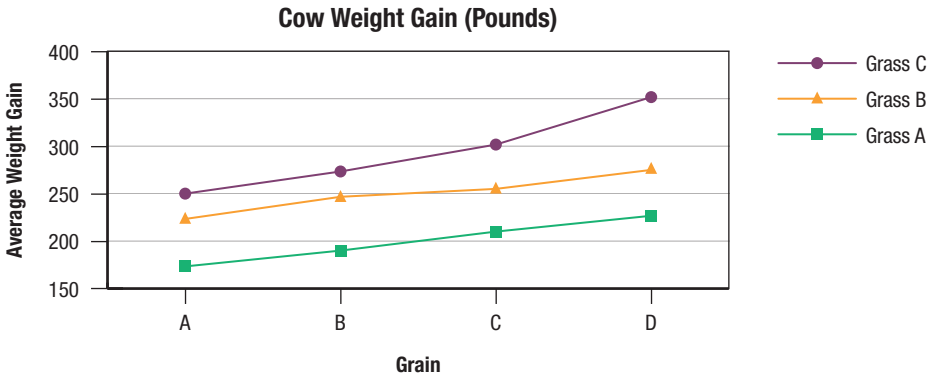
c. Is there sufficient evidence to conclude that there is a significant difference among the average number of defects produced per day for the different machines? Use $\alpha = 0.10$.

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 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**
Grain A	175	225	250
	160	215	240
	185	230	260
Grain B	190	245	275
	185	240	260
	195	255	285
Grain C	210	255	300
	200	245	310
	220	265	295
Grain D	225	275	350
	235	270	360
	220	280	345

a. 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?



b. The results of the two-way ANOVA for the farmer's study are given in the following table.

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	

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 **a.**?

c. If there is no interaction found in part **b.**, 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 $\alpha = 0.05$.

d. 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)." 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. Complete the ANOVA table.

b. Can we conclude, with $\alpha = 0.10$, that there is interaction between power and knowledge?

c. With $\alpha = 0.05$, can we conclude that there is a significant difference in the two levels of power?

d. With $\alpha = 0.05$, can we conclude that there is a significant difference in the two levels of knowledge?

In addition to the formal assumptions previously stated, a linear model should only be used to fit data that appear to be reasonably linear. Because of the wide availability of computer programs that calculate least squares estimates, you will not need to manually calculate estimates very often.

13.1 Exercises

Basic Concepts

1. What is regression analysis?
2. Give two examples of why businesses might be interested in studying the relationship between two variables.
3. What is the difference between a dependent and an independent variable?
4. What is a simple linear regression model? Give the equation that describes a simple linear regression model and define all terms in the equation.
5. What is the estimated simple linear regression equation and how is it used?
6. What is $\hat{y}$? How does this differ from y?
7. What is the technique used to estimate the simple linear regression coefficients?
8. What is the relationship between scatterplots and simple linear regression?
9. Why is it often difficult to accurately describe real world situations using a simple linear regression equation?
10. What is the correlation coefficient? Why is the correlation coefficient insufficient when describing an exact linear relationship between x and y?
11. What is the residual of a model?
12. What is the sum of squared errors and what does it measure?
13. Explain why the best line is referred to as the least squares line.
14. What measure should be minimized in order to find the least squares line?
15. What is the equation for finding the slope of the least squares line?
16. What is the equation for finding the intercept of the least squares line?
17. When finding the least squares line manually, which must be calculated first: the slope or the y-intercept?
18. Interpret the intercept coefficient, b_0.
19. Interpret the slope coefficient, b_1.
20. Why is the magnitude of the prediction errors important when estimating a regression model?
21. What is the mean error for a least squares model?
22. Describe what the magnitude of the variation in the error terms tells us about the reliability of the regression model.
23. What is mean square error?
24. How many degrees of freedom are associated with the error term in a simple linear regression model?
25. What is the square root of the mean square error known as?
26. Describe where the summary statistics for the standard error and mean square error are found in a standard regression summary output in Microsoft Excel.

27. Is there a universal rule on how large is *large* with regard to standard error in a model?

28. What is estimated by the mean square error and what is estimated by the standard error?

29. Why is there an error term incorporated in the simple linear model?

30. What does the error term represent?

31. List the four assumptions about the error term in the simple linear model.

32. List the parameters of the simple linear regression model, and identify their estimates.

Exercises

33. Consider the following simple linear regression model. Write the estimated simple linear regression equation that corresponds to this model.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

34. Consider the following estimated simple linear regression equation.

$$\hat{y}_i = b_0 + b_1 x_i$$

a. What population parameter does b_0 estimate?

b. What population parameter does b_1 estimate?

c. Is error incorporated into the estimated model? Explain.

35. 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.

$$\text{Estimated Sales Volume} = 49.25 + 0.51\left(\text{Advertising Expenditures}\right)$$

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?

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

$$\text{Estimated Salary} = 25689.10 + 2148.35\left(\text{Years of Experience}\right)$$

a. What is the dependent variable?

b. What is the independent variable?

c. What is the value that estimates β_0 in this particular equation?

d. What is the value that estimates β_1 in this particular equation?

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

37. Plot the following lines.

a. $y = 2 + 3x$

b. $y = 4 + 8x$

c. $y = 9 - 2x$

d. $y = x$

38. Plot the following lines.

a. $y = 100 + 50x$

b. $y = 0.5 + 0.7x$

c. $y = 20 - 5x$

39. Consider the following estimated regression equation.

$$\hat{y}_i = 10x_i - 5$$

a. Complete the following table.

Predicted Values	
x	$\hat{y}$
2	
5	
7	
9	
10	

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.

40. Consider the following data.

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

a. Draw a scatterplot of the data.

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

c. Suppose that $\hat{y}_i = 3 + 0.8x_i$ 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 these data?

41. Consider the following data regarding home sale prices and square footage.

Housing Prices and Square Footage	
Selling Price (Thousands of Dollars)	Square Footage
199.9	1065
228.0	1254
235.0	1300
285.0	1577
239.0	1600
293.0	1750
285.0	1800
365.0	1870
295.0	1935
290.0	1948
385.0	2254
505.0	2600
425.0	2800
415.0	3000

a. Suppose we want to predict selling price based on square footage. Write the estimated regression equation in terms of selling price and square footage. (Assume the parameters of this model have not been estimated.)

b. Create a scatterplot of the data and draw a line of best fit.

c. Suppose we determine that an equation that fits the data reasonably well is

$$\text{Estimated Selling Price} = 52.35 + 0.14(\text{Square Footage}).$$

Complete the following table.

Housing Prices and Square Footage				
Observed Selling Price (Thousands of Dollars)	Observed Square Footage	Predicted Selling Price (Thousands of Dollars)	Error	Squared Error
199.9	1065			
228.0	1254			
235.0	1300			
285.0	1577			
239.0	1600			
293.0	1750			
285.0	1800			
365.0	1870			
295.0	1935			
290.0	1948			
385.0	2254			
505.0	2600			
425.0	2800			
415.0	3000			

d. Compute the sum of squared errors for these data.

42. Consider the following data.

x	1	2	3	4	5
y	1	3	4	4	6

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.

43. 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.

44. Comparing the least squares lines in Exercises 42 and 43, which line fits the data better? Explain your answer.

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

$$\text{Estimated Current Value of Home} = 12331 + 3.14\left(\text{Annual Salary}\right)$$

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 $32,000?

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.

46. 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.

$$\text{Estimated Points Scored} = -97.2 + 0.645\left(\text{Minutes Played}\right)$$

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.

47. 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 **b.** 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 **b.**?

d. Use the model created in part **d.** 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?

48. 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 mean square error for these data?

b. What is the standard error of the model?

49. Consider the following data.

Observed Values	
x	*y*
15	110
18	135
25	150
24	149
26	158
40	169

a. Suppose that, using statistical software, we determine that $b_0 = 93.2922$ and $b_1 = 2.1030$. Complete the following table.

Observed versus Predicted Values				
Observed *x*	Observed *y*	Predicted *y*	Error	Squared Error
15	110			
18	135			
25	150			
24	149			
26	158			
40	169			

b. Compute the sum of squared errors.

c. Compute the mean square error.

d. Compute the standard error of the model.

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

50. Consider the following data regarding students' college GPAs and high school GPAs.

GPAs	
College GPA	High School GPA
2.80	3.42
3.54	3.56
2.88	3.13
2.15	3.27
2.22	3.38
3.31	4.13
2.13	3.95
2.39	3.81
3.01	4.33
2.68	2.85

a. Suppose we want to predict college GPA based on high school GPA. Write the estimated regression equation in terms of college GPA and high school GPA. (Assume the parameters of the model have not been estimated.)

b. Suppose we determine, using statistical software, that the estimated regression equation is

$$\text{Estimated College GPA} = 1.88 + 0.2319(\text{High School GPA}).$$

Complete the following table.

GPAs				
Observed College GPA	Observed High School GPA	Predicted College GPA	Error	Squared Error
2.80	3.42			
3.54	3.56			
2.88	3.13			
2.15	3.27			
2.22	3.38			
3.31	4.13			
2.13	3.95			
2.39	3.81			
3.01	4.33			
2.68	2.85			

c. Compute the sum of squared errors for the model.

d. Compute the standard error of the model.

51. The regression equation that relates the delivery time with number of pizzas and distance is given by $\widehat{\text{Delivery Time}} = 1.79 + 1.95(\text{Number of Pizzas}) + 1.57(\text{Distance})$.

a. Estimate the delivery time to deliver 5 pizzas at a distance of 2 miles.

b. The observed data shows that the time taken to deliver 5 pizzas at a distance of 2 miles is 16 minutes. Find the residual.

c. Interpret the meaning of the residual in the context of the problem.

13.2 Residual Analysis

When performing regression analysis, residual analysis is a useful technique to help us determine if the model we are using is appropriate. By studying the estimated errors (i.e., the residuals), we can check the underlying assumptions of the regression model. Before one can adequately make predictions with the estimated regression model, the analyst should ensure that the assumptions of the model are valid. Residual analysis is the method used to validate those assumptions.

All estimates, intervals, and hypothesis tests in regression analysis are based on assuming that the model is correct. If the model is not correct (i.e., at least one of the assumptions is not valid), the formulas and methods will also be incorrect.

Validating the assumptions of the simple linear regression model revolve around the error term (ε). You may recall that the assumptions of the simple linear regression model are:

1. The average response at each value of the independent variable is a linear function. That is, there is a linear relationship between x and y.
2. The errors, ε_i, are assumed to be independent of each other.
3. The errors, ε_i, at each value of x_i are normally distributed.
4. The errors, ε_i, at each value of x_i have equal variances, σ_ε^2.

One method of validating the assumptions is by performing a graphical analysis of the residuals. The most frequently used graph is that of the residuals (e_i) vs. the fitted values ($\hat{y}_i$). It is a scatter plot with the residuals on the vertical axis and the fitted values on the horizontal axis. This plot is used to detect non-linearity, unequal variances, and outliers.

In Figure 13.2.1, (a) is a scatterplot of the raw data, y vs. x, with a simple linear regression line fit through the data. Note that the scatterplot is somewhat curvilinear. When we plot the

that hypothesis tests, confidence intervals, and prediction intervals are sensitive to departures from independence and departures from equal variance. Hypothesis tests and confidence intervals for the slope and intercept are robust against departures from normality. Lastly, prediction intervals are very sensitive to departures from normality.

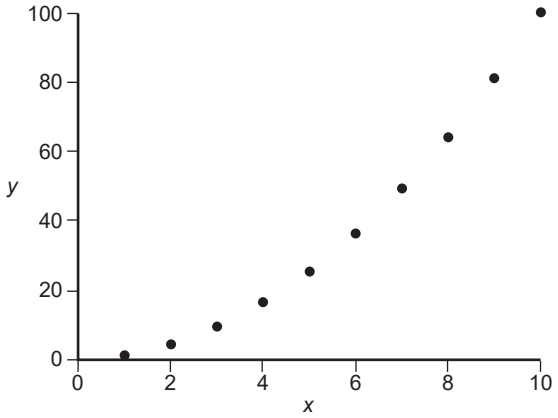
13.2 Exercises

Basic Concepts

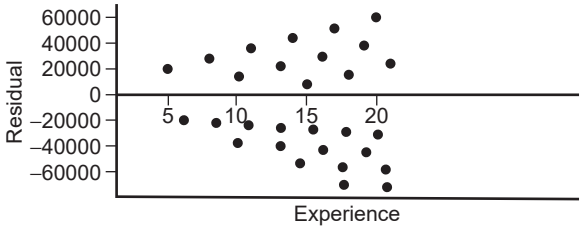
1. What are the assumptions of the simple linear model that need to be validated when doing a residual analysis?
2. What should a well-behaved residual plot look like?
3. List three ways to determine if the errors are normally distributed in a regression analysis.
4. How should a normal probability plot look to indicate normality?

Exercises

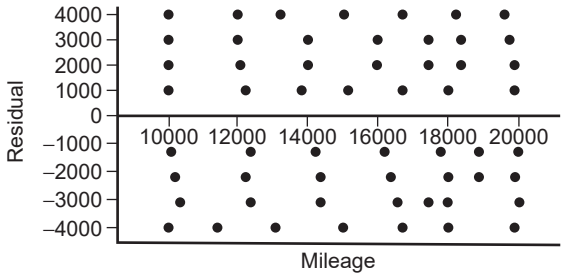
5. A scatterplot of y versus x for a dataset is given below. Which regression assumption is violated in this plot?



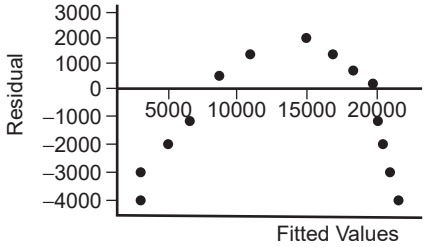
6. Based on the above plot, would you recommend fitting the regression model to predict the response given the predictor? Please explain.
7. A linear regression model was fitted to estimate the salary of an employee based on his/her experience. The plot of residuals of the regression model against the experience is given below. Which regression assumption is violated in this plot?



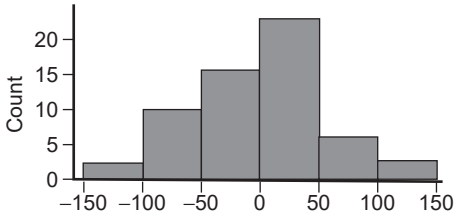
8. A regression model was fitted to predict the price of a used car using the mileage as the predictor. The plot of residuals of the regression model against the mileage is given below. State the regression assumption, if any, violated in this plot.



9. Observe the residuals vs. the fitted plot for the regression model of the price of a car against the age of the car. Is this model appropriate for predicting the price of the car using the age of the car? Explain.



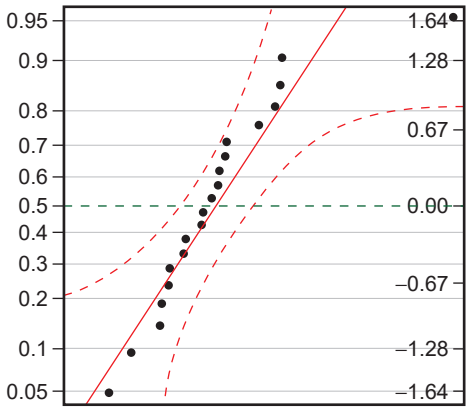
A simple regression model was fitted to estimate the credit score of customers based on their income. The histogram of residuals of the regression model is shown below. Use the histogram to answer the next two exercises.



10. Which assumption of the regression model can be checked using this plot?

11. Based on this plot, what can you say about the validity of the regression model?

A simple regression model was fitted to estimate the price of a used Honda Civic using the mileage as the predictor. The normal probability plot of regression residuals is shown below. Use this to answer the next two exercises.



12. Which assumption of the regression model can be checked using this plot?

13. Based on this plot, what can you say about the validity of the regression model?

14. Download the Pizza Delivery Data, which describes the relationship between Delivery Time (Minutes), the Number of Pizzas delivered, and the Distance (Miles). Use the data to answer the following questions.

a. Create a scatterplot of Delivery Time vs. Number of Pizzas. By examining the scatterplot, do you believe that the data follow a linear pattern?

b. Perform a residual analysis to check the assumptions of linearity, independence, normality, and equal variance. Are any of the assumptions violated? Justify your answer.

Data

The data can be found by visiting stat.hawkeslearning.com and navigating to **Discovering Business Statistics, Second Edition > Data Sets > Pizza Delivery Data**.

15. Download the Marathon Time Data, which has the finishing Marathon Times of 44 runners along with the total number of kilometers they run in training the 4 weeks prior to the race. Use the data to answer the following questions.

a. Create a scatterplot of Marathon Time vs. Km Run in 4 Weeks Prior. By examining the scatterplot, do you believe that the data follow a linear pattern?

b. Perform a residual analysis to check the assumptions of linearity, independence, normality, and equal variance. Are any of the assumptions violated? Justify your answer.

Data

The data can be found by visiting stat.hawkeslearning.com and navigating to **Discovering Business Statistics, Second Edition > Data Sets > Marathon Time**.

13.3 Evaluating the Fit of the Linear Regression Model

The goal in constructing most linear models is to use the independent variable, x, to explain or predict the dependent variable, y. The question we want to consider is, how much of the variation in y can be explained with the model? Before determining how much variation the model explains, it will be necessary to evaluate how much variability exists in the y-variable. This quantity is called the **total sum of squares (TSS)** and represents the total variation in the dependent variable, y.

Formula

Total Sum of Squares (TSS)

The total variation in y is given by the **total sum of squares (TSS)**.

$$\text{TSS} = \sum\left(y_i - \bar{y}\right)^2$$

If you think TSS looks a great deal like the numerator of the formula for the sample variance, you are right. TSS is the sum of the squared deviations about the mean of the dependent variable, y. If TSS were divided by $n - 1$ it would be the sample variance of y.

What is an Error?

An error $\left(y_i - \hat{y}_i\right)$ represents the model's inability to predict the variation in the dependent variable, y. If y didn't vary, for example if all y's were 6, its value would be easy to predict and the model's errors would all be zero. Adding all of the squared errors accumulates the total of all *unexplained* variation.

$$\text{SSE} = \sum\left(y_i - \hat{y}_i\right)^2$$

Because R^2 is a unit-free measure, it can be used to compare the fit of two models. The GPA model only explains approximately 16% of the variation in the dependent variable. Compared to the production model, which had an R^2 of approximately 0.7507, this model seems dramatically inferior. Using R^2 as a criterion, the production model seems to have a substantially better fit (0.7507 versus 0.1597) than the GPA model. The real question is whether you can predict more accurately with the model than other available alternatives. If so, models with relatively low coefficients of determination (such as the GPA model) are useful. For example, if you could develop a model to predict stock prices, minute-by-minute, achieving an R^2 value of only 0.20, you could be a very wealthy person.

R^2 can also be found using the following computational formula.

Technology

For instructions on how to find the coefficient of determination using technology, visit stat.hawkeslearning.com and navigate to **Discovering Business Statistics, Second Edition > Technology Instructions > Regression > Coefficient of Determination.**

Formula

Coefficient of Determination

The **coefficient of determination**, R^2, can be calculated using the equation

$$R^2 = \left(\frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right)\left(n\sum y_i^2 - \left(\sum y_i\right)^2\right)}} \right)^2.$$

Technology

For instructions on how to find the correlation coefficient using technology, visit stat.hawkeslearning.com and navigate to **Discovering Business Statistics, Second Edition > Technology Instructions > Regression > Correlation Coefficient**.

Normally you will not have to use this formula since calculators and computer programs can calculate the coefficient of determination. Recall the computational formula for the correlation coefficient, discussed in Section 4.7, that measures the degree of linear relationship between two variables.

$$r = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}$$

The coefficient of determination is the square of the correlation coefficient. The correlation coefficient can be found by either using the formula given previously or by taking the square root of the coefficient of determination and adding the sign corresponding to the slope coefficient. Remember that the correlation coefficient takes on values between −1 and 1, where negative values indicate a downward sloping relationship and positive values indicate an upward sloping relationship. The coefficient of determination takes on values between 0 and 1, where values close to 0 indicate a weak linear relationship and values close to 1 indicate a strong linear relationship.

13.3 Exercises

Basic Concepts

1. What is the total sum of squares?
2. How are the total sum of squares and the sample variance related?
3. Define error in terms of a regression model.
4. What part of the simple linear regression model captures the unexplained variation?
5. Describe the total sum of squares in terms of explained and unexplained variation.
6. What is the sum of squares of regression?
7. Express SSR in terms of the total sum of squares and the sum of squared errors. Interpret this in terms of model variation.
8. Why will there be errors in virtually all regression models?

9. What is the coefficient of determination? What kinds of values can the coefficient of determination take?

10. Suppose that regression analysis is performed and the resulting model has an R^2 value of 0.856. Interpret this value.

11. How is the coefficient of determination related to the correlation coefficient?

Exercises

12. A direct mail marketing company has been experimenting with the effect of price on sales. Five different direct mail 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.

Direct Mail

Proportion That Purchased Product	Price of Product ($)
0.032	29.95
0.028	34.95
0.026	39.95
0.015	44.95
0.009	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 mean square error.

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

k. Consider exercise 12 parts f and j. Use the information in these two parts to compute the correlation coefficient between the Proportion that Purchased Product and the Price of Product.

13. 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)	Savings (Thousands of Dollars)
28	0.2
25	0
34	0.8
43	1.2
48	3.1
39	2.1
74	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?

14. The Road Warrior Trucking Company has kept careful records on ten hauls. The traffic manager has recorded the haul weight of each truck and its miles per gallon during ten runs with the intent of building a regression model. He wants to predict the miles per gallon for a haul based on the haul weight. The haul weights and miles per gallon information is given in the following table. Haul weights are given in thousands of pounds.

Trucking

Miles per Gallon	Haul Weight (Thousands of Pounds)
4.6	36
4.8	33
5.1	31
4.0	42
4.7	33
5.2	30
4.5	37
4.6	37
4.2	40
4.5	36

a. What is the dependent variable in the model?

b. What is the independent variable in the model?

c. Construct a scatterplot of the data. Based on the scatterplot, does a linear model seem appropriate?

d. Write the model in terms of miles per gallon and haul weight. (Assume the parameters of the model have not been estimated.)

e. Use the data provided and estimate the coefficients of the linear model.f. Interpret the coefficient of the independent variable.

g. Use the model to predict the miles per gallon for a truck hauling 38,000 pounds.

h. Do you believe there is a causal relationship between haul weight and the miles per gallon? If so, which direction is the causality? Do greater haul weights cause reduced mileage, or vice versa? Does the regression analysis prove the causality?

15. An agricultural research station is trying to determine the relationship between the yield of sunflower seeds and the amount of fertilizer applied. To determine the relationship, three different fields were planted. In each field four different plots were defined. In each plot a different amount of fertilizer was used. The plot assignments for the fertilizer application were randomly selected in each field.

Agricultural Research

Pounds of Fertilizer (per Acre)	Pounds of Sunflower Seeds (per Acre)
200	420
200	445
200	405
400	580
400	540
400	550
600	580
600	600
600	610
800	630
800	620
800	626

a. Are the data developed through a controlled experiment or are the data observational?

b. Draw a scatterplot of the data.

c. If a linear model is developed, which of the variables will be the dependent variable? Why?

d. Use the method of least squares to estimate the appropriate model.

e. Interpret the meaning of the slope coefficient in the model.

f. What fraction of the variation in pounds of sunflower seeds per acre can be explained by the amount of fertilizer used?

g. Predict the sunflower seed yield per acre if 500 pounds of fertilizer are applied.

16. 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 (%)
12	4.00
24	4.40
36	5.24
12	3.43
24	4.40
36	5.79
36	5.98
48	6.58
36	5.31
36	5.91

Lengths of Loans and Interest Rates (cont.)

Months Financed	Interest Rate (%)
48	6.51
48	6.68
60	7.13
60	7.48
72	8.31
60	7.85
72	8.07
72	8.48
48	6.12
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?

17. A sample data shows that the correlation coefficient between the number of pizzas and the delivery time is 0.64. If you would fit a regression model for the data to predict the delivery time given the number of pizzas, what percentage of the variation in the delivery time would be explained by the regression model?

13.4 Fitting a Linear Time Trend

Definition

Linear Time Trend Model

A **linear time trend model** is a linear model used to model the changes in some phenomenon over time; the independent variable is always a time index.

In Chapter 4 we discussed the notion that the mean is not a reasonable descriptor for nonstationary time series data. Recall that nonstationary time series do not meander around some central value. Instead, the data tend to get larger or smaller over time. How can you describe time series data that possess a trend? For some time series, a **linear time trend** is a useful model. A linear time trend is nothing but a line that is used to model the changes in some phenomenon measured over time. In a linear time trend model, the independent variable is always a time index. The following example will illustrate the estimation of a **linear time trend model**.

Example 13.4.1

Modeling Data with a Linear Time Trend

Data

This data set can be found on stat.hawkeslearning.com under **Discovering Business Statistics, Second Edition > Data Sets > Tuition Consumer Price Index**.

Many analysts believe that college tuition prices may soon be in the same situation as housing prices were when the housing bubble burst (causing home prices to drop significantly). Table 13.4.1 contains data for the Tuition Consumer Price Index (TCPI) from 1978 to 2020. Use a linear time trend to model the data.

Table 13.4.1 – Tuition Consumer Price Index, 1978-2020

Year	TCPI
1978	59.9
1979	64.7

Linear Fit

TCPI = –41295.44 + 20.854732*Year

Summary of Fit

RSquare	0.961187
RSquare Adj	0.96024
Root Mean Square Error	53.25903
Mean of Response	393.1699
Observations (or Sum Wgts)	43

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-Ratio
Model	1	2880039.2	2880039	1015.341
Error	41	116297.5	2837	Prob > F
C. Total	42	2996336.7		< .0001*

Parameter Estimates

Team	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	–41295.44	1308.338	-31.56	< .0001*
Year	20.854732	0.654483	31.86	< .0001*

Figure 13.4.2

The estimate of the slope, 20.8547, tells us that on average the TCPI is increasing at a rate of 20.8547 per year. Given how well the line fits the data ($R^2 \approx 0.9612$), the trend line is a good descriptor of the data.

The trend line can also be used for short-term prediction. Suppose you wanted to estimate the TCPI for 2021. If the data are not available, the trend model can be used.

$$\text{Estimated TCPI} = -41295.44 + 20.8547(2021) = 851.9087$$

(Prediction of the TCPI for 2021)

One of the problems with this prediction, as with all predictions in this chapter, is that the accuracy of the prediction is unknown. It might be very close to the true value or it could be very inaccurate. If there were some knowledge about the accuracy of the prediction it would be more useful. In later sections, we will return to this topic and study inferential methods.

13.4 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?
5. Identify a problem with predictions that are made using a time trend model.

Exercises

6. Consider the following table containing the Consumer Price Index (CPI) for all urban consumers in the United States from 1990 to 2010. The index is based on 1982–84 prices.

Consumer Price Index			
Year	Consumer Price Index (CPI)	Year	Consumer Price Index (CPI)
1990	130.7	2001	177.1
1991	136.2	2002	179.9
1992	140.3	2003	184.0
1993	144.5	2004	188.9
1994	148.2	2005	195.3
1995	152.4	2006	201.6
1996	156.9	2007	207.34
1997	160.5	2008	215.30
1998	163.0	2009	214.54
1999	166.6	2010	218.06
2000	172.2		

Source: Bureau of Labor Statistics

Data

This data set can be found on stat.hawkeslearning.com under **Discovering Business Statistics, Second Edition > Data Sets > Consumer Price Index**.

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

b. Suppose we are interested in constructing a linear trend model for these 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 CPI.

d. Using statistical software, the following least squares model was determined.

$$\text{Estimated CPI} = -8647.4245 + 4.4107(\text{Year})$$

Use this model to predict the price level in 2015.

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

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

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

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

Write the estimated regression equation.

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?

8. Consider the following table containing unemployment rates for North Carolina and South Carolina in 2000 through 2010.

Unemployment Rates 2000–2010

	Unemployment Rate (%)	
Year	North Carolina	South Carolina
2000	3.7	3.6
2001	5.6	5.2
2002	6.6	6.0
2003	6.5	6.7
2004	5.5	6.8
2005	5.3	6.8
2006	4.8	6.4
2007	4.7	5.6
2008	6.2	6.8
2009	10.8	11.3
2010	10.6	11.2

Source: Bureau of Labor Statistics

a. Using statistical software, estimate the following linear time trend model:

$$\text{N.C. Unemployment Rate} = \beta_0 + \beta_1(\text{Year}) + \varepsilon_i.$$

Write the estimated regression equation using the least squares estimates for β_0 and β_1.

b. Using statistical software, estimate the following linear time trend model:

$$\text{S.C. Unemployment Rate} = \beta_0 + \beta_1(\text{Year}) + \varepsilon_i.$$

Write the estimated regression equation using the least squares estimates for β_0 and β_1.

c. Use the equations in parts **a.** and **b.** to estimate the unemployment rates for North and South Carolina in the year 2013.

d. What is the coefficient of determination for the regression model in part **a.**?

e. What is the coefficient of determination for the regression model in part **b.**?

f. Do you think that these regression models are reliable in predicting future unemployment rates? Of the two models, which seems to fit the data better?

13.5 Inference Concerning the Slope

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 estimates of b_0 and b_1. The confidence interval will serve two purposes: to place bounds on the location of β_1 and to provide information about the quality of the point estimate, b_1. The form of the confidence interval is familiar.

$$\begin{matrix}\text{Sample}\\ \text{estimate of}\\ \text{parameter}\end{matrix} \quad \pm \quad \left(\begin{matrix}\text{A certain number of standard}\\ \text{deviation units depending on}\\ \text{the desired confidence}\end{matrix}\right) \cdot \left(\begin{matrix}\text{The standard}\\ \text{deviation of the}\\ \text{sample estimate}\end{matrix}\right)$$

The sample estimate of β_1 is b_1. The variance of b_1 is given by

$$\sigma_{b_1}^2 = \frac{\sigma_\varepsilon^2}{\sum (x_i - \bar{x})^2}$$

but like all population measurements, $\sigma_{b_1}^2$ usually has to be estimated from the data. Notice that the denominator of the expression above is equal to the variance of x, multiplied by the sample size, n. This indicates that the variance of b_1 is reduced if the variance of the error terms decreases, the sample size increases, or the variance of x increases.

The sample estimate of the variance of b_1 is given by

$$s_{b_1}^2 = \frac{s_e^2}{\sum (x_i - \bar{x})^2}.$$

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_ε^2, with the corresponding sample statistic, s_e^2. The standard deviation (standard error) of the sample estimate b_1 is

$$s_{b_1} = \sqrt{\frac{s_e^2}{\sum (x_i - \bar{x})^2}}.$$

Formula

100(1–α)% Confidence Interval for β_1

The 100(1 – α)% confidence interval for β_1 is given by

$$b_1 \pm t_{\alpha/2,df} s_{b_1},$$

where $t_{\alpha/2,df}$ is the critical value for a t-distribution with $n - 2$ degrees of freedom.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.866441198
R Square	0.75072035
Adjusted R Square	0.719560393
Standard Error	223.0681781
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1198827.203	1198827	24.09247121	0.00118116
Residual	8	398075.2967	49759.41		
Total	9	1596902.5			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2227.955922	370.1487713	6.019082	0.000316596	1374.391324	3081.520519
Items Produced	53.8835069	10.97779658	4.908408	0.00118116	28.5686626	79.1983512

Figure 13.5.4

The *P*-value measures the probability that the test statistic is as large as it is (in magnitude) under the assumption that the null hypothesis is true. Specifically, a *P*-value is the probability of observing a test statistic as large or larger (in absolute value) than what has been observed, given that the null hypothesis is true. In Example 13.5.4, the value of the test statistic was 4.908. The probability of observing a test statistic this large (in absolute value) or larger, given that the true value of the slope is zero, is very small. Fortunately, virtually all statistical analysis programs that perform regression analysis calculate *P*-values for the two-tailed test of hypothesis

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0.$$

Figure 13.5.4 shows the *P*-value of b_1 to be approximately 0.0012. A *P*-value of 0.0012 is persuasive evidence that $\beta_1 \neq 0$. The null hypothesis is rejected if the *P*-value is less than or equal to α. In Example 13.5.4, the significance level of the test was set at $\alpha = 0.05$. Since the *P*-value = $0.0012 \leq 0.05$, the null hypothesis is rejected in favor of the alternative hypothesis ($H_a: \beta_1 \neq 0$).

If the *P*-value is used, how does the procedure for testing a hypothesis change? In **Step 4** all that is necessary is to specify α. It serves as a critical value. In **Step 6**, the *P*-value is compared to α. Everything else remains the same, provided the *P*-value has been calculated for you.

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 the computer output that will be examined.

13.5 Exercises

Basic Concepts

1. Give an example of a practical application of the confidence interval for β_1.
2. Identify two purposes that confidence intervals for the estimated regression coefficients serve.
3. What is the sampling distribution for b_1? Give the mean and standard deviation.
4. What is the expression for determining the $100(1-\alpha)\%$ confidence interval for β_1?

5. Suppose a 95% confidence interval for β_1 is found to be $(15.11,\ 20.11)$. Give two interpretations of this interval.
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 associated with the simple linear regression model?
9. Can we make inferences about β_0? Explain why we are more interested in inferences about β_1.
10. Describe why the P-value corresponding to b_1, which is displayed by many regression summary outputs, is one of the first values examined by data analysts.

Exercises

11. Consider the summary output for the monthly sales data given in Exercise 7 in Section 13.4.

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
Age	40.9090901	4.282219283	9.55324522	2.41268E-06

   a. Compute a 90% confidence interval for β_1.

   b. Interpret this interval.

12. Consider the 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 ($)
1	11,875
1	10,995
2	9995
2	8500
3	8995
4	6995
5	4450
5	5500
6	4400
6	4800

   a. Using statistical software, determine the sample estimate of β_1.

b. What is the standard error of b_1?

c. Find a 99% confidence interval for β_1.

d. Interpret the confidence interval found in part **c.**

13. 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)	IRA Contribution (Thousands of Dollars)
28	0.3
25	0
34	1.0
43	1.3
48	3.3
39	2.2
74	8.5

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.

$$\text{IRA Contribution} = \beta_0 + \beta_1(\text{Income}) + \varepsilon_i$$

c. Calculate and interpret a 95% confidence interval for β_1.

d. What assumptions are being made in the construction of the confidence interval for β_1?

e. Use the confidence interval you obtained to test the hypothesis that the IRA contribution increases with the increase in income of the subject.

14. Consider the following summary output, which was generated from a 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 significance level?

c. Does the decision in part **b.** change at the 0.01 significance level? Explain.

d. What percentage of the variation in annual salary is explained by the model?

15. The college placement office is developing a model to relate grade point average (GPA) to starting salary for liberal arts majors. Ten recent graduates have been randomly selected, and their graduating GPAs and starting salaries were recorded.

GPA and Starting Salary

GPA	Starting Salary (Thousands of Dollars)
2.2	35.1
3.5	45.2
2.1	36.3
2.8	39.3
3.2	41.4
2.5	37.6
2.4	34.8
2.9	25.7
3.1	40.1
3.7	39.5

a. Plot the data. Describe the relationship you observe between GPA and starting salary.

b. Using statistical software, estimate the parameters of the model

$$\text{Starting Salary} = \beta_0 + \beta_1(\text{GPA}) + \varepsilon_i.$$

c. Is there evidence of a linear relationship between GPA and starting salary? Test at the 0.05 significance 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 fraction of the variation in starting salaries is explained by GPA?

g. To perform statistical inference on the model, what assumptions are being made?

16. An experienced census official feels that she can accurately estimate the number of inhabitants of a city block by simply noting the size of the block and the types of buildings (single family homes, apartments, businesses, etc.) that are found on the block. This procedure, if accurate, would be much quicker and cheaper than visiting each residence and taking a survey of its inhabitants. The data below are estimates of block populations provided by the official for 10 blocks in a large city. Also given are the actual numbers of inhabitants for the same 10 blocks. These were found at a later point in time by conventional methods.

Inhabitants of City Blocks

Estimate	115	234	215	97	78	134	78	129	170	67
Actual	100	225	190	99	92	125	75	130	155	82

a. Draw a scatterplot of points of the actual number against the estimated number of inhabitants. Does the relationship appear to be linear?

b. Estimate the slope and intercept of the following regression equation using statistical software.

$$\text{Actual Inhabitants} = \beta_0 + \beta_1(\text{Estimated Inhabitants}) + \varepsilon_i$$

c. Is there evidence of a linear relationship between the actual number and the estimated number? Test at the $\alpha = 0.01$ significance level.

d. Interpret the regression coefficient for the estimated number of inhabitants.

e. Construct a 95% confidence interval for the slope of the regression equation. Interpret the interval.

f. Compute and interpret the R^2 value.

g. Predict the actual number of inhabitants on a block when the estimated number is 150. Round your answer to the nearest whole number.

17. 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. Using statistical software, estimate the parameters of the model

$$\text{Second Test Grade} = \beta_0 + \beta_1\left(\text{First Test Grade}\right) + \varepsilon_i.$$

b. What fraction of the variation in the grades on the second test is explained by the grades on the first test?

c. Is there a linear relationship between the first test grades and the second test grades? Test at the 0.05 significance level.

d. 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.

13.6 Inference Concerning the Model's Prediction

Many regression models are developed for predictive purposes. For example, if you built the model relating the number of items produced to total cost, it was probably because you want to use it to predict total cost. While it is important to evaluate b_1, the estimate of the slope, the real concern of the model builder is the accuracy of the model's predictions. In the case of the production model, how accurate are the costs 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 production data in Table 13.5.1 reveals two weeks in which 30 items were produced. For a given value of items produced (say 30) the costs of producing 30 items were \$3800 and \$3600. For anyone who has observed a production process, price variation is not unexpected. If you use the model

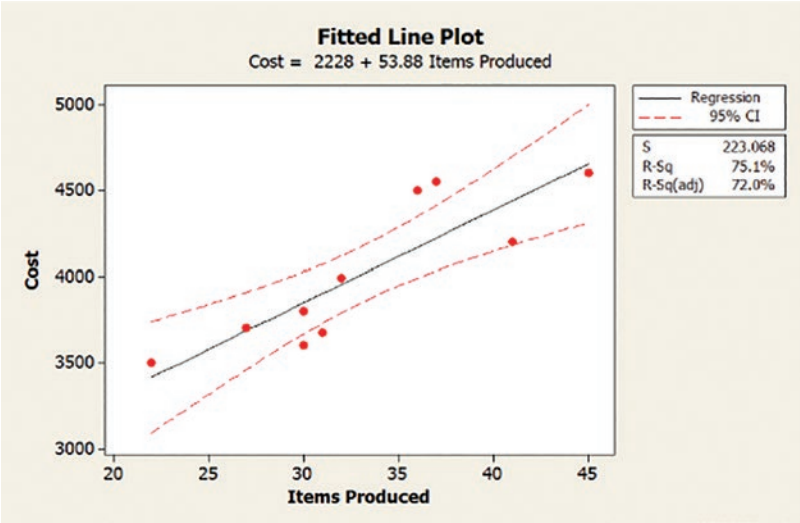


Figure 13.6.4

13.6 Exercises

Basic Concepts

1. What is the main concern for the model builder when performing regression analysis?
2. Distinguish between the mean value of y given x and the predicted value of y given x.
3. Distinguish between a confidence interval and a prediction interval. Which interval is wider? Explain why.
4. Is there a particular range of x-values for which using the regression model for prediction is appropriate? Explain your answer.

Exercises

5. 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:

$$\text{Actual Point Spread} = \beta_0 + \beta_1(\text{Betting Spread}) + \varepsilon_i.$$

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 significance level of a linear relationship between the betting spread and the actual spread?

d. What fraction 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.

6. 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.

$$\text{Net Income} = \beta_0 + \beta_1 \left(\text{Net Sales}\right) + \varepsilon_i$$

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 fraction 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 significance 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.)

7. 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.

$$\text{Sick Days} = \beta_0 + \beta_1(\text{Age}) + \varepsilon_i$$

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 fraction 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 significance 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.

8. A manufacturing company that produces laminate for countertops is interested in studying the relationship between the number of hours of training that an employee receives and the number of defects per countertop produced. Ten employees are randomly selected. The number of hours of training each employee has received is recorded and the number of defects on the most recent countertop produced is determined. The results are as follows.

Training Hours and Countertop Defects

Hours of Training	Defects per Countertop
1	1
4	4
7	0
3	3
2	5
2	4
5	3
5	2
1	5
6	1

A regression analysis has been performed to estimate the model, and the following output is produced.

$$\text{Defects per Countertop} = \beta_0 + \beta_1 \left(\text{Hours of Training}\right) + \varepsilon_i$$

Regression Analysis: Defects per Countertop versus Hours of Training

The regression equation is
Defects per Countertop = 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 defects per countertop 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 fraction of the variation in the number of defects per countertop is explained by the hours of training? What other factors might affect the number of defects?

e. Is there evidence of a linear relationship between the number of hours of training and the number of defects per countertop? Test at the 0.05 significance level and the 0.10 significance level.

f. Construct and interpret a 95% confidence interval for β_1, the slope coefficient.

g. Find the predicted value of the number of defects per countertop for an employee who has had 4 hours of training.

h. Find and interpret the 95% confidence interval for the average number of defects per countertop for employees who have had 4 hours of training.

i. Suppose a new employee has had 4 hours of training. What would be the 95% prediction interval for the number of defects per countertop?

j. Use the model to predict the number of defects per countertop for an employee who has had 7 hours of training. Round your answer to the nearest whole number.

9. Use the following data regarding the age of a particular model of car and the asking price for that car. Construct a confidence interval for the slope to test if there is a significant relation between the age of the car and its price. Use 1% level of significance.

Car Data	
Age (Years)	**Asking Price ($)**
1	11,875
1	10,995
2	9995
2	8500
3	8995
4	6995
5	4450
5	5500
6	4400
6	4800

Interpreting the Coefficients of the Multiple Regression Model

In interpreting the coefficients of the model, we ask the question, *Do the signs and magnitudes of the estimated coefficients appear to be reasonable*?

In the simple linear regression model, the estimated coefficient, b_1, is the slope of the line. It is interpreted to be the average change in the dependent variable associated with a one-unit change in the independent variable. This interpretation remains basically valid for the multiple regression model as well. For the pizza delivery model, the coefficient b_1 is the estimated change in delivery time for a one-unit increase in the number of pizzas, given that distance traveled is constant. Is it reasonable to believe that each additional pizza would add approximately 1.6 minutes to the delivery time? While the delivery time varies, 1.6 minutes seems sensible.

The coefficient b_2 is the estimated change in delivery time for a one-unit increase in distance (measured in miles), given a specific number of pizzas (i.e., the number of pizzas to be delivered is constant). That is, for each additional mile added to the delivery, we should expect the average delivery time to increase by approximately 1.57 minutes. All other conditions being equal, do we find that the signs of the coefficients are reasonable? If we add more pizzas and distance traveled to the delivery route, it seems reasonable to expect the delivery time to increase. Thus, the positive signs on the coefficients seem to make sense. Were the signs of the coefficients unexpected, a reasonable question would be, *Is the estimate accurate?* Are there other factors that have not been considered that could reasonably change the signs of the coefficients? We will consider this question later in Section 14.6.

14.1 Exercises

Basic Concepts

1. Explain why a simple linear regression model might not always suffice when attempting to establish a relationship between variables in a business environment.
2. What is the general multiple regression model?
3. 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?
4. What method is used to find the estimated regression equation? Is this method different from the one used to find the simple linear regression equation?
5. What is the greatest challenge in building a multiple regression model?
6. What are some questions that should be asked once a multiple regression model is estimated? Give at least four.
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?
9. What two aspects of the model coefficients are usually analyzed first when studying a multiple regression model?

Exercises

10. 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?

11. The manager of a publishing company would like to conduct cost analysis on the most recent books the company has published. He 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?

12. 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?

13. 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?

14. Consider the following estimated multiple regression equation relating the number of study hours and GPA to a student's ACT score.

$$\text{Estimated ACT Score} = 8.35 + 1.53(\text{Study Hours}) + 0.30(\text{GPA})$$

- **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 10.

$$\text{Estimated Salary} = 11214.20 + 2854.89(\text{Education}) + 839.64(\text{Experience})$$

- **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 his estimated annual salary?
- **e.** How would you expect his salary to change if he stays 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? How much more money would you expect him to make?

16. Suppose the owner of a car dealership wishes to study how certain factors affect the number of new cars sold. Specifically, he wishes to construct a multiple regression model relating car price, average income per capita in the surrounding area, and the interest rate to the quantity of new cars sold. After compiling historical data, he obtains the following summary output for the multiple regression model. In the original data, car price and average income per capita were in thousands of dollars, and interest rates were reported as percentages.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.581495041
R Square	0.338136482
Adjusted R Square	0.305043307
Standard Error	227.5372802
Observations	64

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	2361116.922	787038.9742	10.21771025	1.57438E-05
Residual	60	4621616.515	77026.94192		
Total	63	6982733.438			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-308.6097834	971.3332832	-0.317717707	0.751802188	-2251.565634	1634.346068
Price	-18.29845156	6.196928308	-2.952826085	0.004489897	-30.69415375	-5.902749358
Income	458.1641011	137.8510094	3.323618036	0.001518198	182.4210273	733.907175
Interest Rate	-26.29956832	11.93107528	-2.204291541	0.031350669	-50.16527221	-2.43386442

a. Write the estimated multiple regression equation.

b. Consider the coefficient of the price variable. Do you think the magnitude and sign of this coefficient make sense? Explain your answer.

c. Consider the coefficient of the income variable. Do you think the magnitude and sign of this coefficient make sense? Explain your answer.

d. Consider the coefficient of the interest rate variable. Do you think the magnitude and sign of this coefficient make sense? Explain your answer.

e. How many additional cars would the dealership owner expect to sell if per capita income for the area increased by $1000?

f. How would the dealership owner expect the quantity of cars sold to change if the interest rate increased by 2 percentage points?

14.2 The Coefficient of Determination and Adjusted R^2

Just as for simple linear regression, we will discuss methods that can be used to evaluate the overall effectiveness of multiple regression models. For the pizza delivery model in the previous section, one of the questions to ask is, how do we determine whether the model explains a substantial portion of the variation in the delivery times? The overall effectiveness and usefulness of multiple regression models can be addressed using the coefficient of determination (R^2) and the adjusted R^2 (R_a^2) statistics.

Coefficient of Determination (R^2)

Recall our discussion about the coefficient of determination (R^2) in the previous chapter. In Section 13.3, we defined the R^2 statistic as the statistic that directly measures the degree to

distance that is driven. The interpretation of R_a^2 is the same—96.08% of the variation in delivery time is explained by the two independent variables in the model.

c. With both number of pizzas and distance in the model, the value of R^2 increased by 0.0374, indicating that adding the variable distance to the model helped explain more variability in delivery times. The value of R_a^2 increased by slightly more (0.0375). Using both variables in the model explained nearly 4% more variability in delivery time.

As the number of independent variables increases, the difference between the R^2 and adjusted R^2 values also increases. 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 is the purpose of the R^2 and adjusted R^2 statistics?

2. What values can the coefficient of determination take?

3. If a particular regression model explains 68% of variation in the dependent variable, what is the value of R^2?

4. If the coefficient of determination has a value of zero, is it possible for a regression coefficient to have a value other than zero? Explain why.

5. If the coefficient of determination has a value of one, what is the relationship between the sum of squares of regression and the total sum of squares? Explain why.

6. Does a large value of R^2 always indicate that the fitted model is useful? Explain.

7. Explain the difference between R^2 and adjusted R^2.

8. Explain why the adjusted R^2 statistic is sometimes a better measure to use to evaluate the fit of a regression model.

9. Will there ever be a situation in which the adjusted R^2 statistic is greater than R^2 statistic? Explain your answer.

Exercises

10. Consider the following ANOVA table for a multiple regression model relating housing prices (in thousands of dollars) to the number of bedrooms in the house and the size of the lot on which the house was built (in square feet). There were 88 total observations.

$$\text{Estimated Price} = 63.26 + 57.31(\text{Bedrooms}) + 0.0029(\text{Lot Size})$$

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	309148.8902	154574.4451	21.58486376	2.62677E-08
Residual	85	608705.618	7161.242565		
Total	87	917854.5083			

a. Identify the values of SSR, SSE, and TSS from the table.

b. What is the coefficient of determination for this model? Interpret this value in terms of the problem.

c. What is R_a^2 ? Interpret this value.

d. 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.

11. Suppose an additional variable, Square Feet, was added to the housing price model from Exercise 10. The summary output is given below.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.819976968
R Square	0.672362228
Adjusted R Square	0.660660879
Standard Error	59.83347988
Observations	88

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	617130.7018	205710.2339	57.46023188	2.69597E-20
Residual	84	300723.8065	3580.045315		
Total	87	917854.5083			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-21.7703086	29.47504196	-0.738601446	0.462207782	-80.38466199	36.84404478
Bedrooms	13.85252186	9.010145446	1.537435988	0.127945059	-4.065140472	31.7701842
Lot Size	0.002067707	0.000642126	3.220095719	0.001822929	0.000790769	0.003344644
Square Feet	0.122778185	0.013237407	9.275092996	1.65802E-14	0.096454149	0.149102222

a. What is R_a^2 for this model?

b. How does the adjusted R^2 value for this model compare to the adjusted R^2 value for the model in Exercise 10?

c. Do you think adding the additional independent variable, Square Feet, improved the model? Explain your answer.

12. 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.

e. How could the restaurant owner improve this model? Are there other independent variables that he should consider including?

13. The owner of the pizzeria discussed in Exercise 12 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 12? 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.

In order to compute the t-value, the degrees of freedom must be determined.

$$df = n - (k + 1) = 25 - (2 + 1) = 22$$

For a 95% confidence interval, $t_{\alpha/2,df}$ will be $t_{0.025,22} = 2.074$. The resulting confidence interval will be

$$1.5891 \pm 2.074(0.1563)$$
$$1.5891 \pm 0.3242$$
$$1.2649 \text{ to } 1.9133$$

We are 95% confident that the true value of β_1, the increase in the delivery time for each additional pizza (given that distance is held constant), will be between 1.2649 and 1.9133 minutes.

Notice that JMP automatically generates confidence intervals for each individual coefficient. The endpoints for the 95% confidence intervals are given in the Lower 95% and Upper 95% columns of the output. Compare the upper and lower limits given by JMP to the ones just calculated by hand. JMP has the ability to calculate confidence intervals for individual coefficients for any level of significance.

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. Why is the R^2 value not used in the test statistic for a hypothesis test to determine if a multiple regression model is significant?

5. 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?

6. Give two equivalent formulas for the test statistic in a hypothesis test about an overall regression model.

7. Explain the significance of the ratio of the mean square regression to the mean square error.

8. 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.

9. How are the degrees of freedom calculated for a multiple regression model?

10. When testing the overall model for significance, do you perform a one or two-tailed test?

11. What is the rejection rule in tests of hypothesis for model significance?

12. What is the expression for a confidence interval for an individual coefficient, β_i?

13. Outline the three pieces of information needed to compute a confidence interval for an individual coefficient.

14. 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?

15. 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.

Exercises

16. An article appearing in the *Journal of Wildlife Management* summarized the percent fat of 75 arctic foxes. According to the authors, "Storage of fat to provide energy during regular periods of food shortage, or to insulate against low ambient temperatures, is essential for survival in severe arctic homeotherms." Computing percent fat was a laborious process. In one analysis, the author regressed y = percent fat on rump fat thickness (RFT), which was measured in millimeters and was much easier to determine than percent fat. It was noted that a plot of percent fat versus RFT was indeed linear, and the resulting regression equation was $y = 7.40 + 1.36(\text{RFT})$. R^2 and s_e^2 were determined to be 0.88 and 3.70, respectively.

ANOVA

	df	*SS*	*MS*	*F*
Regression	k	SSR	SSR/k	MSR/MSE
Residual	$n-(k+1)$	SSE	SSE/$(n-(k+1))$	
Total	$n-1$	SST		

Use the table shown to help answer the following questions.

a. Compute the sum of squares of regression and the sum of squared errors. Note that R^2 is SSR/TSS and s_e^2 is the same as MSE.

b. Give the degrees of freedom for the regression and for the error (residual).

c. Compute MSR and MSE.

d. Compute the F ratio for testing the significance of the regression line. With $\alpha = 0.05$, can we conclude that the relationship between percent fat and RFT is significant?

e. Compute a point estimate for percent fat if RFT = 10 millimeters.

f. For a 2-millimeter increase in RFT, what would be the expected change in percent fat?

17. Consider the model from Exercise 10 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 at the 5% level of significance 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?

18. Consider the printing cost model discussed in Exercise 11 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	133.8204656	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 price 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.

19. The following table contains data from selected cities regarding rental rates of two-bedroom apartments, city populations, and median incomes. 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.

Monthly Rent, Population, and Median Income in Selected Cities

City	Monthly Rent ($)	2010 Population (Thousands)	2010 Median Income (Thousands of Dollars)
Denver, CO	868	600.158	45.438
Birmingham, AL	711	212.237	31.704
San Diego, CA	1414	1307.402	61.962
Gainesville, FL	741	124.354	28.653
Winston-Salem, NC	707	229.617	41.979
Memphis, TN	819	646.889	36.535

Monthly Rent, Population, and Median Income in Selected Cities (cont.)

City	Monthly Rent ($)	2010 Population (Thousands)	2010 Median Income (Thousands of Dollars)
Austin, TX	966	790.390	50.236
Seattle, WA	1219	608.660	58.990
Richmond, VA	735	204.214	37.735
Charleston, SC	812	120.083	47.799
College Park, MD	1407	30.413	66.900
Savannah, GA	789	136.286	33.778
Minneapolis, MN	988	382.578	45.625
Detroit, MI	805	713.777	29.447
Baton Rouge, LA	827	229.493	35.436

Source: U.S. Census Bureau

a. Write the multiple regression model in terms of rent, population, and income. 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.

20. Using the information from Exercise 19, estimate the simple linear regression equation relating monthly rent to median income only.

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 19.

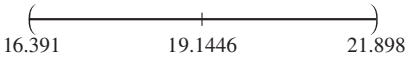
e. Which model do you think is a better model to use to predict monthly rental rates? Explain your answer.

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. Recall the estimated multiple regression equation for our pizza delivery model:

$$\text{Delivery Time} = 1.7929 + 1.5891(\text{Number of Pizzas}) + 1.5677(\text{Distance}).$$

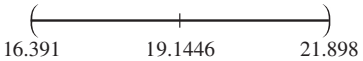
According to the Minitab output given in Figure 14.4.1, the 95% confidence interval for the mean delivery time for 5 pizzas being delivered 6 miles away is 16.391 minutes to 21.898 minutes.



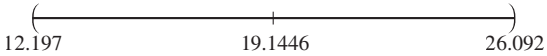
Confidence Interval for the Predicted Value of *y* Given *x*

A caller asks to speak to the manager of the pizza restaurant. The caller wants to know how long it would take to deliver five pizzas to his specific location, which is six miles away. As the manager, you would like to guarantee how long it will take to make this delivery of 5 pizzas to this customer. You are not especially interested in the *average* delivery time for such a delivery. Instead, it would be preferable to create a confidence interval for the time it is going to take for this particular order to be delivered. Once again, for multiple regression, the expression for the prediction interval is beyond the scope of this text. Fortunately, statistical analysis programs such as Minitab will also produce a prediction interval. Using the output shown in Figure 14.4.1, the 95% prediction interval for the delivery of 5 pizzas to a location 6 miles away is 12.197 minutes to 26.092 minutes. As we observed in Section 13.6, to account for individual variation, the prediction interval for y given x is substantially wider than the confidence interval for the mean value of y given x.

Confidence Interval for Average Delivery Time



Prediction Interval for Individual Delivery Time



Can the model make a useful prediction of the delivery time? Although the model has an R^2 of 0.9640, the prediction interval is fairly wide. This indicates that not a great deal of confidence can be placed in the estimated value of 19.1446 minutes as a delivery time for 5 pizzas traveling 6 miles.

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, $\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i}$, 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

6. Consider the multiple regression model predicting graduating GPA using both the SAT critical reading score and the SAT math score. Computer output of the model

$$\text{GPA} = \beta_0 + \beta_1 (\text{SAT Reading}) + \beta_2 (\text{SAT Math}) + \varepsilon_i$$

is given.

Session

Regression Analysis: Graduating GPA versus SAT Critical Reading, SAT Math

The regression equation is
Graduating GPA = 1.35 + 0.00151 SAT Critical Reading + 0.00134 SAT MATH

Predictor	Coef	SE Coef	T	P
Constant	1.3480	0.5760	2.34	0.028
SAT Critical Reading	0.001510	0.001757	0.86	0.399
SAT Math	0.001341	0.001337	1.00	0.326

S = 0.452511 R-Sq = 19.5% R-Sq (adj) = 12.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	1.1869	0.5934	2.90	0.075
Residual Error	24	4.9144	0.2048		
Total	26	6.1013			

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	2.7736	0.1416	(2.4814, 3.0659)	(1.7950, 3.7522)

Values of Predictors for New Observations

New Obs	SAT Critical Reading	SAT Math
1	500	500

a. Find the standard deviation of the error terms in the output.

b. Interpret the coefficient of SAT Critical Reading. What would it mean if the coefficient was negative?

c. Determine if the overall model is useful in explaining GPA. Test at the 0.05 level.

d. What fraction of the variation in GPA is explained by the model?

e. Determine if the SAT Critical Reading variable is a useful predictor of GPA. Test at the 0.05 level.

f. The output includes a predicted GPA for someone scoring 500 on both the SAT Critical Reading and SAT Math portions. Find the predicted value in the output.

g. What is the average GPA for an individual who scored 500 on both the SAT Critical Reading and SAT Math sections? Find the 95% confidence interval for this average. Interpret this interval.

h. Suppose your nephew scored 500 on both the critical reading and 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.

i. Why is the prediction interval so much wider than the confidence interval in part **g.**?

j. 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.

$$\text{Daughter Height} = \beta_0 + \beta_1(\text{Mother Height}) + \beta_2(\text{Father Height}) + \varepsilon_i$$

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, January 2, 2011, 16 games were played in the National Football League. The number of rushing yards, passing yards, first downs, and points for the 32 teams participating in these games is given in the table.

NFL Teams 2011				
Team	**Rushing Yards**	**Passing Yards**	**First Downs**	**Points**
Miami	44	240	16	7
New England	181	321	24	38
Buffalo	37	130	6	7
New York (Jets)	276	119	17	38
Cincinnati	90	305	20	7
Baltimore	98	125	10	13
Pittsburgh	100	325	24	41
Cleveland	43	210	17	9
Oakland	209	160	21	31
Kansas City	115	142	17	10
Minnesota	74	145	16	13
Detroit	107	258	22	20
Carolina	137	182	12	10
Atlanta	99	256	24	31

Data

This data set can be found at stat.hawkeslearning.com under **Discovering Business Statistics, Second Edition > Data Sets > NFL Teams 2011**.

NFL Teams 2011 (cont.)

Team	Rushing Yards	Passing Yards	First Downs	Points
Tampa Bay	84	255	18	23
New Orleans	106	212	20	13
Jacksonville	198	140	23	17
Houston	244	253	22	34
Dallas	159	127	14	14
Philadelphia	121	162	14	13
New York (Giants)	82	243	14	17
Washington	67	336	20	14
San Diego	164	313	20	33
Denver	146	205	18	28
Arizona	78	242	19	7
San Francisco	100	276	16	38
Chicago	110	168	13	3
Green Bay	60	229	14	10
Tennessee	51	300	17	20
Indianapolis	101	264	24	23
Saint Louis	47	155	10	6
Seattle	141	192	19	16

Source: National Football League

a. In order to predict a team's points from rushing yards, passing yards, and first downs, a multiple regression analysis is performed on the data with points as the dependent variable. 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	0.774540403
R Square	0.599912836
Adjusted R Square	0.557046354
Standard Error	7.508433133
Observations	32

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	2366.956093	788.9853643	13.99492	9.23535E-06
Residual	28	1578.543907	56.37656811		
Total	31	3945.5			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-15.6150327	5.982932568	-2.609929582	0.014378
Rushing Yards	0.127133936	0.029615758	4.292780033	0.000191
Passing Yards	0.081411959	0.029307161	2.777886231	0.009655
First Downs	0.121492053	0.460140758	0.264032366	0.793689

b. Find the standard deviation of the error terms in the output.

c. Determine if the overall model is useful in predicting points scored. Use $\alpha = 0.05$.

d. What fraction of the total variation in points is explained by the model?

e. Is the rushing yards variable useful in predicting points scored at the 0.01 level?

f. Is the passing yards variable useful in predicting points scored at the 0.01 level?

g. Is the first downs variable useful in predicting points scored at the 0.01 level?

h. The coefficient of rushing yards in the regression equation is 0.1271. Interpret this value.

i. 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

$$\text{Points} = \beta_0 + \beta_1(\text{Rushing Yards}) + \beta_2(\text{Passing Yards}) + \varepsilon_i.$$

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 63 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.

14.5 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 in business and economics, however, are not quantitative, but qualitative. Examples of qualitative variables are gender (male or female), firm size (small, medium, or large), and type of investment (stock or mutual fund).

Definition

Indicator (or Dummy) Variable

An **indicator (or dummy) variable** is created to assign numerical values to classes of a categorical variable. The dummy variable allows one to use a single variable to represent multiple categories.

In order to use qualitative variables in regression analysis, we need to identify the classes of the qualitative variable quantitatively. To do this, we will use **indicator** (or **dummy**) **variables** that take on values of 0 and 1. If we have a qualitative variable with c classes, that variable will be represented by $c - 1$ indicator variables in the regression model, with each indicator variable taking on a value of 0 or 1.

Suppose we added a variable to our pizza model that asked the customer if they lived in town or out of town. The variable, let's call it town (in or out), has $c = 2$ classes. Thus, the variable town will be represented by $c - 1 = 1$ indicator variable in the model. Town could be modeled as follows.

$$x_3 = \begin{cases} 1 & \text{if In Town} \\ 0 & \text{otherwise} \end{cases}$$

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.

Definition

Polynomial Regression

Polynomial regression models are used when the relationship between the independent variable and the dependent variable are modeled using an n^{th} degree polynomial in the independent variable. For example, the regression model could resemble $y = \beta_0 + \beta_1 x^3 + \varepsilon$.

14.5 Exercises

Basic Concepts

1. Explain why qualitative variables have not been used in the regression models we have discussed in previous sections.

2. Give three examples of qualitative independent variables that may be of interest to someone performing regression analysis to predict annual salary.

3. Explain how qualitative variables are transformed into quantitative variables in order to estimate a regression model.

4. If a qualitative variable has c classes, how many indicator (dummy) variables will there be in the model? Explain why this is the case.

5. When an indicator (dummy) variable is equal to one, does this represent a difference in the slope or the intercept of the model? Explain.

6. What is a base level variable? Interpret the value of an estimated coefficient for an indicator variable in terms of the base level variable.

7. Identify three potential issues to keep in mind when constructing regression models involving indicator variables. Also suggest how these issues can be addressed.

Exercises

8. 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).

$$\text{GPA} = -0.8777 + 0.0672(\text{Attendance}) + 0.0678(\text{Exam Score}) - 0.1436(\text{Freshman})$$

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?

9. 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 GPA using a 5% significance level.

b. What percentage of the variation in GPA is explained by the three independent variables?

c. Is the qualitative independent variable, freshman, useful in predicting GPA? Use $\alpha = 0.05$.

d. Do you believe this is a good model to use to predict GPA? Why or why not?

e. 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.

10. A personnel director is interested in studying the effects which age, experience, and gender have on salary. Eight employees are randomly selected and each employee's salary, age, experience, and gender (= 0 if male, = 1 if female) are recorded.

Employee Data

Salary ($)	Age	Experience (Years)	Gender
27,000	25	2	1
50,000	55	20	0
48,000	27	5	0
35,000	30	7	1
29,000	22	3	1
58,000	33	8	1
23,000	19	1	0
43,000	45	15	1

a. Create three scatterplots using salary with age, salary with experience, and salary with gender. Does each of the plots have a linear relationship?

b. Using statistical software, estimate the parameters of the following regression model:

$$\text{Salary} = \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{Experience}) + \beta_3(\text{Gender}) + \varepsilon_i.$$

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 gender useful in explaining salary? Test at the 0.10 level.

g. Interpret each of the regression coefficients.

h. Construct and interpret 95% confidence intervals for each of the regression coefficients. Do you think that the company discriminates in the salary paid based on gender?

i. Predict the salary of a female employee who is 35 years old with 10 years of experience.

j. Construct and interpret a 95% prediction interval for a female employee who is 35 years old with 10 years of experience. How useful is this interval?

k. Construct and interpret a 95% confidence interval for the average salary of a female employee who is 35 years old with 10 years of experience. How useful is this interval?

Data

This data set can be found at stat.hawkeslearning.com by navigating to **Discovering Business Statistics, Second Edition > Data Sets > Campus Crime**.

11. 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.

Campus Crime Data

Number of Crimes	Number of Police	Total Enrollment	Private School
64	12	1131	Yes
138	21	12,954	No
141	32	16,009	No
84	22	1682	Yes
86	35	2888	Yes
141	45	17,407	No
135	42	3028	Yes
174	50	4306	Yes
201	75	34,511	No
203	84	37,240	No
125	36	2918	Yes
234	109	39,414	No
143	45	4000	Yes
148	50	20,950	No
152	48	4277	Yes
158	52	26,519	No
174	69	27,687	No
84	26	2810	Yes
173	58	27,619	No
193	56	4563	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.

$$\text{Crimes} = \beta_0 + \beta_1(\text{Police}) + \beta_2(\text{Enrollment}) + \beta_3(\text{Private}) + \varepsilon$$

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 $\alpha = 0.01$?

12. 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.

Data

This data set can be found at stat.hawkeslearning.com by navigating to **Discovering Business Statistics, Second Edition > Data Sets > Weekly Sales by Quarter**.

Weekly Sales by Quarter

Weekly Sales ($)	Number of Sales Reps	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
4272.90	3	1	0	0	0
5069.70	9	1	0	0	0
6067.70	11	1	0	0	0
6680.55	17	1	0	0	0
9725.05	20	1	0	0	0
4107.10	3	0	1	0	0
7520.25	9	0	1	0	0
12,135.00	11	0	1	0	0
13,016.55	17	0	1	0	0
13,673.90	20	0	1	0	0
3272.05	3	0	0	1	0
5074.40	9	0	0	1	0
7505.45	11	0	0	1	0
8272.75	17	0	0	1	0
10,020.40	20	0	0	1	0
4925.75	3	0	0	0	1
10,018.10	9	0	0	0	1
12,505.85	11	0	0	0	1
15,329.05	17	0	0	0	1
19,477.20	20	0	0	0	1

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.

$$\text{Sales} = \beta_0 + \beta_1(\text{Reps}) + \beta_2(\text{Quarter 1}) + \beta_3(\text{Quarter 2}) + \beta_4(\text{Quarter 3}) + \varepsilon$$

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?

wasted when estimating unnecessary parameters. Therefore, it would be best to include only important variables in the model or variables that are clearly necessary. A solution to avoid over-fitting a model is to utilize a model-building procedure such as **stepwise regression**. Stepwise regression involves selecting independent variables using an automated procedure, which is beyond the scope of this text.

Some other issues with fitting multiple regression models are **extrapolation** (which was discussed in Chapter 13) and **correlated errors**. Extrapolation can be a concern when the regression model is used to predict values outside the range of the data used to estimate the model. Be sure to only use the model within an appropriate range of x-values. The problem with correlated errors arises when measurements of the dependent variable are correlated. That is, since the observations (the responses) of the regression model are assumed to be independent, it is problematic if there is a relationship between responses. One will often see this type of dependency with time series data. Current measurements are often dependent on measurements in the previous time period.

14.6 Exercises

Basic Concepts

1. Define multicollinearity. How can you detect if multicollinearity exists in a regression model?
2. Why is multicollinearity a concern when performing regression analysis?
3. How can you attempt to correct the problem of multicollinearity?
4. What is parameter estimability?
5. How can you alleviate concerns about parameter estimability?
6. Why is variable selection difficult when building a multiple regression model?
7. Does the R^2 value always increase as additional variables are added? Does this mean that adding additional variables always produces a more useful model? Explain.
8. What is extrapolation? Why is this a concern?
9. With what type of data do you often encounter issues with correlated errors?

Exercises

10. In Exercise 8 of Section 14.4, we modeled the relationship between total points and rushing yards, passing yards, and first downs.

    a. Using the correlation matrix below, discuss whether collinearity might play a role in estimating total points using rushing yards, passing yards, and first downs in the model.

Correlation Matrix			
	Rushing Yards	**Passing Yards**	**First Downs**
Rushing Yards	1.0000	–0.1943	0.3789
Passing Yards		1.0000	0.5744
First Downs			1.0000

    b. How would you determine if there is a relationship (and if so, the strength of such relationship) between the independent variables in the model?

    c. Given the multiple regression model that was fit in Exercise 8, what would the total points be if a team had 30 rushing yards, 100 passing yards, and 5 first downs?

    d. Should you have any concerns about the estimate in part **c.**? Explain your answer.

11. In Exercise 10 of Section 14.5, salary was modeled as a function of age, experience, and gender.

a. Discuss how collinearity might play a role in estimating salary using age, experience, and gender in the model.

b. How would you determine if there is a relationship (and if so, the strength of such relationship) between the independent variables in the model?

c. Given the multiple regression model that was fit in the problem, what would the expected salary be for a 60-year-old male employee with 25 years of experience?

d. Should you have any concerns about the estimate in part **c.**? Explain your answer.

12. In Exercise 11 of Section 14.5, we attempted to predict the number of crimes on a college/university campus based on the number of police, the enrollment at the university, and if it was a private institution.

a. Examine the correlation matrix below and discuss whether collinearity might play a role in estimating the number of crimes based on the number of police, enrollment, and if the institution is private.

Correlation Matrix			
	Police	**Enrollment**	**Private**
Police	1.0000	0.8042	–0.4942
Enrollment		1.0000	–0.8795
Private			1.0000

b. How would you determine if there is a relationship (and if so, the strength of such relationship) between the independent variables in the model?

c. Given the multiple regression model that was fit in the problem, what would the expected number of crimes be for a private university with a police force of 100 officers and an enrollment of 50,000?

d. Should you have any concerns about the estimate in part **c.**? Explain your answer.

13. Suppose you fit a multiple regression model of the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \varepsilon_i$$

The correlation matrix for the pairs of independent variables is given in the following table. Discuss if you detect multicollinearity between any of the variables.

Correlation Matrix				
	x_1	x_2	x_3	x_4
x_1	1.00	0.18	0.86	0.45
x_2		1.00	0.35	0.22
x_3			1.00	0.50
x_4				1.00

Time Series Profile Based on Pegel's and Gardner's Classification (cont.)

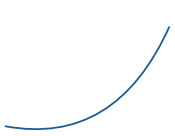
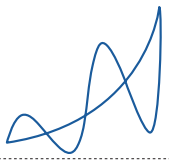
	Nonseasonal	Additive seasonality	Multiplicative seasonality
Exponential trend			
Dampled trend			

Figure 15.1.8

15.1 Exercises

Basic Concepts

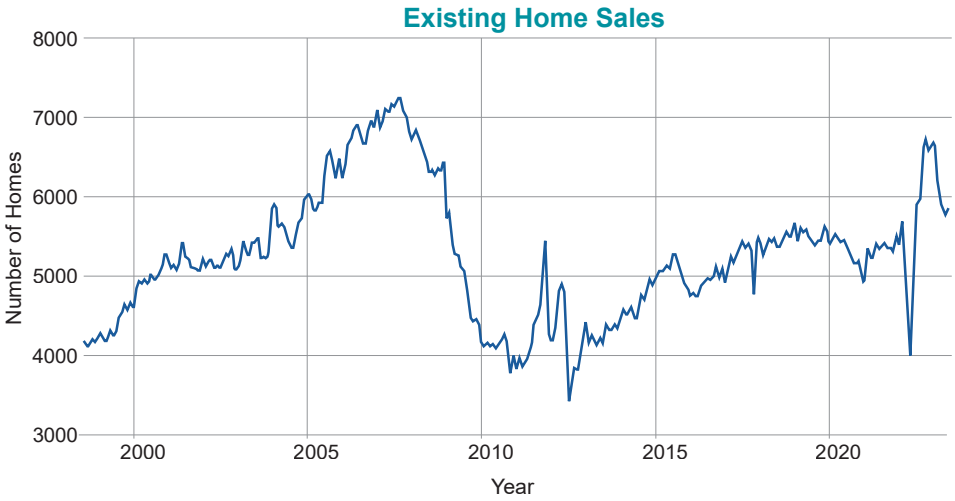
1. What is the difference between seasonal variation and cyclical variation?
2. What is timeframe?
3. Give three examples of business variables that can be represented using a time series plot.
4. What is stationary data?
5. Give three examples of seasonal data in the business world.
6. Suppose a variable is exhibiting a significant upward trend over time. What type of time series data would this represent?
7. What are two ways to determine the best time series method to make a forecast?

Exercises

Data

The data set can be found by visiting stat.hawkeslearning.com and navigating to **Discovering Business Statistics, Second Edition > Data Sets > Border Crossings**.

8. Use the Border Crossings data set. Plot the truck crossings across the U.S.-Canada border at Detroit, MI and identify any time series patterns. Look at the data at the following time frequencies and explain your findings: monthly and yearly.
9. Use the Border Crossings data set. Plot the passenger vehicle crossings across the U.S.-Canada border at Detroit, MI and identify any time series patterns. Look at the data in the following time frequencies and explain your findings: monthly and yearly.
10. Use the Border Crossings data set. Plot the truck and passenger vehicle crossings across the U.S.-Mexico border at Laredo, TX and identify any time series patterns. Look at the data in the following time frequencies and explain your findings: monthly and yearly. In addition, compare the findings with the border crossings at Detroit, MI.
11. What patterns does the existing (not new) home sales time series plot depict?



12. Use the Monthly Average Retail Gas Prices data set, which includes the average gas prices in the U.S. from April 1993 to July 2021.

a. What patterns do you see in the data?

b. Is monthly the right frequency to explore the data, or would you prefer quarterly or yearly? Explain your reasons.

13. Use the Mortgage Rates data set, which includes the yearly mortgage rate in the U.S. from 1971. Currently, there is a belief that the mortgage rate is at an all-time low; do you agree? What is the current trend showing?

Data

The data set can be found by visiting stat.hawkeslearning.com and navigating to **Discovering Business Statistics, Second Edition > Data Sets > Monthly Average Retail Gas Prices**.

Data

The data set can be found by visiting stat.hawkeslearning.com and navigating to **Discovering Business Statistics, Second Edition > Data Sets > Mortgage Rates**.

15.2 Moving Averages

Simple Moving Average (SMA)

The first method we are going to talk about is the **Simple Moving Average (SMA)**. The simple moving average method uses several values (two or more) from the recent past to develop a forecast. It is a smoothing technique because we are taking two to three observations, or even more, and predicting one. Therefore, we are averaging these observations and smoothing out some of the variability.

When we are using the simple moving average, we actually compute the average from a chosen window of points and the resulting average is the forecast for that next period of time. The wider the window of points, the smoother the fit will be, because we are using more observations and turning them into one, thereby smoothing out the variability.

Definition

Simple Moving Average (SMA)

The **simple moving average (SMA)**, uses the average of several values (two or more) from the recent past to develop a forecast.

Formula

Simple Moving Average

MA_n denotes the **moving average** over n periods, and it is the sum of the most recent n data values in the time series divided by the number of periods (n) that we use to calculate that moving average.

$$MA_n = \frac{\sum_{i=1}^{n} D_i}{n}$$

where n = the number of periods used to compute the moving average and

D_i = the actual data value of the time series in period i.

15.2 Exercises

Basic Concepts

1. What is a simple moving average?
2. What would we do if we wanted to predict or forecast for several periods in the future?
3. Give an advantage and a disadvantage of using a simple moving average.
4. How can you determine the number of periods and the appropriate weights for each of those periods in a weighted moving average?
5. What is the primary disadvantage of moving average methods?
6. As the number of periods increases in the moving average, what happens to the forecasts?

Exercises

Data

The data set can be found by visiting stat.hawkeslearning.com and navigating to **Discovering Business Statistics, Second Edition > Data Sets > Border Crossings**.

7. Use the Border Crossings data set. Provide a 3-month SMA forecast for the Laredo truck crossings and predict the number of truck crossings for January 2019.
8. Use the Border Crossings data set. Provide a 5-month SMA forecast for the Laredo passenger crossings and predict the number of passenger crossings for January 2019.
9. Use the Border Crossings data set. Provide a 5-month SMA forecast for the Detroit truck and passenger crossings and predict the number of truck and passenger crossings for January 2019.
10. Use the Border Crossings data set. Provide a 3-month WMA and 5-month WMA forecast for the Laredo truck crossings and predict the number of truck crossings for January 2019. Note: Use the weights of 0.6, 0.3 and 0.1 for the 3-month WMA and 0.4, 0.3, 0.15, 0.1 and 0.05 for the 5-month WMA.

Data

The data set can be found by visiting stat.hawkeslearning.com and navigating to **Discovering Business Statistics, Second Edition > Data Sets > Mortgage Rates**.

11. Use the Mortgage Rates data set, which includes the yearly mortgage rate in the U.S. from 1971. Predict the U.S. mortgage rate for the year 2020 using a 4-year SMA and 4-year WMA. For WMA use the weights of 0.4, 0.3, 0.2 and 0.1, respectively.

Data

The data set can be found by visiting stat.hawkeslearning.com and navigating to **Discovering Business Statistics, Second Edition > Data Sets > Monthly Average Retail Gas Prices**.

12. Use the Monthly Average Retail Gas Prices data set, which includes the average gas prices in the U.S. from April 1993 to July 2021. Predict the retail gasoline price for August 2021 using a 5-month SMA, compare it with a 3-month SMA.

15.3 Exponential Smoothing Techniques

Simple Exponential Smoothing

Another technique that we will use is called **Simple Exponential Smoothing**. With simple exponential smoothing, we weight the most recent observation more than the past using a convex combination of weights. This weighting scheme allows the forecast to react more strongly to quick changes in the data based on the smoothing constant α, which is used as the weight. Small values of α do not react well to changes in the data, whereas large values of α react quickly to changes in the data.

Definition

Simple Exponential Smoothing

In **simple exponential smoothing**, we weight the most recent observation more than the past using a convex combination of weights.

Month	Actual Sales	Forecast Component	Trend Component	Adjusted Forecast
April	10	11.11	0.35	11.46
May	12	11.02	0.26	11.28
June	-	11.50	0.30	11.80

Therefore, the adjusted exponential smoothing forecast for June is 11.80.

Figure 15.3.4 contains the simple exponentially smoothed and the adjusted exponentially smoothed forecasts for the yearly Laredo truck crossings using $\alpha = 0.3$ and $\beta = 0.4$, respectively. Note that the forecast is very close to the actual data.

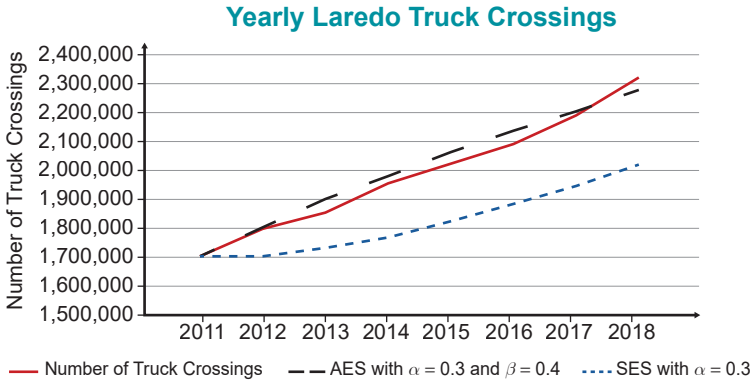


Figure 15.3.4

15.3 Exercises

Basic Concepts

1. What is a simple exponential smoothing?
2. How do different values of α react to changes in the data in simple exponential smoothing?
3. Describe the relationship between weight and period in simple exponential smoothing.
4. Give an advantage and a disadvantage of using simple exponential smoothing.
5. What do we adjust for in adjusted exponential smoothing?
6. For an upward trend, which forecast will be higher? The adjusted or the simple exponential smoothing? What about a downward trend?

Exercises

7. Use the Border Crossings data set. Using $\alpha = 0.3$, calculate the simple exponential smoothing yearly forecast of Detroit truck crossings for 2011–2019. Assume the forecast for the year 2011 to be the actual truck crossing value of year 2011.
8. Use the Border Crossings data set. Using $\alpha = 0.3$ and $\beta = 0.4$, calculate the adjusted exponential smoothing yearly forecast of Laredo truck crossings for 2011–2019. Assume the forecast component for the year 2011 to be the actual truck crossing value of year 2011 minus the initial trend. Let the initial trend be 100,000 trucks. (Note: One way to compute the initial trend is to determine the slope of a linear regression line fit to the data. In this case, if you fit a line to the Laredo yearly truck crossing data, you will get a slope of 84,220. This could be used as the initial trend.)

Data

The data set can be found by visiting stat.hawkeslearning.com and navigating to **Discovering Business Statistics, Second Edition > Data Sets > Border Crossings**.

Data

The data set can be found by visiting stat.hawkeslearning.com and navigating to **Discovering Business Statistics, Second Edition > Data Sets > Mortgage Rates**.

Data

The data set can be found by visiting stat.hawkeslearning.com and navigating to **Discovering Business Statistics, Second Edition > Data Sets > Monthly Average Retail Gas Prices**.

9. Use the Mortgage Rates data set, which contains the yearly mortgage rate in the U.S. since 1971.

a. Calculate the simple exponential smoothing forecast and the adjusted exponential smoothing forecast for the yearly mortgage rates for 2020. Assume $\alpha = 0.2$ and $\beta = 0.3$ for the respective methods.

b. Create a time series plot with both forecasts. What conclusions can you draw from it?

10. Use the Monthly Average Retail Gas Prices data set, which includes the average gas prices in the U.S. from April 1993 to July 2021.

a. Compare the simple exponential smoothing forecast and adjusted exponential smoothing forecast for the monthly gas price and their respective forecasts for August 2020. Assume $\alpha = 0.3$ and $\beta = 0$ for the respective methods and forecast the gasoline price for August 2021. Assume the first period forecasted gas price as the original gas price, and the initial trend is the slope of the linear regression line that fits the data.

b. Create a time series plot with both forecasts. What conclusions can you draw from it?

SOLUTION

$$E = 2 + 1 - 2 + 13 - 3 - 5 = 6$$

The cumulative error (E) for this dataset is 6, which indicates that the forecasts are consistently underestimating the actual data. However, a closer look at the forecast errors reveals that there are ups and downs and the positive value of E is primarily a result of the forecast for Period 4 which underestimates the actual value by 13 units. Thus, examining just E is somewhat of a disadvantage given that three of the forecasts are overestimations and three are underestimations. This drawback is overcome by the tracking signal error metric.

Tracking Signal (*TS*)

As can be seen in Example 15.4.4, bias should never be confirmed using a single measure. Bias should be observed over time, which is the primary purpose of the **tracking signal** (*TS*). It is computed for each time period using the following formula.

$$TS = \frac{E}{\text{MAD}}$$

There are typically two ground rules to detect bias using the tracking signal, which are based on the control chart principles (see Chapter 18). The first rule is that any time the *TS* is above +4 or below –4, it is considered to be "out of control" and the forecast is biased. The second rule is to ensure that the *TS* has no trend, increasing upwards or decreasing downwards, for any considerable amount of time. A trend in the *TS* indicates that the forecasts are inching towards being biased and some corrective measures should be taken, such as using a different forecasting method.

Example 15.4.5

Calculating the Tracking Signal for Forecasts

Using the data in Example 15.4.1, compute the tracking signal (*TS*).

Time Period, t	Actual data, D_t	Forecast, F_t	Error, FE_t	E	MAD	$TS = \frac{E}{\text{MAD}}$
1	134	132	2	2	2.00	**1.00**
2	142	141	1	3	1.500	**2.00**
3	143	145	–2	1	1.667	**0.60**
4	156	143	13	14	4.500	**3.11**
5	151	154	–3	11	4.200	**2.62**
6	145	150	–5	6	4.333	**1.38**

SOLUTION

As you can see from the above table, the forecast is biased as indicated by E. But the *TS* is within +4 and –4 and there is no discernible trend; it goes up and down randomly.

There are several error metrics in forecasting analysis. They all let you make the same decisions—whether the forecast is good or not. Depending upon on our need, we could choose the appropriate error metric on which to focus. For example, MAPD is suggested for accuracy, *TS* is suggested for bias, and MSE is suggested for outliers.

15.4 Exercises

Basic Concepts

1. What is forecast error?
2. If the forecast error is positive, what does it mean? What if it is negative?
3. Mean absolute deviation relies on which variable in order to assess if the forecast is good or not?

4. Describe one key difference between mean absolute deviation and mean absolute percentage error.
5. What is one draw back from the mean absolute percentage error that makes this error metric not so popular?
6. What is the only way that mean absolute percentage deviation is indeterminate?
7. When is the mean squared error especially useful?
8. Having a lower (or higher) error forecast, but consistently below (or above) the time series actual value is an indicator of what?
9. What does a positive/negative cumulative error indicate?
10. What is the drawback in cumulative error that is overcome by the tracking signal?

Exercises

Data

The data set can be found by visiting stat.hawkeslearning.com and navigating to **Discovering Business Statistics, Second Edition > Data Sets > Border Crossings**.

11. Use the Border Crossings data set. For the 3-month SMA forecast of Laredo truck crossings, compute the MAPD and plot the *TS*. Is the forecast good?
12. Use the Border Crossings data set. For the simple exponential smoothing forecast of Detroit truck crossings, compute the MAPD and plot the *TS*. Is the forecast good? (Use $\alpha = 0.2$.)
13. Use the Border Crossings data set. For the adjusted exponential smoothing forecast of Laredo truck crossings, compute the MAPD and plot the *TS*. Is the forecast good? (Use $\alpha = 0.2$; $\beta = 0.4$.)

Data

The data set can be found by visiting stat.hawkeslearning.com and navigating to **Discovering Business Statistics, Second Edition > Data Sets > Monthly Average Retail Gas Prices**.

14. Use the Monthly Average Retail Gas Prices data set, which includes the average gas prices in the U.S. from April 1993 to July 2021. Perform a simple exponential smoothing forecast of retail gasoline price and compute the MSE. (Use $\alpha = 0.3$.)
15. Use the Monthly Average Retail Gas Prices data set, which includes the average gas prices in the U.S. from April 1993 to July 2021. Perform an adjusted exponential smoothing forecast of retail gasoline price and compute the MAPD. (Use $\alpha = 0.3$; $\beta = 0.4$.)
16. Use the Monthly Average Retail Gas Prices data set, which includes the average gas prices in the U.S. from April 1993 to July 2021. Calculate the best α, β combination that minimizes the MAPD of the forecast for retail gasoline price.
17. Which error metric(s) should we concentrate on for each forecasting objective?

Objective	Error Metrics
Minimize Outliers	
Minimize Overall Forecast Errors	
Minimize Bias	
Minimize Overall Forecast Errors of intermittent items	

15.5 Exercises

Basic Concepts

1. What is seasonality?
2. Give a business scenario where we would see seasonality.
3. What assumption do we make with the additive seasonal forecasting method?
4. What is a disadvantage of the additive seasonal model? Which model can overcome this?
5. What is a seasonal factor? What does it represent?
6. What are the four steps to perform a multiplicative seasonal forecast?

Exercises

Data

The data set can be found by visiting stat.hawkeslearning.com and navigating to **Discovering Business Statistics, Second Edition > Data Sets > Border Crossings**.

7. Use the Border Crossings data set. Calculate the monthly seasonality factor for the Laredo Truck Crossings data.
8. Use the Border Crossings data set. Calculate the monthly seasonality factor for the Detroit Truck Crossings data.
9. Use the Border Crossings data set. Compute the additive quarterly seasonal forecast model for the Detroit Truck Crossing data without trend.
10. Use the Border Crossings data set. Compute the additive quarterly seasonal forecast model for the Detroit Truck Crossing data with trend.

Data

The data set can be found by visiting stat.hawkeslearning.com and navigating to **Discovering Business Statistics, Second Edition > Data Sets > Monthly Average Retail Gas Prices**.

11. Use the Monthly Average Retail Gas Prices data set, which includes the average gas prices in the U.S. from April 1993 to July 2021. Using the data from January 1994 to December 2020, compute the monthly seasonality factor for the gas prices. Which month or months generally have the highest gas prices in the U.S.?
12. Use the Monthly Average Retail Gas Prices data set, which includes the average gas prices in the U.S. from April 1993 to July 2021. Use the data from January 2016 to December 2020 to compute the monthly forecast of gas prices for 2021 if the same trend and seasonality pattern continue.

Suppose a significance level of $\alpha = 0.01$ has been specified and our sample size is 14. If we reject the null hypothesis for large values of the test statistic, we look in the table under the column labeled $\chi^2_{0.010}$ and find the critical value corresponding to 14 – 1, or 13 degrees of freedom. The corresponding critical value is 27.688, as shown in Figure 16.1.3.

df	…	$\chi^2_{0.050}$	$\chi^2_{0.025}$	$\chi^2_{0.010}$
1		3.841	5.024	6.635
2		5.991	7.378	9.210
3		7.815	9.348	11.345
…				
13		22.362	24.736	27.688
14		23.685	26.119	29.141
…				

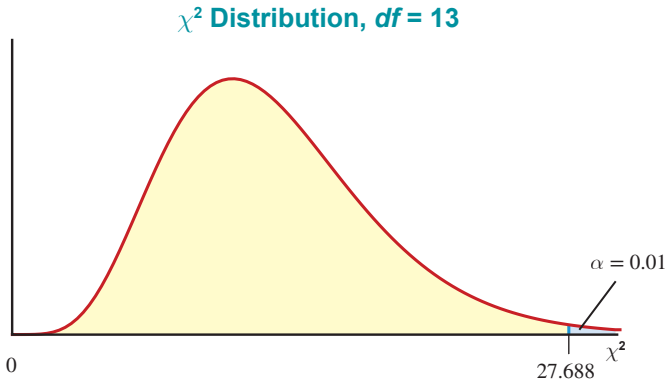


Figure 16.1.3

The analysis in the remainder of the chapter deals with comparing the *actual* number of observations falling into a particular category with the number of observations that is *expected* to fall in that category, based on our hypothesis. In certain circumstances, this type of formulation can be evaluated with a chi-square distribution.

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 χ^2_α.
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. $\alpha = 0.005,\ df = 21$

b. $\alpha = 0.025,\ df = 16$

c. $\alpha = 0.025,\ df = 1$

d. $\alpha = 0.10,\ df = 90$

e. $\alpha = 0.10,\ df = 17$

10. Find the chi-square critical value for each of the following.

a. $\alpha = 0.01,\ df = 10$

b. $\alpha = 0.01,\ df = 21$

c. $\alpha = 0.05,\ df = 6$

d. $\alpha = 0.05,\ df = 11$

e. $\alpha = 0.005,\ df = 29$

11. 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. $\alpha = 0.10,\ df = 50$

12. Suppose that a marketing manager is studying sales data for products that are not available in stores and only sold on television. She collects the following weekly sales data for 10 products not sold in stores. Assume the population standard deviation for these data is $5000.

Weekly Sales Figures

Product	Weekly Sales ($)	Product	Weekly Sales ($)
1	26,259	6	22,511
2	18,514	7	29,753
3	21,579	8	20,235
4	18,739	9	16,258
5	27,821	10	15,990

a. Compute the sample standard deviation for these 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 $\chi^2_{0.05}$ for these data?

13. Michael is studying 30-year fixed mortgage rates in Myrtle Beach, SC. He got 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	3.918
AimLoan	3.925
Great Western	4.062
Greenlight	4.353
Flagstar	4.350
AuroraBank	4.040
Quicken	4.458
Roundpoint	4.125

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 $\chi^2_{0.025}$ for these data.

Step 6: Make the decision and state the conclusion in terms of the original question.

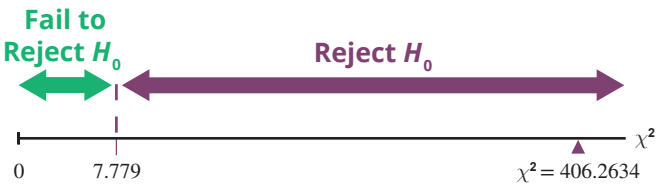


Figure 16.2.4

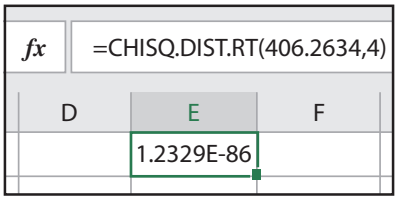
If the null hypothesis is true, the test statistic will be greater than or equal to the critical value of 7.779 only 10% of the time. Since $\chi^2 \approx 406.2634$ is larger than 7.779, we will reject H_0. Large values of the test statistic indicate that the proportion of Americans falling into a category changed from January 2017 to January 2021, and the change is too great to be due to ordinary sampling variation.

The *P*-value for a test statistic value of 406.2634 and degrees of freedom equal to 4 is approximately 0. Therefore, we reject the null hypothesis.

Conclusion and Interpretation: At the 10% level of significance, there is sufficient evidence to conclude that American sentiment towards support for basic research changed from January 2017 to January 2021. The difference in sentiment over the 4-year period is much too great to be attributed to ordinary sampling variation alone.

Technology

The *P*-value can be found in Excel using the CHISQ.DIST.RT function. For instructions, please visit stat.hawkeslearning.com and navigate to **Discovering Business Statistics, Second Edition > Technology Instructions > Chi-Square Distribution > Right-Tailed Probability (cdf)**.



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. A telephone company 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. The market research firm Nielson recently published market share figures for the operating systems in smartphones. The report stated the following results.

Market Share for Smartphone Operating Systems

Operating System	Market Share (%)
Android OS	29
iPhone OS	27
Blackberry OS	27
Microsoft Windows Mobile	10
HP Palm/WebOS	4
Symbian OS	2
Other	1

Source: Nielson.com

Suppose that a marketing manager for a telecommunications company that uses one of the above operating systems doubts the Nielson findings. He collects his own data by surveying 400 people at a local mall. His findings are given in the following table.

Survey Results

Operating System	Number of People
Android OS	125
iPhone OS	115
Blackberry OS	99
Microsoft Windows Mobile	56
HP Palm/WebOS	5
Symbian OS	0
Other	0

a. Compute the expected number of observations for each category for the survey conducted by the telecommunications marketing manager.

b. State the null and alternative hypotheses for the chi-square test for goodness of fit.

c. Using $\alpha = 0.05$, perform a goodness of fit test to determine if the survey conducted by the marketing manager is evidence that the market shares reported by Nielson have changed.

d. What assumptions were made in the test for part **c.**?

e. Do you have any concerns about the way in which the marketing manager's survey was conducted? Explain.

13. 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.**?

16.3 Exercises

Basic Concepts

1. Give two examples of relationships between qualitative variables that would be of interest to a manager in a business setting.
2. Explain the difference between the chi-square test for goodness of fit and the chi-square test for association.
3. What is a contingency table?
4. Describe the information that each cell in a contingency table gives.
5. What properties must the two categories of the contingency table possess?
6. What level(s) of measurement may the categories of a contingency table have?
7. Consider the variable income. Describe how this variable could be transformed to be included in a contingency table. Is information lost during the transformation?
8. Explain why a test for association is not valid if single data points are allowed to belong to more than one category.
9. Restate the multiplication rule for independent events. Explain how this rule pertains to the chi-square test for association.
10. State the null and alternative hypotheses for a chi-square test for association between two qualitative variables.
11. What is the test statistic for the chi-square test for association?
12. How many degrees of freedom are associated with the test statistic given in Exercise 11?

Exercises

13. 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.**?

14. A sociologist is interested in studying the relationship between education and crime. She randomly selects 150 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	2	35
9 Years to 12 Years	4	31
12 Years to 16 Years	1	31
16+ Years	4	42

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.**?

15. A psychologist is preparing his thesis on child abuse. He thinks that there may be a relationship between various types of child abuse and the age of the child. To study this, he randomly selects the records of 197 abused children and determines both the age group in which each child falls (Tweens (ages 9-12) and Teens (ages 13-17)) and the documented type of child abuse. The results of the study are as follows.

Child Abuse

	Age Group	
Type of Abuse	Tweens	Teens
Neglect	50	50
Physical	20	30
Sexual	10	19
Emotional	10	8

a. Can the psychologist conclude that the type of child abuse and age group in which the child falls are dependent at $\alpha = 0.05$?

b. What assumptions were made in the test for part **a.**?

16. 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	40	35
Heating Equipment	25	20	18	19
Arson	17	15	16	15
Electrical	12	13	12	13
Children at Play	10	11	12	11
Other	27	28	29	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.**?

Since X = the number of times the less frequent sign occurs, $X = 15$, and the calculated value of the test statistic is

$$z = \frac{15 + 0.5 - \left(\frac{50}{2}\right)}{\frac{\sqrt{50}}{2}} \approx -2.69 .$$

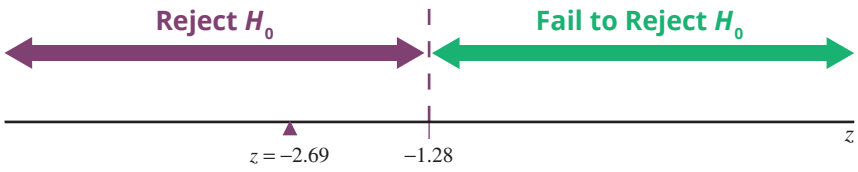


Figure 17.1.4

Step 6: Make the decision and state the conclusion in terms of the original question.

As shown in Figure 17.1.4, 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 $\alpha = 0.10$.

Conclusion and Interpretation: There is sufficient evidence for the Montgomery County Hospital CEO to conclude at $\alpha = 0.10$ that the median length of stay for her patients is significantly shorter than the median length of stay for the United States as a whole.

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 $\alpha = 0.10$.

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. Do 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 ($)			Price ($)			Price ($)	
Textbook	Bookstore	Online Retailer	Textbook	Bookstore	Online Retailer	Textbook	Bookstore	Online Retailer
1	70	60	10	97	86	19	49	40
2	38	36	11	140	130	20	149	127
3	88	89	12	40	30	21	126	130
4	165	149	13	175	150	22	92	93
5	80	136	14	85	75	23	144	129
6	103	95	15	100	85	24	98	84
7	42	50	16	68	62	25	40	52
8	98	111	17	67	69			
9	89	65	18	140	142			

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 cash register will enable cashiers to process a larger number of items on average than the cash register 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 cash register and the new cash register. 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 Register	60	70	55	75	62	52	58
New Register	65	71	55	75	65	57	57

Without making any assumptions about the distribution, can management conclude that the new cash register will allow cashiers to process a significantly larger number of items on average than the old cash register 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 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 also be used to evaluate a paired difference experiment. 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 randomly selected, and it must be possible to rank the differences.

An advantage of the Wilcoxon signed-rank test is that it does not ignore the magnitudes of the differences. However, it does not take the magnitude directly into account. Instead, the ranks of the data are analyzed.

Ranking is nothing new. It simply requires putting the data in order from smallest to largest and attaching a rank to each data item. In general, the lowest value is assigned a rank of one and the highest value is assigned a rank of n, where n is the number of nonzero differences. How do we handle ties? If there are two or more values with the same magnitude, these values will each be assigned the same rank, which is equal to the average of the ranks which would have been assigned to these values if they had slightly different consecutive values. The ranking procedure is explained more fully in the following example.

Example 17.2.1

Ranking Quantitative Data

Rank the following stocks, traded on the New York Stock Exchange, from smallest price to largest price.

Table 17.2.1 – Stock Prices	
Stock	**Price per Share ($)**
Merck & Co., Inc.	78.25
AT&T, Inc.	27.15
SecureWorks Corp.	28.04
Micron Technology, Inc.	73.21
HP Inc.	25.16
AMC Entertainment Holdings, Inc.	30.50
CitiGroup, Inc.	60.12
Exxon Mobil Co.	55.30
AutoCanada Inc.	30.50

Procedure (cont.)

Wilcoxon Signed-Rank Test

Critical Value(s):

If $n \le 25$, reject H_0 if $T \le T_c$, where T_c is the critical value found in Appendix A, Table J.

If $n > 25$:

One-Tailed Test: reject H_0 if $z \le -z_\alpha$.

Two-Tailed Test: reject H_0 if $z \le -z_{\alpha/2}$.

Assumptions:

Pairs of data have been randomly selected and are such that the absolute values of their differences can be ranked.

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 signed-rank test.

Emerging Markets Mutual Funds

Mutual Fund	Price ($)	Mutual Fund	Price ($)
American Funds	24.40	DWS Investments	15.57
Columbia Management	9.41	UBS	12.15
Morgan Stanley	23.74	Prudential Investments	9.23
Fidelity Investments	24.40	Value Line Funds	32.82
John Hancock	9.41	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.

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
Education	4.4
Tobacco and Smoking Products	2.4
Gasoline	33.3
New and Used Motor Vehicles	3.6

Source: Bureau of Labor Statistics

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, do 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, do the data indicate that the DHEA-S level of women increases after practicing transcendental mediation 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 cash register will enable cashiers to process a larger number of items on average than the cash register 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 cash register and the new cash register. 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 Cash Register	60	70	55	75	62	52	58
New Cash Register	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, do the data provide conclusive evidence that the new cash register enables cashiers to process a significantly larger number of items than the old cash register 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 11.3, 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, do 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 11.3, 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.

Procedure (cont.)

Wilcoxon Rank-Sum Test

If $n_1 > 10$, and the smaller sample size, n_1, is associated with Population X, then

$$z = \frac{T - \frac{n_1(n_1 + n_2 + 1)}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$, where T is defined just as when $n_1 \le 10$.

Critical Value(s):

If $n_1 \le 10$:

If H_a is > One-Tailed: then reject H_0 if $T \ge T_U$, the critical value in Table K.

If H_a is < One-Tailed: then reject H_0 if $T \le T_L$, the critical value in Table K.

If H_a is ≠ Two-Tailed: then reject H_0 if $T \le T_L$, or $T \ge T_U$ the critical values in Table K.

If $n_1 > 10$:

if H_a is > One-Tailed, then reject H_0 if $z \ge z_\alpha$.

if H_a is < One-Tailed, then reject H_0 if $z \le -z_\alpha$.

if H_a is ≠ Two-Tailed, then reject H_0 if $z \le -z_{\alpha/2}$ or if $z \ge z_{\alpha/2}$.

Assumptions:

The data are such that they can be ranked. The two samples are selected in an independent and random fashion.

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 ($)	South Side ($)
50,000	43,000
45,000	45,000
55,000	42,000
25,000	50,000
75,000	36,000
35,000	48,000
65,000	38,000
55,000	43,000
45,000	43,000

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 salesmen, Mr. Ellis and Mr. Ford. The auditor randomly selects seven days and determines the daily expenses for each of the salesmen.

Daily Expenses	
Mr. Ellis ($)	Mr. Ford ($)
55	60
53	55
58	65
54	50
56	70
55	55
55	65

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 salesmen, Mr. 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	Program B
85	87
95	96
75	78
100	100
70	74
90	92
80	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. Tom Anderson, 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	Old Battery
655	745
730	675
670	730
715	690
685	760
745	660

a. Using the Wilcoxon rank-sum test, do 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	Bran Flakes Plus
5	6
6	8
4	4
7	9
3	3
5	7
5	5
6	8
4	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	Comedy
180	150
240	190
120	110
220	170
140	130

a. Using the Wilcoxon rank-sum test, do 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	Brand B
800	780
828	805
772	755
830	807
770	753
826	803
774	757

a. Using the Wilcoxon rank-sum test, do 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	Service City
8.50	7.90
9.00	8.35
8.00	7.45
9.07	8.40
7.93	7.40
9.14	8.45
7.86	7.35
8.50	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 The Rank Correlation Test

In Section 4.7 we studied the correlation coefficient as a measure of association between two random variables. The parametric correlation coefficient gives a direct correlation between the variables. In this section we will transform the data from two variables into ranks and develop a method for detecting an association between them.

Since our sample size is 12 ($n < 30$), the value of the statistic is compared with the critical value obtained using Appendix A, Table L.

n	$\alpha = 0.10$	$\alpha = 0.05$	⋯	$\alpha = 0.01$
⋯				
11	0.523	0.623		0.818
12	0.497	0.591		0.780
13	0.475	0.566		0.745
14	0.457	0.545		0.716
15	0.441	0.525		0.689
⋯				

Note that Table L in Appendix A is constructed for a two-tailed test. So, we want to reject H_0 if $r_s \leq -0.497$ or $r_s \geq 0.497$. Since $0.5944 > 0.497$, we reject the null hypothesis at the 10% level of significance. Hence, there seems to be an association between SAT scores and GPAs.

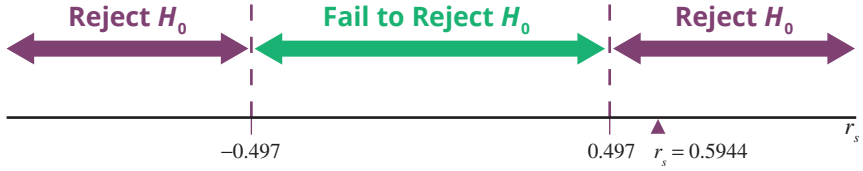


Figure 17.4.1

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 cash register in a supermarket. Each day a sample of purchases at that register is examined and a percent of pricing errors is recorded along with the total number of customers who used that register. Do the following data indicate an association between Chris' performance and how busy his register 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
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 consist 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. Do 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. Do 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

17.5 The Runs Test for Randomness

Randomness is an important concept in probability and statistics. In this section we are going to discuss a method for determining whether a sequence of observations exhibits randomness. To illustrate this concept, we will use the familiar coin-tossing experiment. One characteristic of the coin-tossing experiment is that in the long run there should be approximately equal numbers of heads and tails. In an ordered sequence, however, randomness implies more than compliance with this frequency criterion. For example, if the outcomes of 20 tosses of a coin were recorded as

H H H H H T T T T T T T T T T H H H H H,

we would suspect that the process was flawed. We would be equally surprised if the ordered outcomes were

H T H T H T H T H T H T H T H T H T H T,

but be reasonably happy with the sequence

H H T H T T T H T H H T H T H H H T T H.

A characteristic that reflects our reservations about the first two sequences is the number of **runs**, where a run is a subsequence of one or more heads (or tails).

Definition

Run

A **run** is a series of increasing values, a series of decreasing values, or a sequence of at least one symbol.

In the first sequence, there are three runs: a run of 5 heads, then 10 tails, then 5 heads.

H H H H H | T T T T T T T T T T | H H H H H

Run Run Run

In the second sequence, there are 20 runs, each consisting of a single head or tail.

H | T | H | T | H | T | H | T | H | T | H | T | H | T | H | T | H | T | H | T

R R R R R R R R R R R R R R R R R R R R

Example 17.5.3

Detecting Randomness of a Set of Numbers

Are the following data random? Test at $\alpha = 0.05$.

16, 25, 52, 11, 38, 47, 12, 98, 4

SOLUTION

How do you test randomness with a numerical 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 are random.

H_a: The data are not random.

Median = 25

16	25	52	11	38	47	12	98	4
B	∅	A	B	A	A	B	A	B

$m = 4$ (the number of A's)

$n = 4$ (the number of B's)

$R = 7$ (the number of runs)

Using Appendix A, Table M the rejection region is $R \leq 1$ or $R \geq 9$ at the 0.05 level of significance. Since $R = 7$, we fail to reject H_0 and conclude that there is not sufficient evidence of non-randomness.

Figure 17.5.6

Technology

For instructions on performing the runs test for randomness using technology, please visit stat.hawkeslearning.com and navigate to **Discovering Business Statistics, Second Edition > Technology Instructions > Nonparametrics > Runs Test for Randomness**.

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. Suppose that in your city the number of deaths due to traffic accidents involving drunk driving from 1999 to 2011 were 75, 91, 54, 85, 79, 63, 12, 55, 63, 49, 89, 98, and 71. Use the runs test to examine non-randomness at the 0.05 level.

9. A sociologist designs a study that involves a procedure of selecting families randomly from a phone book and then calling 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?

17.6 The Kruskal–Wallis Test

In this section we present a procedure where k random samples are obtained, one from each of the k possibly different populations, and we are interested in testing whether all of the populations have identical distributions. Suitable hypotheses for this test would be as follows.

H_0: The populations from which the samples are drawn have identical distributions.

H_a: Not all populations have the same distribution.

Definition

Kruskal-Wallis Test

The **Kruskal-Wallis test** is a nonparametric procedure that can be used to determine if two or more distributions are different.

The **Kruskal-Wallis test** is a method that can be used instead of the ANOVA F-test. The Kruskal-Wallis test does not need the assumption of normality of the populations. It does require that independent, random samples be drawn.

The Kruskal-Wallis test is similar to the Wilcoxon rank-sum test in that the test statistic will be based on the sums of the ranks of the groups being compared.

The data consist of k random samples (not necessarily the same size) drawn from their respective populations. The data set may be arranged as follows.

Group 1	Group 2	…	Group k
$x_{1,1}$	$x_{2,1}$	…	$x_{k,1}$
$x_{1,2}$	$x_{2,2}$	…	$x_{k,2}$
…	…	…	…
x_{1,n_1}	x_{2,n_2}	…	x_{k,n_k}

Let N be the total number of observations, that is, $N = \sum_{i=1}^{k} n_i$.

SOLUTION

The null and alternative 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.

In Table 17.6.3 we are given the ranks of the observations. We then need to determine R_1, R_2, and R_3, corresponding to the sums of the ranks assigned to the observations for brake pads A, B, and C. The sum of the ranks for each brake pad is as follows.

$$R_1 = 21$$
$$R_2 = 55$$
$$R_3 = 44$$

Having the ranks, we can now calculate the test statistic, H.

$$\begin{aligned} H &= \frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(N+1) \\ &= \frac{12}{15(15+1)}\left(\frac{21^2}{5}+\frac{55^2}{5}+\frac{44^2}{5}\right) - 3(15+1) \\ &= 6.02 \end{aligned}$$

Referring to Appendix A, Table G, we see that $\chi^2_{0.05} = 5.991$. Thus, we want to reject the null hypothesis if the test statistic, H, is greater than or equal to 5.991. Since the test statistic exceeds the critical value ($6.02 > 5.991$), we reject the null hypothesis and conclude that the braking distances are sufficiently different.

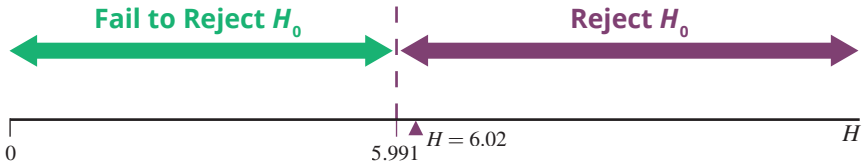


Figure 17.6.2

Technology

For instructions on performing the Kruskal-Wallis test using technology, please visit stat.hawkeslearning.com and navigate to **Discovering Business Statistics, Second Edition > Technology Instructions > Nonparametrics > Kruskal-Wallis Test**.

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 are 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 these 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 show 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?

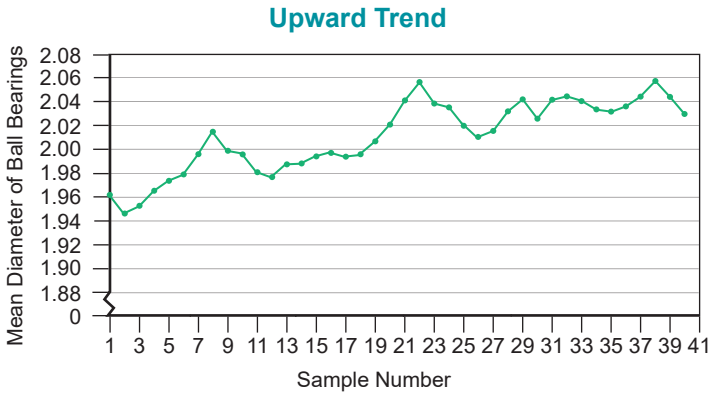


Figure 18.1.3

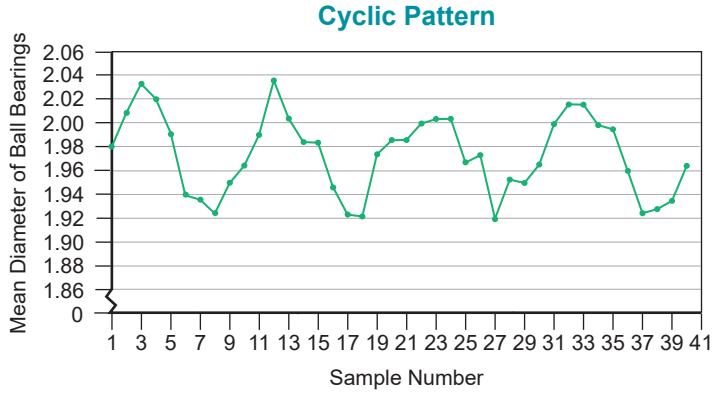


Figure 18.1.4

NOTE

Unstable patterns on a runs chart are often the result of special causes — perhaps a systematic environmental change (e.g. maintenance, employee fatigue, or equipment rotation).

Definition

Cycle

A **cycle** is a systematic repeating pattern observed in a run chart.

Notice how the data in the graph in Figure 18.1.4 move systematically up and down in a repeating pattern. This pattern (called a **cycle**) is also an indication of a process that is unstable.

18.1 Exercises

Basic Concepts

1. Describe the contributions made by Dr. Walter A. Shewhart in the field of quality control.
2. What contributions did W. Edwards Deming make in the field of quality control? Of Deming's 14 Points, give five that you believe are most important.
3. To which types of organizations do Deming's 14 Points apply?
4. What is the common philosophy of the fathers of modern quality control?
5. What is Six Sigma? Briefly describe the methodology behind Six Sigma.
6. What is a flowchart? Why are flowcharts important in quality control?
7. What is a Pareto chart? Explain how a Pareto chart can be used in the field of quality control.
8. What is a run chart? Is a run chart the same as a time series plot? Explain.
9. Explain how a run chart can be used in the field of quality control.

Exercises

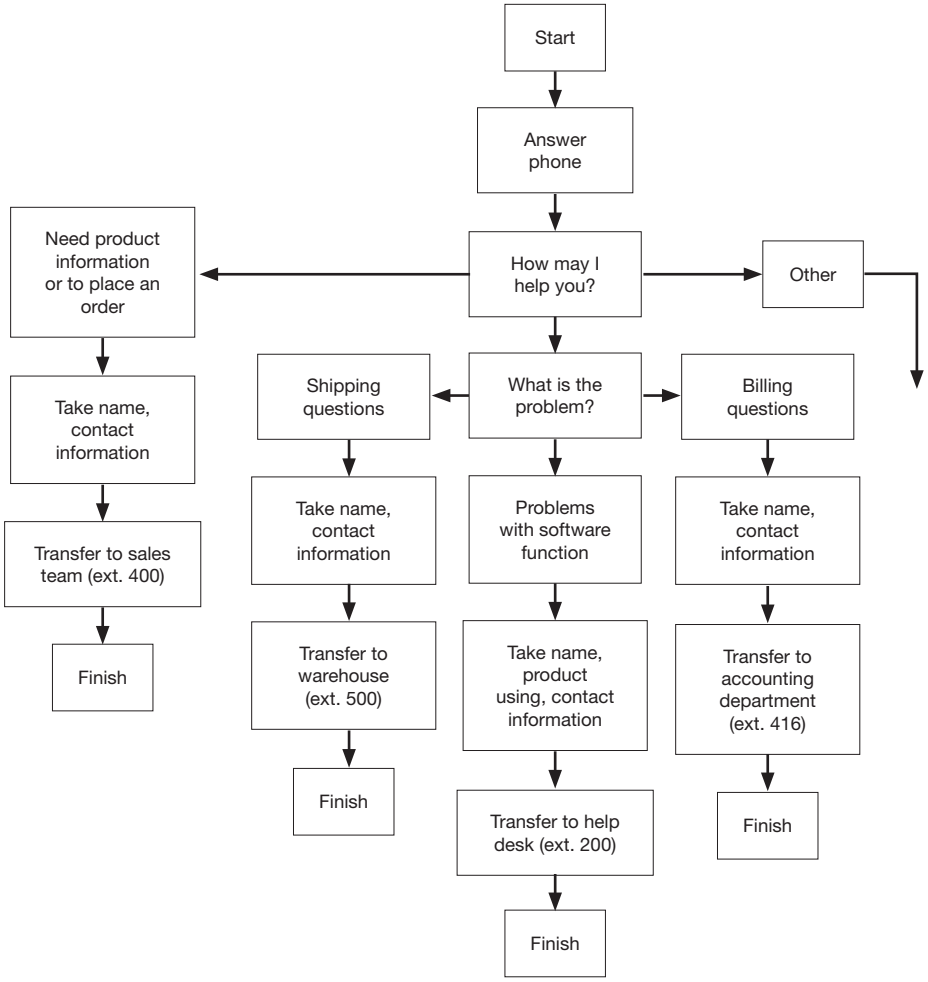
10. Create a flowchart for each of the following processes.
    a. Getting ready for work in the morning

b. Getting married

c. Going on a week-long vacation in Bermuda

d. Going to a job interview

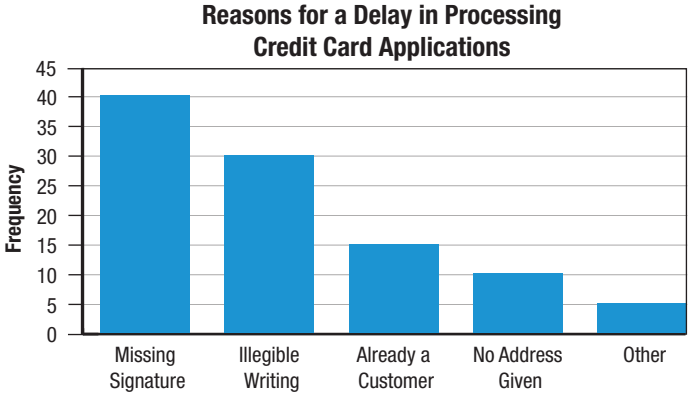
11. Consider the following flowchart regarding incoming call routing for a software company.



a. Explain why it is important for a new receptionist working at the company to understand this flowchart.

b. In what ways do you think this flowchart could be improved?

12. Consider the following Pareto chart regarding reasons for a delay in processing credit card applications.



a. What percentage of delays is caused by a missing signature on the application?

b. What percentage of delays is caused by a missing signature or illegible writing on the application?

c. From the chart, what would you identify as the "vital few" problems?

d. How do you think the credit card company could attempt to correct these problems?

e. Does the 80/20 notion seem to apply here? Explain.

13. Consider the following data regarding customer complaints for a clothing store.

Customer Complaints

Complaint	Frequency
Not Enough Parking	80
Rude Personnel	50
Poor Lighting	42
Confusing Store Layout	28
Limited Sizes	15
Clothing Unattractive	10

a. Compute the relative frequencies for the complaints listed in the table.

b. Create a Pareto chart for the data.

c. Which problem(s) would you identify as the "vital few"?

d. Does the 80/20 notion seem to apply here? Explain.

e. How should the clothing store proceed in attempting to improve customer satisfaction? In your opinion, what should be done first? Explain why.

14. Consider the following data regarding the number of returned products for a large online retailer, by month, for the years 2003 and 2011.

Number of Returned Products in 2003 and 2011

Month	2003	2011
January	79	100
February	81	105
March	92	96
April	101	84
May	111	72
June	120	80
July	119	64
August	125	60
September	137	55
October	120	59
November	140	42
December	145	56

a. Create a run chart for the number of returned items, by month, in 2003.

b. Analyze the run chart for returned items in 2003. Is there a downward or upward trend or is the pattern cyclic?

c. Would you consider the process to be stable or unstable? Explain why.

d. Create a run chart for the number of returned items, by month, in 2011.

e. Analyze the run chart for returned items in 2011. Is there a downward or upward trend or is the pattern cyclic? Does the process appear to be stable or unstable? Explain.

f. Does it appear that the retailer has improved the process from 2003 to 2011? Explain.

15. Consider the following data regarding revenues, by quarter, for a popular local restaurant.

Quarterly Revenues (Thousands of Dollars)

Year	Quarter	Revenue
2008	1	270
2008	2	369
2008	3	468
2008	4	306
2009	1	285
2009	2	354
2009	3	525
2009	4	330
2010	1	261
2010	2	288
2010	3	375
2010	4	366
2011	1	303
2011	2	420
2011	3	471
2011	4	414

a. Create a run chart for the quarterly revenues.

b. Analyze the run chart. Is there a downward trend or an upward trend? Do revenues appear to be cyclical? Explain.

c. Do you think restaurant sales depend on the time of year?

d. Can you think of any reasons why there would be this type of trend for restaurant sales?

e. Does this process appear to be unstable? If so, suggest ways that quality control could help the restaurant manager control the process.

18.2 Basic Concepts

The scientific method for attaining quality relies on two basic concepts.

1. No matter what the specifications are for a product, the process that produces the product will create output that has *variation*. (For example, suppose a manufacturer desires to produce ball bearings with a diameter of two inches. If a process is set up to produce ball bearings with a diameter of two inches, then each item, when actually measured, will show deviation from the *ideal* of two inches.)
2. Improving a process requires removing variation from it. (Though the ideal would be to remove all variation, this cannot be achieved. The goal then is to move towards the ideal. This notion is known as *continuous improvement*.)

With these two powerful ideas in mind, let's look at some important definitions.

Definition

Control Chart Terminology

A **control chart** for a process consists of values plotted over time. This chart has an upper bound and a lower bound called the **upper control limit (UCL)** and the **lower control limit (LCL)**, respectively. The process is **out of control** when a measurement falls either above the UCL or below the LCL. The control chart also contains a **centerline** that represents the average value of the quality characteristic corresponding to the in-control state.

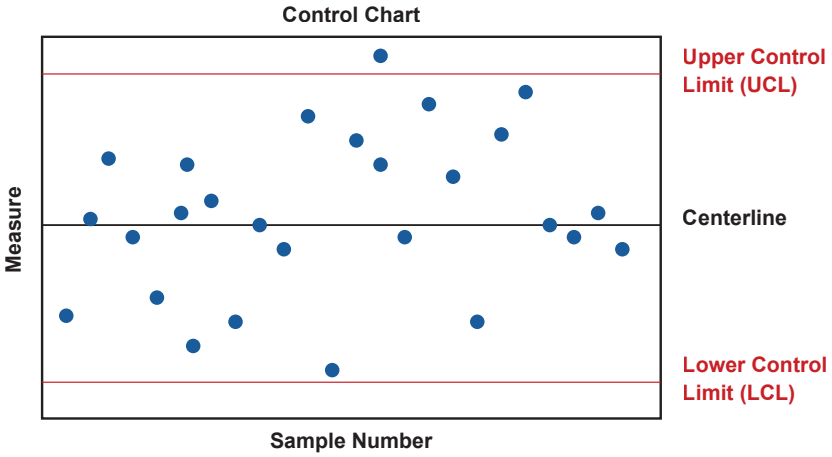


Figure 18.2.1

Definition

Types of Variation

Normal process variation is normal variation in a process in which the data falls within the control limits. **Assignable variation** is random variation that causes data to fall outside the control limits but can be reduced by determining the root cause of the variation.

The Highway Control Chart

When you are driving a car and you stay in your lane, you could say that you are operating the car "in control." The white lines that define your lane are similar to the UCL and LCL. The car would be expected to move around within the lane, which would be normal process variation. Veering outside your lane might have assignable causes such as cell phone usage, children fighting in the back seat, or any number of other distractions.

The control chart with its limits tells us when to stop the process (if it is out of control) and when not to interrupt the process. When data points fall within the UCL and LCL, we think the variation is due to **normal process variation** (also called **common cause variation** or **chance variation**). But when a point or points fall outside the control limits, the cause is said to be **assignable variation** (or **special cause variation**). This type of variation is not random and can be eliminated (or reduced) by investigating the problem and determining the root cause(s). Reducing system variation is the surest path toward continuous improvement. For assignable causes, the system should be stopped and the cause(s) should be found and removed before the process is resumed.

It is important to emphasize that a control chart focuses on the process, not on the product. It does not ensure good quality, but instead allows management to check a quality characteristic of the process at regular intervals in order to determine if the statistical distribution of the characteristic has changed. If it has, then modifications may be needed to correct the process.

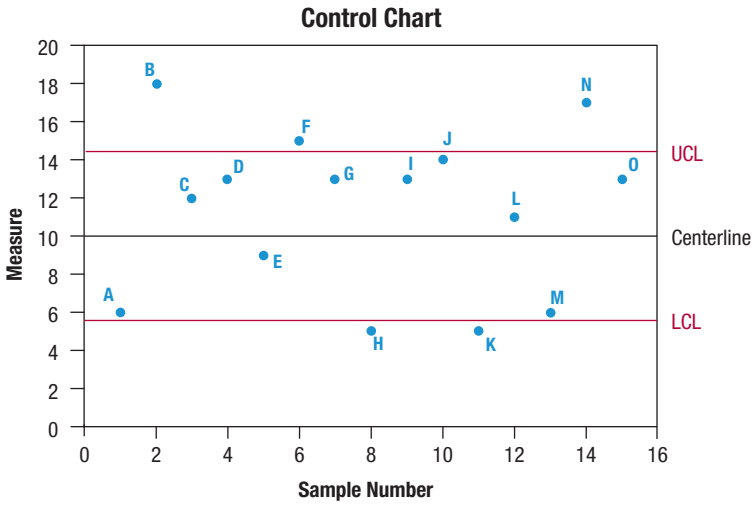
18.2 Exercises

Basic Concepts

1. Identify and describe the two concepts on which the scientific method for attaining quality is based on.
2. What is a control chart? What is the basic purpose of control charts?
3. Identify and define the three basic components of a control chart.
4. What is normal process variation?
5. What is assignable variation?
6. Does a control chart give information about a process or a product? Explain.
7. What does it mean to say that a process is in control?
8. Does an in-control state guarantee quality output? Explain.

Exercises

9. Consider the following control chart.



a. From the chart, estimate the values of the UCL, LCL, and centerline.

b. Interpret the estimated values of the UCL, LCL, and centerline.

c. Which points, if any, are out of control?

10. Consider the following values of the UCL, LCL, and centerline from a control chart.

13.56, 16.56, 10.56

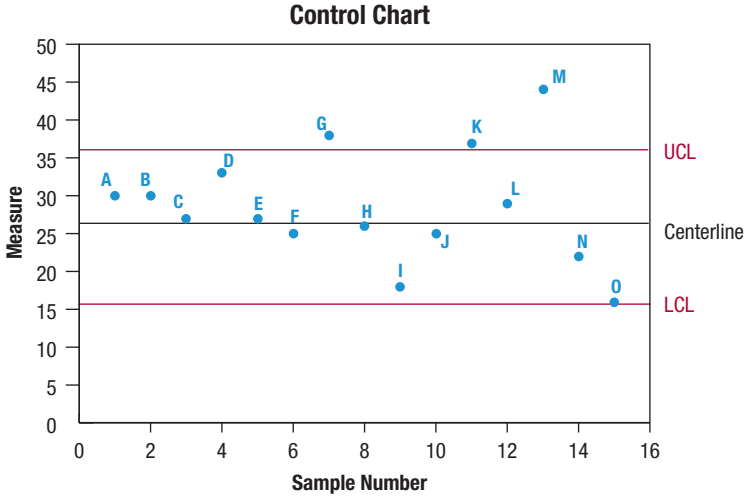
a. Identify which value is the LCL, which is the UCL, and which is the centerline.

b. Plot the UCL, LCL, and centerline on a control chart.

c. Identify three points that would be considered in control and three points that would be considered out of control.

d. Plot these points on the chart that you made in part **b.**

11. Consider the following control chart.



a. From the chart, estimate the values of the UCL, LCL, and centerline.

b. Interpret the estimated values of the UCL, LCL, and centerline.

c. Identify any points that can be attributed to assignable variation.

Computing the control limits for the R chart, we have

$$\begin{aligned}\text{UCL} &= \bar{R}D_4 \\ &= 0.0575(2.574) \\ &\approx 0.1480 \\ \text{LCL} &= \bar{R}D_3 \\ &= 0.0575(0) \\ &= 0 \\ \text{Centerline} &= \bar{R} = 0.0575.\end{aligned}$$

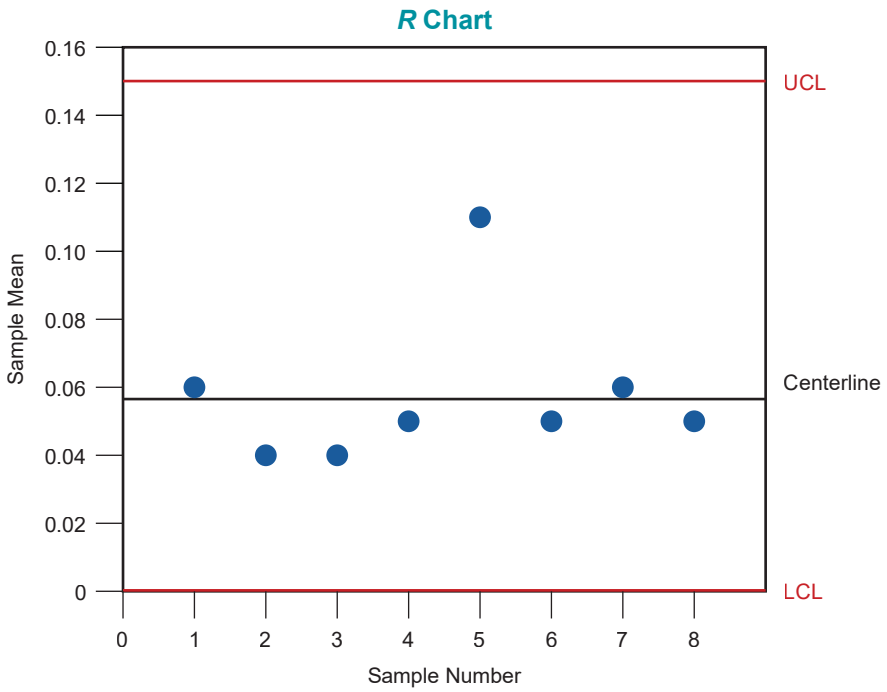


Figure 18.3.4

Examining the $\bar{x}$ chart and R chart, one can see that both the process mean and variability are within the control limits. That is, each of the sample means and sample ranges falls within the control limits of the $\bar{x}$ chart and R chart, respectively, indicating that the process is in control.

18.3 Exercises

Basic Concepts

1. How are control charts associated with hypothesis testing? Identify the null and alternative hypotheses that can be tested using control charts.
2. What is an $\bar{x}$ chart?
3. What is an R chart?
4. What does the Central Limit Theorem have to do with statistical quality control?
5. What is a common interval used to measure in-control and out-of-control processes? What is the probability that random variation would cause a sample statistic to fall outside this interval?

6. How are the upper and lower control limits for an $\bar{x}$ chart calculated if the process mean and standard deviation are known?

7. How do the calculations for the control limits change for an $\bar{x}$ chart when the process mean and standard deviation are unknown?

8. When studying a control chart, how do you determine if the process is out of control?

9. Consider the process in Example 17.1. Give an example of something that might cause this process to be out of control.

10. Why do managers study $\bar{x}$ charts and R charts together?

Exercises

11. An automobile manufacturer requires the fuel injections to be adjusted so that its cars get an average of 30 miles per gallon with a standard deviation of 3 miles per gallon over a long period of time. Calculate the upper and lower control limits for an $\bar{x}$ chart if the quality control department starts sampling 64 cars each day during the month.

12. A pharmaceutical manufacturer requires the average active ingredient of allergy pills to be 0.03 grams, with a standard deviation of 0.002 grams. An FDA inspector inspects 10 batches of 100 pills and finds the following sample means.

0.032 0.028 0.031 0.032 0.026 0.027 0.030 0.033 0.034 0.026

a. Determine which batches, if any, are out of control using an $\bar{x}$ chart.

b. Does the process appear to be in statistical control? Explain.

13. A manufacturer of auto windows employs a constant quality control technique where the thickness of glass is checked every hour. A perfect piece of glass will have a thickness of 4 mm. From past experience, it is known that the standard deviation of thickness is 0.25 mm. The result of one shift's production is given in the following table.

Glass Thickness (mm)

Sample	Observations					Sample	Observations				
1	4	3	4	2	4	9	5	4	3	2	5
2	5	3	5	4	2	10	4	4	4	4	4
3	3	3	3	3	4	11	1	6	4	4	2
4	4	5	5	4	5	12	3	3	4	5	4
5	4	2	2	3	2	13	4	4	5	6	5
6	4	5	2	5	4	14	3	2	5	4	2
7	3	5	4	4	5	15	3	3	3	2	1
8	2	4	4	4	3	16	4	4	4	5	3

a. Construct an $\bar{x}$ chart for these data.

b. Construct an R chart for these data.

c. According to the $\bar{x}$ chart constructed in part **a.**, which samples, if any, are out of control?

d. According to the R chart constructed in part **b.**, which samples, if any, are out of control?

14. Princeton Manufacturing produces air conditioning units designed to maintain 45 degrees. Samples of 10 units are taken to monitor the process, and it is found that the units are maintaining 45 degrees as designed. The mean of the sample ranges is found to be 2 degrees.

a. Find the UCL, LCL, and centerline for the $\bar{x}$ chart.

b. Find the UCL, LCL, and centerline for the R chart.

15. A trucking company tries to deliver its freight in 24 hours. Ten samples of 20 customers are taken with the following sample means.

22.6 24.5 24.1 23.8 25.3 25.0 23.8 23.6 23.0 25.2

The average range for these deliveries is 5.8 hours.

a. Compute the 3σ control limits for the mean delivery time. Which samples, if any, are out of control?

b. Compute the 3σ control limits for the process range.

16. Natural Life produces a variety of natural food products. The quality control department samples one cereal to ensure proper net weight. In the past when taking samples of 15 boxes, the average range was 0.45 ounces. Find the upper and lower control limits for an R chart.

17. A paper products manufacturer makes 60-inch cores that are later cut into smaller lengths in the production of bathroom tissue. To monitor the production and to make sure the cores are acceptable for the cutting state, a sample of 25 cores is taken each hour of the day. Along with the core length, the range of the core length is recorded for each sample (as shown in the following table). Determine the upper and lower control limits for an R chart and indicate which samples, if any, are out of control.

Core Lengths

Sample Number	Sample Range	Sample Number	Sample Range
1	0.10	13	0.19
2	0.20	14	0.08
3	0.22	15	0.10
4	0.08	16	0.07
5	0.06	17	0.16
6	0.23	18	0.19
7	0.20	19	0.21
8	0.09	20	0.14
9	0.25	21	0.16
10	0.17	22	0.19
11	0.14	23	0.12
12	0.18	24	0.13

18. A manufacturer of small electric motors has a model that draws 1300 watts when working properly. To ensure conformity with this standard, samples of 20 motors are taken each hour during the day shift. Along with the average wattage, the range of wattage is recorded for each sample (as shown in the following table). Determine the upper and lower control limits for an R chart and indicate which samples, if any, are out of control.

Electric Motor Wattage

Sample Number	Sample Range	Sample Number	Sample Range
1	8.8	5	4.1
2	12.2	6	9.6
3	11.6	7	8.8
4	8.0	8	5.3

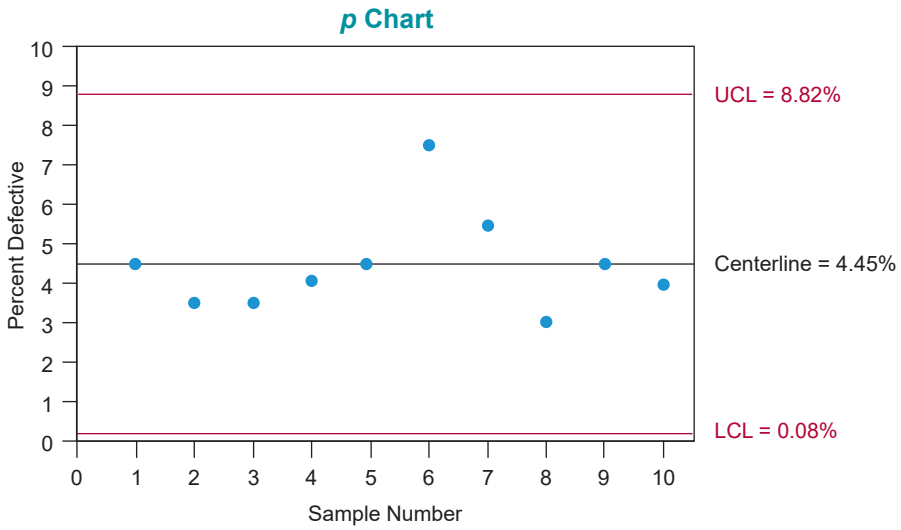


Figure 18.4.2

From the p chart, we can see that for this data, when the standard percent defective is unknown, all of the samples fall within the control limits.

18.4 Exercises

Basic Concepts

1. Explain the difference between control charts for attributes and control charts for variables.
2. What is a p chart?
3. How are the upper and lower control limits calculated for a p chart when the process proportion is known?
4. Is the allowable variation for a process involving a p chart larger, smaller, or the same as the allowable variation for a process involving a mean chart or range chart? Explain.
5. Suppose the LCL for a process is computed to be -0.07. What value should be used for the LCL in the p chart? Explain.
6. How does the procedure for constructing a p chart change if the process proportion is not known?
7. What is $\bar{p}$? What other hypothesis testing procedure uses the concept of $\bar{p}$? Are these measures the same? Explain.
8. When examining a p chart, how do you determine if samples are out of control?

Exercises

9. In a paper products plant, 100 product samples are taken each hour and tested for being either acceptable or defective. In the past, 1% defective was considered normal. Find the upper and lower control limits for a 3σ p chart.
10. To monitor the production of sheet metal screws by a particular machine in a large manufacturing company, a sample of 100 screws is examined each hour for three shifts of eight hours each. Each screw is inspected and designated as conforming or nonconforming according to specifications. Historically, the proportion of nonconforming screws has been 5%. Use the following results of one day's sampling to construct a 3σ p chart. Which samples, if any, are out of control?

Nonconforming Screws					
Sample Number	Number Defective	Sample Number	Number Defective	Sample Number	Number Defective
1	4	9	10	17	9
2	7	10	5	18	11
3	9	11	5	19	14
4	10	12	4	20	5
5	8	13	12	21	6
6	6	14	6	22	12
7	5	15	7	23	15
8	1	16	13	24	5

11. A production process involves the manufacture of rubber gaskets for windows. When these gaskets are inspected, they are classified as conforming or nonconforming based on a number of different characteristics, such as thickness, consistency, overall size, and so on. To monitor the percentage of nonconforming gaskets being produced, a sample of 25 gaskets is inspected each hour. Management predetermines the acceptable fraction of nonconforming gaskets as 10%.

a. Determine the UCL, LCL, and centerline.

b. Use the following table to plot the samples on your control chart.

Nonconforming Gaskets			
Sample Number	Percent Defective	Sample Number	Percent Defective
1	16	13	12
2	16	14	8
3	16	15	8
4	12	16	12
5	8	17	8
6	8	18	12
7	4	19	12
8	0	20	4
9	8	21	8
10	4	22	12
11	4	23	16
12	4	24	4

c. Are any samples out of control? If so, identify which ones.

12. The academic dean decides to sample 200 students each semester to study the drop rate at his institution. The numbers of drops for the last eight semesters are shown in the following table. Find the upper and lower control limits and construct a *p* chart. Indicate which semesters, if any, are out of control.

Number of Drops			
Semester Number	Number of Drops	Semester Number	Number of Drops
1	10	5	8
2	12	6	6
3	14	7	13
4	9	8	15

13. The Thompson Company makes voltage protectors at its Midland, Georgia plant. During each shift, 10 protectors are tested until failure, with some rated defective and others rated non-defective. The numbers of defective protectors for the last 20 shifts are given in the following table. Find the upper and lower control limits and construct a *p* chart. Indicate which shifts, if any, are out of control.

Defective Voltage Protectors			
Sample Number	Number Defective	Sample Number	Number Defective
1	1	11	1
2	1	12	1
3	0	13	1
4	2	14	2
5	3	15	0
6	1	16	0
7	0	17	2
8	2	18	1
9	3	19	1
10	0	20	1