Tenth Edition

Elementary Statistics

A STEP BY STEP APPROACH

ALLAN G. BLUMAN

Mc Graw Hill Education

1

The Nature of Probability and Statistics

© Shutterstock/Monkey Business Images RF

STATISTICS TODAY

Is Higher Education "Going Digital"?

Today many students take college courses online and use eBooks. Also, many students use a laptop, smartphone, or computer tablet in the classroom. With the increased use of technology, some questions about the effectiveness of this technology have been raised. For example,

How many colleges and universities offer online courses?

Do students feel that the online courses are equal in value to the traditional classroom presentations?

Approximately how many students take online courses now?

Will the number of students who take online courses increase in the future?

Has plagiarism increased since the advent of computers and the Internet?

Do laptops, smartphones, and tablets belong in the classroom?

Have colleges established any guidelines for the use of laptops, smartphones, and tablets?

To answer these questions, Pew Research Center conducted a study of college graduates and college presidents in 2011. The procedures they used and the results of the study are explained in this chapter. See Statistics Today—Revisited at the end of the chapter.

OUTLINE

OBJECTIVES

After competing this chapter, you should be able to:

1. Demonstrate knowledge of statistical terms.
2. Differentiate between the two branches of statistics.
3. Identify types of data.
4. Identify the measurement level for each variable.
5. Identify the four basic sampling techniques.
6. Explain the difference between an observational and an experimental study.
7. Explain how statistics can be used and misused.
8. Explain the importance of computers and calculators in statistics.

Introduction

You may be familiar with probability and statistics through radio, television, newspapers, and magazines. For example, you may have read statements like the following found in newspapers.

- A recent survey found that 76% of the respondents said that they lied regularly to their friends.
- *The Tribune Review* reported that the average hospital stay for circulatory system ailments was 4.7 days and the average of the charges per stay was $52,574.
- Equifax reported that the total amount of credit card debt for a recent year was $642 billion.
- A report conducted by the SAS Holiday Shopping Styles stated that the average holiday shopper buys gifts for 13 people.
- The U.S. Department of Agriculture reported that a 5-foot 10-inch person who weighs 154 pounds will burn 330 calories for 1 hour of dancing.
- The U.S. Department of Defense reported for a recent year that the average age of active enlisted personnel was 27.4 years.

Unusual Stats

Of people in the United States, 14% said that they feel happiest in June, and 14% said that they feel happiest in December.

Statistics is used in almost all fields of human endeavor. In sports, for example, a statistician may keep records of the number of yards a running back gains during a football game, or the number of hits a baseball player gets in a season. In other areas, such as public health, an administrator might be concerned with the number of residents who contract a new strain of flu virus during a certain year. In education, a researcher might want to know if new methods of teaching are better than old ones. These are only a few examples of how statistics can be used in various occupations.

Interesting Fact

Every day in the United States about 120 golfers claim that they made a hole-in-one.

Furthermore, statistics is used to analyze the results of surveys and as a tool in scientific research to make decisions based on controlled experiments. Other uses of statistics include operations research, quality control, estimation, and prediction.

Statistics is the science of conducting studies to collect, organize, summarize, analyze, and draw conclusions from data.

Historical Note

A Scottish landowner and president of the Board of Agriculture, Sir John Sinclair introduced the word *statistics* into the English language in the 1798 publication of his book on a statistical account of Scotland. The word *statistics* is derived from the Latin word *status*, which is loosely defined as a statesman.

There are several reasons why you should study statistics.

1. Like professional people, you must be able to read and understand the various statistical studies performed in your fields. To have this understanding, you must be knowledgeable about the vocabulary, symbols, concepts, and statistical procedures used in these studies.
2. You may be called on to conduct research in your field, since statistical procedures are basic to research. To accomplish this, you must be able to design experiments; collect, organize, analyze, and summarize data; and possibly make reliable predictions or forecasts for future use. You must also be able to communicate the results of the study in your own words.
3. You can also use the knowledge gained from studying statistics to become better consumers and citizens. For example, you can make intelligent decisions about what products to purchase based on consumer studies, about government spending based on utilization studies, and so on.

It is the purpose of this chapter to introduce the goals for studying statistics by answering questions such as the following:

What are the branches of statistics?

What are data?

How are samples selected?

1–1 Descriptive and Inferential Statistics

OBJECTIVE 1

Demonstrate knowledge of statistical terms.

To gain knowledge about seemingly haphazard situations, statisticians collect information for *variables*, which describe the situation.

A **variable** is a characteristic or attribute that can assume different values.

Data are the values (measurements or observations) that the variables can assume. Variables whose values are determined by chance are called **random variables.**

Suppose that an insurance company studies its records over the past several years and determines that, on average, 3 out of every 100 automobiles the company insured were involved in accidents during a 1-year period. Although there is no way to predict the specific automobiles that will be involved in an accident (random occurrence), the company can adjust its rates accordingly, since the company knows the general pattern over the long run. (That is, on average, 3% of the insured automobiles will be involved in an accident each year.)

A collection of data values forms a **data set.** Each value in the data set is called a **data value** or a **datum.**

In statistics it is important to distinguish between a sample and a population.

Historical Note

The 1880 Census had so many questions on it that it took 10 years to publish the results.

A **population** consists of all subjects (human or otherwise) that are being studied.

When data are collected from every subject in the population, it is called a *census*.

For example, every 10 years the United States conducts a census. The primary purpose of this census is to determine the apportionment of the seats in the House of Representatives.

The first census was conducted in 1790 and was mandated by Article 1, Section 2 of the Constitution. As the United States grew, the scope of the census also grew. Today the Census limits questions to populations, housing, manufacturing, agriculture, and mortality. The Census is conducted by the Bureau of the Census, which is part of the Department of Commerce.

Most of the time, due to the expense, time, size of population, medical concerns, etc., it is not possible to use the entire population for a statistical study; therefore, researchers use samples.

Historical Note

The origin of descriptive statistics can be traced to data collection methods used in censuses taken by the Babylonians and Egyptians between 4500 and 3000 B.C. In addition, the Roman Emperor Augustus (27 B.C.–A.D. 17) conducted surveys on births and deaths of the citizens of the empire, as well as the number of livestock each owned and the crops each citizen harvested yearly.

A **sample** is a group of subjects selected from a population.

If the subjects of a sample are properly selected, most of the time they should possess the same or similar characteristics as the subjects in the population. See Figure 1–1.

However, the information obtained from a statistical sample is said to be *biased* if the results from the sample of a population are radically different from the results of a census of the population. Also, a sample is said to be biased if it does not represent the population from which it has been selected. The techniques used to properly select a sample are explained in Section 1–3.

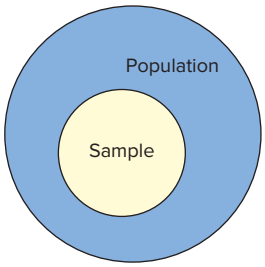
OBJECTIVE 2

Differentiate between the two branches of statistics.

The body of knowledge called statistics is sometimes divided into two main areas, depending on how data are used. The two areas are

1. Descriptive statistics
2. Inferential statistics

FIGURE 1–1
Population and Sample



Descriptive statistics consists of the collection, organization, summarization, and presentation of data.

In *descriptive statistics* the statistician tries to describe a situation. Consider the national census conducted by the U.S. government every 10 years. Results of this census give you the average age, income, and other characteristics of the U.S. population. To obtain this information, the Census Bureau must have some means to collect relevant data. Once data are collected, the bureau must organize and summarize them. Finally, the bureau needs a means of presenting the data in some meaningful form, such as charts, graphs, or tables.

The second area of statistics is called *inferential statistics.*

Inferential statistics consists of generalizing from samples to populations, performing estimations and hypothesis tests, determining relationships among variables, and making predictions.

Historical Note

Inferential statistics originated in the 1600s, when John Graunt published his book on population growth, *Natural and Political Observations Made upon the Bills of Mortality*. About the same time, another mathematician/astronomer, Edmond Halley, published the first complete mortality tables. (Insurance companies use mortality tables to determine life insurance rates.)

Here, the statistician tries to make inferences from *samples* to *populations.* Inferential statistics uses **probability,** i.e., the chance of an event occurring. You may be familiar with the concepts of probability through various forms of gambling. If you play cards, dice, bingo, or lotteries, you win or lose according to the laws of probability. Probability theory is also used in the insurance industry and other areas.

The area of inferential statistics called **hypothesis testing** is a decision-making process for evaluating claims about a population, based on information obtained from samples. For example, a researcher may wish to know if a new drug will reduce the number of heart attacks in men over age 70 years of age. For this study, two groups of men over age 70 would be selected. One group would be given the drug, and the other would be given a placebo (a substance with no medical benefits or harm). Later, the number of heart attacks occurring in each group of men would be counted, a statistical test would be run, and a decision would be made about the effectiveness of the drug.

Statisticians also use statistics to determine *relationships* among variables. For example, relationships were the focus of the most noted study in the 20th century, "Smoking and Health," published by the Surgeon General of the United States in 1964. He stated that after reviewing and evaluating the data, his group found a definite relationship between smoking and lung cancer. He did not say that cigarette smoking actually causes lung cancer, but that there is a relationship between smoking and lung cancer. This conclusion was based on a study done in 1958 by Hammond and Horn. In this study, 187,783 men were observed over a period of 45 months. The death rate from lung cancer in this group of volunteers was 10 times as great for smokers as for nonsmokers.

Unusual Stat

Twenty-nine percent of Americans want their boss's job.

Finally, by studying past and present data and conditions, statisticians try to make predictions based on this information. For example, a car dealer may look at past sales records for a specific month to decide what types of automobiles and how many of each type to order for that month next year.

EXAMPLE 1–1 Descriptive or Inferential Statistics

Determine whether descriptive or inferential statistics were used.

a. The average price of a 30-second ad for the Academy Awards show in a recent year was 1.90 million dollars.

b. The Department of Economic and Social Affairs predicts that the population of Mexico City, Mexico, in 2030 will be 238,647,000 people.

c. A medical report stated that taking statins is proven to lower heart attacks, but some people are at a slightly higher risk of developing diabetes when taking statins.

d. A survey of 2234 people conducted by the Harris Poll found that 55% of the respondents said that excessive complaining by adults was the most annoying social media habit.

SOLUTION

a. A descriptive statistic (average) was used since this statement was based on data obtained in a recent year.

b. Inferential statistics were used since this is a prediction for a future year.

c. Inferential statistics were used since this conclusion was drawn from data obtained from samples and used to conclude that the results apply to a population.

d. Descriptive statistics were used since this is a result obtained from a sample of 2234 survey respondents.

Applying the Concepts 1–1

Attendance and Grades

Read the following on attendance and grades, and answer the questions.

A study conducted at Manatee Community College revealed that students who attended class 95 to 100% of the time usually received an A in the class. Students who attended class 80 to 90% of the time usually received a B or C in the class. Students who attended class less than 80% of the time usually received a D or an F or eventually withdrew from the class.

Based on this information, attendance and grades are related. The more you attend class, the more likely it is you will receive a higher grade. If you improve your attendance, your grades will probably improve. Many factors affect your grade in a course. One factor that you have considerable control over is attendance. You can increase your opportunities for learning by attending class more often.

1. What are the variables under study?
2. What are the data in the study?
3. Are descriptive, inferential, or both types of statistics used?
4. What is the population under study?
5. Was a sample collected? If so, from where?
6. From the information given, comment on the relationship between the variables.

See page 38 for the answers.

Unusual Stat

Only one-third of crimes committed are reported to the police.

Exercises 1–1

1. Define statistics.

2. What is a variable?

3. What is meant by a census?

4. How does a population differ from a sample?

5. Explain the difference between descriptive and inferential statistics.

6. Name three areas where probability is used.

7. Why is information obtained from samples used more often than information obtained from populations?

8. What is meant by a biased sample?

For Exercises 9–17, determine whether descriptive or inferential statistics were used.

9. Because of the current economy, 49% of 18- to 34- year-olds have taken a job to pay the bills. (*Source:* Pew Research Center)

10. In 2025, the world population is predicted to be 8 billion people. (*Source:* United Nations)

11. In a weight loss study using teenagers at Boston University, 52% of the group said that they lost weight and kept it off by counting calories.

12. Based on a sample of 2739 respondents, it is estimated that pet owners spent a total of 14 billion dollars on veterinarian care for their pets. (*Source:* American Pet Products Association, Pet Owners Survey)

13. A recent article stated that over 38 million U.S. adults binge-drink alcohol.

14. The Centers for Disease Control and Prevention estimated that for a specific school year, 7% of children in kindergartens in the state of Oregon had nonmedical waivers for vaccinations.

15. A study conducted by a research network found that people with fewer than 12 years of education had lower life expectancies than those with more years of education.

16. A survey of 1507 smartphone users showed that 38% of them purchased insurance at the same time as they purchased their phones.

17. Forty-four percent of the people in the United States have type O blood. (*Source:* American Red Cross)

Extending the Concepts

18. Find three statistical studies and explain whether they used descriptive or inferential statistics.

19. Find a gambling game and explain how probability was used to determine the outcome.

1–2 Variables and Types of Data

OBJECTIVE 3

Identify types of data.

As stated in Section 1–1, statisticians gain information about a particular situation by collecting data for random variables. This section will explore in greater detail the nature of variables and types of data.

Variables can be classified as qualitative or quantitative.

Qualitative variables are variables that have distinct categories according to some characteristic or attribute.

For example, if subjects are classified according to gender (male or female), then the variable *gender* is qualitative. Other examples of qualitative variables are religious preference and geographic locations.

Quantitative variables are variables that can be counted or measured.

For example, the variable *age* is numerical, and people can be ranked in order according to the value of their ages. Other examples of quantitative variables are heights, weights, and body temperatures.

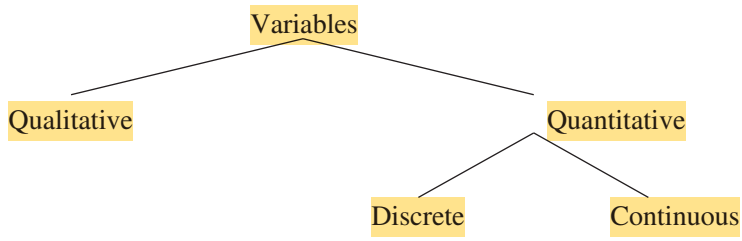
Quantitative variables can be further classified into two groups: discrete and continuous. *Discrete variables* can be assigned values such as 0, 1, 2, 3 and are said to be *countable.* Examples of discrete variables are the number of children in a family, the number of students in a classroom, and the number of calls received by a call center each day for a month.

Discrete variables assume values that can be counted.

Continuous variables, by comparison, can assume an infinite number of values in an interval between any two specific values. Temperature, for example, is a continuous variable, since the variable can assume an infinite number of values between any two given temperatures.

Continuous variables can assume an infinite number of values between any two specific values. They are obtained by measuring. They often include fractions and decimals.

The classification of variables can be summarized as follows:



EXAMPLE 1–2 Discrete or Continuous Data

Classify each variable as a discrete or continuous variable.

a. The number of hours during a week that children ages 12 to 15 reported that they watched television.
b. The number of touchdowns a quarterback scored each year in his college football career.
c. The amount of money a person earns per week working at a fast-food restaurant.
d. The weights of the football players on the teams that play in the NFL this year.

SOLUTION

a. Continuous, since the variable time is measured
b. Discrete, since the number of touchdowns is counted
c. Discrete, since the smallest value that money can assume is in cents
d. Continuous, since the variable weight is measured

Unusual Stat

Fifty-two percent of Americans live within 50 miles of a coastal shoreline.

Since continuous data must be measured, answers must be rounded because of the limits of the measuring device. Usually, answers are rounded to the nearest given unit. For example, heights might be rounded to the nearest inch, weights to the nearest ounce, etc. Hence, a recorded height of 73 inches could mean any measure from 72.5 inches up to but not including 73.5 inches. Thus, the boundary of this measure is given as 72.5–73.5 inches. The **boundary** of a number, then, is defined as a class in which a data value would be placed before the data value was rounded. *Boundaries are written for convenience as 72.5–73.5 but are understood to mean all values up to but not including 73.5.* Actual data values of 73.5 would be rounded to 74 and would be included in a class with boundaries of 73.5 up to but not including 74.5, written as 73.5–74.5. As another example, if a recorded weight is 86 pounds, the exact boundaries are 85.5 up to but not including 86.5, written as 85.5–86.5 pounds. Table 1–1 helps to clarify this concept. The boundaries of a continuous variable are given in one additional decimal place and always end with the digit 5.

TABLE 1–1 Recorded Values and Boundaries

Variable	Recorded value	Boundaries
Length	15 centimeters (cm)	14.5–15.5 cm
Temperature	86 degrees Fahrenheit (°F)	85.5–86.5°F
Time	0.43 second (sec)	0.425–0.435 sec
Mass	1.6 grams (g)	1.55–1.65 g

EXAMPLE 1–3 Class Boundaries

Find the boundaries for each measurement.

a. 17.6 inches
b. 23° Fahrenheit
c. 154.62 mg/dl

SOLUTION

a. 17.55–18.55 inches
b. 22.5–23.5° Fahrenheit
c. 154.615–154.625 mg/dl

In addition to being classified as qualitative or quantitative, variables can be classified by how they are categorized, counted, or measured. For example, can the data be organized into specific categories, such as area of residence (rural, suburban, or urban)? Can the data values be ranked, such as first place, second place, etc.? Or are the values obtained from measurement, such as heights, IQs, or temperature? This type of classification—i.e., how variables are categorized, counted, or measured—uses **measurement scales,** and four common types of scales are used: nominal, ordinal, interval, and ratio.

OBJECTIVE 4

Identify the measurement level for each variable.

The first level of measurement is called the *nominal level* of measurement. A sample of college instructors classified according to subject taught (e.g., English, history, psychology, or mathematics) is an example of nominal-level measurement. Classifying survey subjects as male or female is another example of nominal-level measurement. No ranking or order can be placed on the data. Classifying residents according to zip codes is also an example of the nominal level of measurement. Even though numbers are assigned as zip codes, there is no meaningful order or ranking. Other examples of nominal-level data are political party (Democratic, Republican, independent, etc.), religion (Christianity, Judaism, Islam, etc.), and marital status (single, married, divorced, widowed, separated).

Unusual Stat

Sixty-three percent of us say we would rather hear the bad news first.

The **nominal level of measurement** classifies data into mutually exclusive (nonoverlapping) categories in which no order or ranking can be imposed on the data.

Historical Note

When data were first analyzed statistically by Karl Pearson and Francis Galton, almost all were continuous data. In 1899, Pearson began to analyze discrete data. Pearson found that some data, such as eye color, could not be measured, so he termed such data *nominal data.* Ordinal data were introduced by a German numerologist Frederich Mohs in 1822 when he introduced a hardness scale for minerals. For example, the hardest stone is the diamond, which he assigned a hardness value of 1500. Quartz was assigned a hardness value of 100. This does not mean that a diamond is 15 times harder than quartz. It only means that a diamond is harder than quartz. In 1947, a psychologist named Stanley Smith Stevens made a further division of continuous data into two categories, namely, interval and ratio.

The next level of measurement is called the *ordinal level*. Data measured at this level can be placed into categories, and these categories can be ordered, or ranked. For example, from student evaluations, guest speakers might be ranked as superior, average, or poor. Floats in a homecoming parade might be ranked as first place, second place, etc. *Note that precise measurement of differences in the ordinal level of measurement* does not *exist*. For instance, when people are classified according to their build (small, medium, or large), a large variation exists among the individuals in each class.

Other examples of ordinal data are letter grades (A, B, C, D, F).

The **ordinal level of measurement** classifies data into categories that can be ranked; however, precise differences between the ranks do not exist.

The third level of measurement is called the *interval level.* This level differs from the ordinal level in that precise differences do exist between units. For example, many standardized psychological tests yield values measured on an interval scale. IQ is an example of such a variable. There is a meaningful difference of 1 point between an IQ of 109 and an IQ of 110. Temperature is another example of interval measurement, since there is a meaningful difference of 1°F between each unit, such as 72 and 73°F. *One property is lacking in the interval scale: There is no true zero*. For example, IQ tests do not measure people who have no intelligence. For temperature, 0°F does not mean no heat at all.

The **interval level of measurement** ranks data, and precise differences between units of measure do exist; however, there is no meaningful zero.

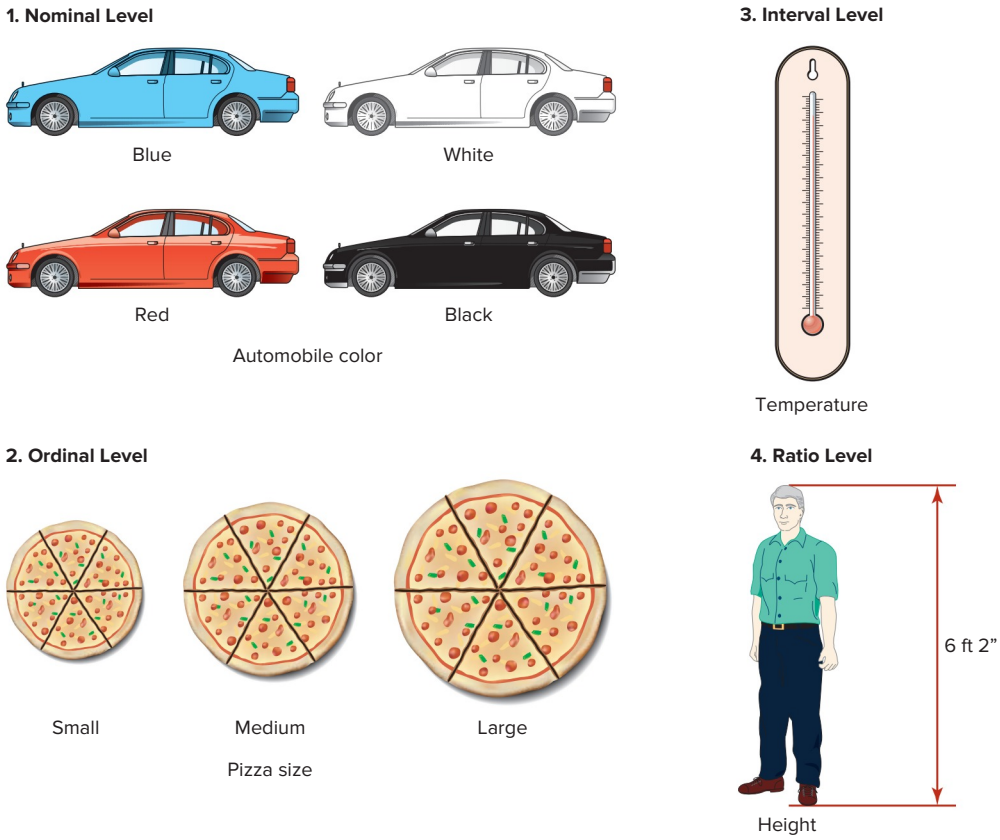
The final level of measurement is called the *ratio level.* Examples of ratio scales are those used to measure height, weight, area, and number of phone calls received. Ratio scales have differences between units (1 inch, 1 pound, etc.) and a true zero. In addition, the ratio scale contains a true ratio between values. For example, if one person can lift 200 pounds and another can lift 100 pounds, then the ratio between them is 2 to 1. Put another way, the first person can lift twice as much as the second person.

The **ratio level of measurement** possesses all the characteristics of interval measurement, and there exists a true zero. In addition, true ratios exist when the same variable is measured on two different members of the population.

TABLE 1–2 Examples of Measurement Scales

Nominal-level data	Ordinal-level data	Interval-level data	Ratio-level data
Zip code Gender (male, female) Eye color (blue, brown, green, hazel) Political affiliation Religious affiliation Major field (mathematics, computers, etc.) Nationality	Grade (A, B, C, D, F) Judging (first place, second place, etc.) Rating scale (poor, good, excellent) Ranking of tennis players	SAT score IQ Temperature	Height Weight Time Salary Age

FIGURE 1–2
Measurement Scales



There is not complete agreement among statisticians about the classification of data into one of the four categories. For example, some researchers classify IQ data as ratio data rather than interval. Also, data can be altered so that they fit into a different category. For instance, if the incomes of all professors of a college are classified into the three categories of low, average, and high, then a ratio variable becomes an ordinal variable. Table 1–2 gives some examples of each type of data. See Figure 1–2.

EXAMPLE 1–4 Measurement Levels

What level of measurement would be used to measure each variable?

a. The ages of authors who wrote the hardback versions of the top 25 fiction books sold during a specific week

b. The colors of baseball hats sold in a store for a specific year

c. The highest temperature for each day of a specific month

d. The ratings of bands that played in the homecoming parade at a college

SOLUTION

a. Ratio
b. Nominal
c. Interval
d. Ordinal

Applying the Concepts 1–2

Fatal Transportation Injuries

Read the following information about the number of fatal accidents for the transportation industry in for a specific year, and answer each question.

Industry	Number of fatalities
Highway accidents	968
Railway accidents	44
Water vehicle accidents	52
Aircraft accidents	151

Source: Bureau of Labor Statistics.

1. Name the variables under study.
2. Categorize each variable as quantitative or qualitative.
3. Categorize each quantitative variable as discrete or continuous.
4. Identify the level of measurement for each variable.
5. The railroad had the fewest fatalities for the specific year. Does that mean railroads have fewer accidents than the other industries?
6. What factors other than safety influence a person's choice of transportation?
7. From the information given, comment on the relationship between the variables.

See page 38 for the answers.

Exercises 1–2

1. Explain the difference between qualitative variables and quantitative variables.

2. Explain the difference between discrete and continuous variables.

3. Why are continuous variables rounded when they are used in statistical studies?

4. Name and define the four types of measurement levels used in statistics.

For Exercises 5–10, determine whether the data are qualitative or quantitative.

5. Sizes of soft drinks sold by a fast-food restaurant (small, medium, and large)

6. Pizza sizes (small, medium, and large)

7. Cholesterol counts for individuals

8. Microwave wattage

9. Number of degrees awarded by a college each year for the last 10 years

10. Ratings of teachers

For Exercises 11–16, determine whether the data are discrete or continuous.

11. Number of phone calls received by a 911 call center each day

12. Systolic blood pressure readings

13. Weights of the suitcases of airline passengers on a specific flight

14. Votes received by mayoral candidates in a city election

15. Number of students in the mathematics classes during the fall semester at your school for a particular school year

16. Temperatures at a seashore resort

For Exercises 17–22, give the boundaries of each value.

17. 24 feet

18. 6.3 millimeters

19. 143 miles

20. 19.63 tons

21. 200.7 miles

22. 19 quarts

For Exercises 23–30, classify each as nominal-level, ordinal-level, interval-level, or ratio-level measurement.

23. Telephone numbers

24. Leap years: . . . 2016, 2020, 2024, . . .

25. Distances communication satellites in orbit are from Earth

26. Scores on a statistical final exam

27. Rating of cooked ribs at a rib cook-off

28. Blood types—O, A, B, AB

29. Online spending in dollars

30. Horsepower of automobile engines

1–3 Data Collection and Sampling Techniques

OBJECTIVE 5

Identify the four basic sampling techniques.

In research, statisticians use data in many different ways. As stated previously, data can be used to describe situations or events. For example, a manufacturer might want to know something about the consumers who will be purchasing his product so he can plan an effective marketing strategy. In another situation, the management of a company might survey its employees to assess their needs in order to negotiate a new contract with the employees' union. Data can be used to determine whether the educational goals of a school district are being met. Finally, trends in various areas, such as the stock market, can be analyzed, enabling prospective buyers to make more intelligent decisions concerning what stocks to purchase. These examples illustrate a few situations where collecting data will help people make better decisions on courses of action.

Data can be collected in a variety of ways. One of the most common methods is through the use of surveys. Surveys can be done by using a variety of methods. Three of the most common methods are the telephone survey, the mailed questionnaire, and the personal interview.

Telephone surveys have an advantage over personal interview surveys in that they are less costly. Also, people may be more candid in their opinions since there is no face-to-face contact. A major drawback to the telephone survey is that some people in the population will not have phones or will not answer when the calls are made; hence, not all people have a chance of being surveyed. Also, many people now have unlisted numbers and cell phones, so they cannot be surveyed. Finally, even the tone of voice of the interviewer might influence the response of the person who is being interviewed.

Mailed questionnaire surveys can be used to cover a wider geographic area than telephone surveys or personal interviews since mailed questionnaire surveys are less expensive to conduct. Also, respondents can remain anonymous if they desire. Disadvantages of mailed questionnaire surveys include a low number of responses and inappropriate answers to questions. Another drawback is that some people may have difficulty reading or understanding the questions.

© Banana Stock Ltd RF

Historical Note

A pioneer in census taking was Pierre-Simon de Laplace. In 1780, he developed the Laplace method of estimating the population of a country. The principle behind his method was to take a census of a few selected communities and to determine the ratio of the population to the number of births in these communities. (Good birth records were kept.) This ratio would be used to multiply the number of births in the entire country to estimate the number of citizens in the country.

2

Frequency Distributions and Graphs

STATISTICS TODAY

How Your Identity Can Be Stolen

Identity fraud is a big business today—more than 12.7 million people were victims. The total amount of the fraud in 2014 was $16 billion. The average amount of the fraud for a victim is $1260, and the average time to correct the problem is 40 hours. The ways in which a person's identity can be stolen are presented in the following table:

Government documents or benefits fraud	38.7%
Credit card fraud	17.4
Phone or utilities fraud	12.5
Bank fraud	8.2
Attempted identity theft	4.8
Employment-related fraud	4.8
Loan fraud	4.4
Other identity theft	9.2

Source: Javelin Strategy & Research; Council of Better Business Bureau, Inc.

Looking at the numbers presented in a table does not have the same impact as presenting numbers in a well-drawn chart or graph. The article did not include any graphs. This chapter will show you how to construct appropriate graphs to represent data and help you to get your point across to your audience.

See Statistics Today—Revisited at the end of the chapter for some suggestions on how to represent the data graphically.

OUTLINE

OBJECTIVES

After completing this chapter, you should be able to

1. Organize data using a frequency distribution.
2. Represent data in frequency distributions graphically, using histograms, frequency polygons, and ogives.
3. Represent data using bar graphs, Pareto charts, time series graphs, pie graphs, and dotplots.
4. Draw and interpret a stem and leaf plot.

Introduction

When conducting a statistical study, the researcher must gather data for the particular variable under study. For example, if a researcher wishes to study the number of people who were bitten by poisonous snakes in a specific geographic area over the past several years, he or she has to gather the data from various doctors, hospitals, or health departments.

To describe situations, draw conclusions, or make inferences about events, the researcher must organize the data in some meaningful way. The most convenient method of organizing data is to construct a *frequency distribution.*

After organizing the data, the researcher must present them so they can be understood by those who will benefit from reading the study. The most useful method of presenting the data is by constructing *statistical charts* and *graphs.* There are many different types of charts and graphs, and each one has a specific purpose.

This chapter explains how to organize data by constructing frequency distributions and how to present the data by constructing charts and graphs. The charts and graphs illustrated here are histograms, frequency polygons, ogives, pie graphs, Pareto charts, and time series graphs. A graph that combines the characteristics of a frequency distribution and a histogram, called a stem and leaf plot, is also explained.

2–1 Organizing Data

OBJECTIVE 1

Organize data using a frequency distribution.

Suppose a researcher wished to do a study on the ages of the 50 wealthiest people in the world. The researcher first would have to get the data on the ages of the people. In this case, these ages are listed in *Forbes Magazine*. When the data are in original form, they are called **raw data** and are listed next.

45	46	64	57	85
92	51	71	54	48
27	66	76	55	69
54	44	54	75	46
61	68	78	61	83
88	45	89	67	56
81	58	55	62	38
55	56	64	81	38
49	68	91	56	68
46	47	83	71	62

Since little information can be obtained from looking at raw data, the researcher organizes the data into what is called a *frequency distribution.*

Unusual Stats

Of Americans 50 years old and over, 23% think their greatest achievements are still ahead of them.

A **frequency distribution** is the organization of raw data in table form, using classes and frequencies.

Each raw data value is placed into a quantitative or qualitative category called a **class.** The **frequency** of a class then is the number of data values contained in a specific class. A frequency distribution is shown for the preceding data set.

Class limits	Tally	Frequency
27–35	/	1
36–44	///	3
45–53	𝍸 ////	9
54–62	𝍸 𝍸	15
63–71	𝍸 𝍸	10
72–80	///	3
81–89	𝍸 //	7
90–98	//	2
		50

Now some general observations can be made from looking at the frequency distribution. For example, it can be stated that the majority of the wealthy people in the study are 45 years old or older.

The classes in this distribution are 27–35, 36–44, etc. These values are called *class limits.* The data values 27, 28, 29, 30, 31, 32, 33, 34, 35 can be tallied in the first class; 36, 37, 38, 39, 40, 41, 42, 43, 44 in the second class; and so on.

Two types of frequency distributions that are most often used are the *categorical frequency distribution* and the *grouped frequency distribution.* The procedures for constructing these distributions are shown now.

Categorical Frequency Distributions

The **categorical frequency distribution** is used for data that can be placed in specific categories, such as nominal- or ordinal-level data. For example, data such as political affiliation, religious affiliation, or major field of study would use categorical frequency distributions.

EXAMPLE 2–1 Distribution of Blood Types

Twenty-five army inductees were given a blood test to determine their blood type. The data set is

A	B	B	AB	O
O	O	B	AB	B
B	B	O	A	O
A	O	O	O	AB
AB	A	O	B	A

Construct a frequency distribution for the data.

SOLUTION

Since the data are categorical, discrete classes can be used. There are four blood types: A, B, O, and AB. These types will be used as the classes for the distribution.

The procedure for constructing a frequency distribution for categorical data is given next.

Step 1 Make a table as shown.

A Class	B Tally	C Frequency	D Percent
A B O AB			

Step 2 Tally the data and place the results in column B.

Step 3 Count the tallies and place the results in column C.

Step 4 Find the percentage of values in each class by using the formula

$$\% = \frac{f}{n} \cdot 100$$

where f = frequency of the class and n = total number of values. For example, in the class of type A blood, the percentage is

$$\% = \frac{5}{25} \cdot 100 = 20\%$$

Percentages are not normally part of a frequency distribution, but they can be added since they are used in certain types of graphs such as pie graphs. Also, the decimal equivalent of a percent is called a *relative frequency.*

Step 5 Find the totals for columns C (frequency) and D (percent). The completed table is shown. It is a good idea to add the percent column to make sure it sums to 100%. This column won't always sum to 100% because of rounding.

A Class	B Tally	C Frequency	D Percent
A	𝍸	5	20
B	𝍸 //	7	28
O	𝍸 ////	9	36
AB	////	4	16
		Total 25	100%

For the sample, more people have type O blood than any other type.

Grouped Frequency Distributions

When the range of the data is large, the data must be grouped into classes that are more than one unit in width, in what is called a **grouped frequency distribution.** For example, a distribution of the blood glucose levels in milligrams per deciliter (mg/dL) for 50 randomly selected college students is shown.

Unusual Stats

Six percent of Americans say they find life dull.

Class limits	Class boundaries	Tally	Frequency
58–64	57.5–64.5	/	1
65–71	64.5–71.5	𝍸 /	6
72–78	71.5–78.5	𝍸 𝍸	10
79–85	78.5–85.5	𝍸 𝍸 ////	14
86–92	85.5–92.5	𝍸 𝍸 //	12
93–99	92.5–99.5	𝍸	5
100–106	99.5–106.5	//	2
			Total 50

The procedure for constructing the preceding frequency distribution is given in Example 2–2; however, several things should be noted. In this distribution, the values 58 and 64 of the first class are called *class limits.* The **lower class limit** is 58; it represents the smallest data value that can be included in the class. The **upper class limit** is 64; it

represents the largest data value that can be included in the class. The numbers in the second column are called **class boundaries.** These numbers are used to separate the classes so that there are no gaps in the frequency distribution. The gaps are due to the limits; for example, there is a gap between 64 and 65.

Students sometimes have difficulty finding class boundaries when given the class limits. The basic rule of thumb is that *the class limits should have the same decimal place value as the data, but the class boundaries should have one additional place value and end in a 5.* For example, if the values in the data set are whole numbers, such as 59, 68, and 82, the limits for a class might be 58–64, and the boundaries are 57.5–64.5. Find the boundaries by subtracting 0.5 from 58 (the lower class limit) and adding 0.5 to 64 (the upper class limit).

$$\text{Lower limit} - 0.5 = 58 - 0.5 = 57.5 = \text{lower boundary}$$
$$\text{Upper limit} + 0.5 = 64 + 0.5 = 64.5 = \text{upper boundary}$$

Unusual Stats

One out of every hundred people in the United States is color-blind.

If the data are in tenths, such as 6.2, 7.8, and 12.6, the limits for a class hypothetically might be 7.8–8.8, and the boundaries for that class would be 7.75–8.85. Find these values by subtracting 0.05 from 7.8 and adding 0.05 to 8.8.

Class boundaries are not always included in frequency distributions; however, they give a more formal approach to the procedure of organizing data, including the fact that sometimes the data have been rounded. You should be familiar with boundaries since you may encounter them in a statistical study.

Finally, the **class width** for a class in a frequency distribution is found by subtracting the lower (or upper) class limit of one class from the lower (or upper) class limit of the next class. For example, the class width in the preceding distribution on the distribution of blood glucose levels is 7, found from 65 – 58 = 7.

The class width can also be found by subtracting the lower boundary from the upper boundary for any given class. In this case, 64.5 – 57.5 = 7.

Note: Do not subtract the limits of a single class. It will result in an incorrect answer.

The researcher must decide how many classes to use and the width of each class. To construct a frequency distribution, follow these rules:

1. *There should be between 5 and 20 classes.* Although there is no hard-and-fast rule for the number of classes contained in a frequency distribution, it is of utmost importance to have enough classes to present a clear description of the collected data.
2. *It is preferable but not absolutely necessary that the class width be an odd number.* This ensures that the midpoint of each class has the same place value as the data. The **class midpoint** X_m is obtained by adding the lower and upper boundaries and dividing by 2, or adding the lower and upper limits and dividing by 2:

$$X_m = \frac{\text{lower boundary} + \text{upper boundary}}{2}$$

or

$$X_m = \frac{\text{lower limit} + \text{upper limit}}{2}$$

For example, the midpoint of the first class in the example with glucose levels is

$$\frac{57.5 + 64.5}{2} = 61 \quad \text{or} \quad \frac{58 + 64}{2} = 61$$

The midpoint is the numeric location of the center of the class. Midpoints are necessary for graphing (see Section 2–2). If the class width is an even number, the

midpoint is in tenths. For example, if the class width is 6 and the boundaries are 5.5 and 11.5, the midpoint is

$$\frac{5.5 + 11.5}{2} = \frac{17}{2} = 8.5$$

Rule 2 is only a suggestion, and it is not rigorously followed, especially when a computer is used to group data.

3. *The classes must be mutually exclusive.* Mutually exclusive classes have nonoverlapping class limits so that data cannot be placed into two classes. Many times, frequency distributions such as this

Age
10–20
20–30
30–40
40–50

are found in the literature or in surveys. If a person is 40 years old, into which class should she or he be placed? A better way to construct a frequency distribution is to use classes such as

Age
10–20
21–31
32–42
43–53

Recall that boundaries are mutually exclusive. For example, when a class boundary is 5.5 to 10.5, the data values that are included in that class are values from 6 to 10. A data value of 5 goes into the previous class, and a data value of 11 goes into the next-higher class.

4. *The classes must be continuous.* Even if there are no values in a class, the class must be included in the frequency distribution. There should be no gaps in a frequency distribution. The only exception occurs when the class with a zero frequency is the first or last class. A class with a zero frequency at either end can be omitted without affecting the distribution.
5. *The classes must be exhaustive.* There should be enough classes to accommodate all the data.
6. *The classes must be equal in width.* This avoids a distorted view of the data.

   One exception occurs when a distribution has a class that is **open-ended.** That is, the first class has no specific lower limit, or the last class has no specific upper limit. A frequency distribution with an open-ended class is called an **open-ended distribution.** Here are two examples of distributions with open-ended classes.

Age	Frequency
10–20	3
21–31	6
32–42	4
43–53	10
54 and above	8

Minutes	Frequency
Below 110	16
110–114	24
115–119	38
120–124	14
125–129	5

The frequency distribution for age is open-ended for the last class, which means that anybody who is 54 years or older will be tallied in the last class. The distribution for minutes is open-ended for the first class, meaning that any minute values below 110 will be tallied in that class.

The steps for constructing a grouped frequency distribution are summarized in the following Procedure Table.

Procedure Table

Constructing a Grouped Frequency Distribution

Step 1 Determine the classes.
- Find the highest and lowest values.
- Find the range.
- Select the number of classes desired.
- Find the width by dividing the range by the number of classes and rounding up.
- Select a starting point (usually the lowest value or any convenient number less than the lowest value); add the width to get the lower limits.
- Find the upper class limits.
- Find the boundaries.

Step 2 Tally the data.

Step 3 Find the numerical frequencies from the tallies, and find the cumulative frequencies.

Example 2–2 shows the procedure for constructing a grouped frequency distribution, i.e., when the classes contain more than one data value.

EXAMPLE 2–2 Record High Temperatures

These data represent the record high temperatures in degrees Fahrenheit (°F) for each of the 50 states. Construct a grouped frequency distribution for the data, using 7 classes.

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

Source: The World Almanac and Book of Facts.

SOLUTION

The procedure for constructing a grouped frequency distribution for numerical data follows.

Unusual Stats

America's most popular beverages are soft drinks. It is estimated that, on average, each person drinks about 52 gallons of soft drinks per year, compared to 22 gallons of beer.

Step 1 Determine the classes.

Find the highest value and lowest value: $H = 134$ and $L = 100$.

Find the range: R = highest value − lowest value $= H - L$, so

$$R = 134 - 100 = 34$$

Select the number of classes desired (usually between 5 and 20). In this case, 7 is arbitrarily chosen.

Find the class width by dividing the range by the number of classes.

$$\text{Width} = \frac{R}{\text{number of classes}} = \frac{34}{7} = 4.9$$

Round the answer up to the nearest whole number if there is a remainder: $4.9 \approx 5$. (Rounding *up* is different from rounding *off*. A number is rounded up if there is any decimal remainder when dividing. For example, $85 \div 6 = 14.167$ and is rounded up to 15. Also, $53 \div 4 = 13.25$ and is rounded up to 14. (Also, after dividing, if there is no remainder, you will need to add an extra class to accommodate all the data.)

Select a starting point for the lowest class limit. This can be the smallest data value or any convenient number less than the smallest data value. In this case, 100 is used. Add the width to the lowest score taken as the starting point to get the lower limit of the next class. Keep adding until there are 7 classes, as shown, 100, 105, 110, etc.

Subtract one unit from the lower limit of the second class to get the upper limit of the first class. Then add the width to each upper limit to get all the upper limits.

$$105 - 1 = 104$$

The first class is 100–104, the second class is 105–109, etc.

Find the class boundaries by subtracting 0.5 from each lower class limit and adding 0.5 to each upper class limit:

99.5–104.5, 104.5–109.5, etc.

Historical Note

Florence Nightingale, a nurse in the Crimean War in 1854, used statistics to persuade government officials to improve hospital care of soldiers in order to reduce the death rate from unsanitary conditions in the military hospitals that cared for the wounded soldiers.

Step 2 Tally the data.

Step 3 Find the numerical frequencies from the tallies.

The completed frequency distribution is

Class limits	Class boundaries	Tally	Frequency
100–104	99.5–104.5	//	2
105–109	104.5–109.5	𝍸 ///	8
110–114	109.5–114.5	𝍸 𝍸 𝍸 ///	18
115–119	114.5–119.5	𝍸 𝍸 ///	13
120–124	119.5–124.5	𝍸 //	7
125–129	124.5–129.5	/	1
130–134	129.5–134.5	/	1
			Total 50

The frequency distribution shows that the class 109.5–114.5 contains the largest number of temperatures (18) followed by the class 114.5–119.5 with 13 temperatures. Hence, most of the temperatures (31) fall between 110 and 119°F.

Sometimes it is necessary to use a *cumulative frequency distribution*. A **cumulative frequency distribution** is a distribution that shows the number of data values less than or equal to a specific value (usually an upper boundary). The values are found by adding the frequencies of the classes less than or equal to the upper class boundary of a specific class. This gives an ascending cumulative frequency. In this example, the cumulative frequency for the first class is $0 + 2 = 2$; for the second class it is $0 + 2 + 8 = 10$; for the third class it is $0 + 2 + 8 + 18 = 28$. Naturally, a shorter way to do this would be to just add the cumulative frequency of the class below to the frequency of the given class. For example, the cumulative frequency for the number of data values less than 114.5 can be

found by adding 10 + 18 = 28. The cumulative frequency distribution for the data in this example is as follows:

	Cumulative frequency
Less than 99.5	0
Less than 104.5	2
Less than 109.5	10
Less than 114.5	28
Less than 119.5	41
Less than 124.5	48
Less than 129.5	49
Less than 134.5	50

Cumulative frequencies are used to show how many data values are accumulated up to and including a specific class. In Example 2–2, of the total record high temperatures 28 are less than or equal to 114°F. Forty-eight of the total record high temperatures are less than or equal to 124°F.

After the raw data have been organized into a frequency distribution, it will be analyzed by looking for peaks and extreme values. The peaks show which class or classes have the most data values compared to the other classes. Extreme values, called *outliers,* show large or small data values that are relative to other data values.

When the range of the data values is relatively small, a frequency distribution can be constructed using single data values for each class. This type of distribution is called an **ungrouped frequency distribution** and is shown next.

EXAMPLE 2–3 Hours of Sleep

The data shown represent the number of hours 30 college students said they sleep per night. Construct and analyze a frequency distribution.

8	6	6	8	5	7
7	8	7	6	6	7
9	7	7	6	8	10
6	7	6	7	8	7
7	8	7	8	9	8

SOLUTION

Step 1 Determine the number of classes. Since the range is small (10 − 5 = 5), classes consisting of a single data value can be used. They are 5, 6, 7, 8, 9, and 10.

Note: If the data are continuous, class boundaries can be used. Subtract 0.5 from each class value to get the lower class boundary, and add 0.5 to each class value to get the upper class boundary.

Step 2 Tally the data.

Step 3 From the tallies, find the numerical frequencies and cumulative frequencies. The completed ungrouped frequency distribution is shown.

Class limits	Class boundaries	Tally	Frequency
5	4.5–5.5	/	1
6	5.5–6.5	𝍸 //	7
7	6.5–7.5	𝍸 𝍸 /	11
8	7.5–8.5	𝍸 ///	8
9	8.5–9.5	//	2
10	9.5–10.5	/	1

In this case, 11 students sleep 7 hours a night. Most of the students sleep between 5.5 and 8.5 hours.

The cumulative frequencies are

	Cumulative frequency
Less than 4.5	0
Less than 5.5	1
Less than 6.5	8
Less than 7.5	19
Less than 8.5	27
Less than 9.5	29
Less than 10.5	30

When you are constructing a frequency distribution, the guidelines presented in this section should be followed. However, you can construct several different but correct frequency distributions for the same data by using a different class width, a different number of classes, or a different starting point.

Interesting Fact

Male dogs bite children more often than female dogs do; however, female cats bite children more often than male cats do.

Furthermore, the method shown here for constructing a frequency distribution is not unique, and there are other ways of constructing one. Slight variations exist, especially in computer packages. But regardless of what methods are used, classes should be mutually exclusive, continuous, exhaustive, and of equal width.

In summary, the different types of frequency distributions were shown in this section. The first type, shown in Example 2–1, is used when the data are categorical (nominal), such as blood type or political affiliation. This type is called a categorical frequency distribution. The second type of distribution is used when the range is large and classes several units in width are needed. This type is called a grouped frequency distribution and is shown in Example 2–2. Another type of distribution is used for numerical data and when the range of data is small, as shown in Example 2–3. Since each class is only one unit, this distribution is called an ungrouped frequency distribution.

All the different types of distributions are used in statistics and are helpful when one is organizing and presenting data.

The reasons for constructing a frequency distribution are as follows:

1. To organize the data in a meaningful, intelligible way.
2. To enable the reader to determine the nature or shape of the distribution.
3. To facilitate computational procedures for measures of average and spread (shown in Sections 3–1 and 3–2).

4. To enable the researcher to draw charts and graphs for the presentation of data (shown in Section 2–2).
5. To enable the reader to make comparisons among different data sets.

The factors used to analyze a frequency distribution are essentially the same as those used to analyze histograms and frequency polygons, which are shown in Section 2–2.

Applying the Concepts 2–1

Ages of Presidents at Inauguration

The data represent the ages of our Presidents at the time they were first inaugurated.

57	61	57	57	58	57	61	54	68
51	49	64	50	48	65	52	56	46
54	49	51	47	55	55	54	42	51
56	55	51	54	51	60	62	43	55
56	61	52	69	64	46	54	47	

1. Were the data obtained from a population or a sample? Explain your answer.
2. What was the age of the oldest President?
3. What was the age of the youngest President?
4. Construct a frequency distribution for the data. (Use your own judgment as to the number of classes and class size.)
5. Are there any peaks in the distribution?
6. Identify any possible outliers.
7. Write a brief summary of the nature of the data as shown in the frequency distribution.

See page 108 for the answers.

Exercises 2–1

1. List five reasons for organizing data into a frequency distribution.
2. Name the three types of frequency distributions, and explain when each should be used.
3. How many classes should frequency distributions have? Why should the class width be an odd number?
4. What are open-ended frequency distributions? Why are they necessary?

For Exercises 5–8, find the class boundaries, midpoints, and widths for each class.

5. 58–62
6. 125–131
7. 16.35–18.46
8. 16.3–18.5

For Exercises 9–12, show frequency distributions that are incorrectly constructed. State the reasons why they are wrong.

9.

Class	Frequency
10–19	1
20–29	2
30–34	0
35–45	5
46–51	8

10.

Class	Frequency
5–9	1
9–13	2
13–17	5
17–20	6
20–24	3

11.

Class	Frequency
162–164	3
165–167	7
168–170	18
174–176	0
177–179	5

12.

Class	Frequency
9–13	1
14–19	6
20–25	2
26–28	5
29–32	9

13. Favorite Coffee Flavor A survey was taken asking the favorite flavor of a coffee drink a person prefers. The responses were V = Vanilla, C = Caramel, M = Mocha, H = Hazelnut, and P = Plain. Construct a categorical frequency distribution for the data. Which class has the most data values and which class has the fewest data values?

V	C	P	P	M	M	P	P	M	C
M	M	V	M	M	M	V	M	M	M
P	V	C	M	V	M	C	P	M	P
M	M	M	P	M	M	C	V	M	C
C	P	M	P	M	H	H	P	H	P

14. Trust in Internet Information A survey was taken on how much trust people place in the information they read on the Internet. Construct a categorical frequency distribution for the data. A = trust in all that they read, M = trust in most of what they read, H = trust in about one-half of what they read, S = trust in a small portion of what they read. (Based on information from the *UCLA Internet Report.*)

M	M	M	A	H	M	S	M	H	M
S	M	M	M	M	A	M	M	A	M
M	M	H	M	M	M	H	M	H	M
A	M	M	M	H	M	M	M	M	M

15. Eating at Fast Food Restaurants A survey was taken of 50 individuals. They were asked how many days per week they ate at a fast-food restaurant. Construct a frequency distribution using 8 classes (0–7). Based on the distribution, how often did most people eat at a fast-food restaurant?

1	3	4	0	4
5	2	2	3	1
2	2	2	2	2
2	2	2	2	3
2	2	5	2	4
2	4	5	2	1
4	1	3	2	2
2	0	7	2	3
2	2	2	5	2
3	3	4	1	3

16. Ages of Dogs The ages of 20 dogs in a pet shelter are shown. Construct a frequency distribution using 7 classes.

5	8	7	6	3
9	4	4	5	8
7	4	7	5	7
3	5	8	4	9

17. Maximum Wind Speeds The data show the maximum wind speeds in miles per hour recorded for 40 states. Construct a frequency distribution using 7 classes.

59	78	62	72	67
76	92	77	64	83
64	70	67	75	75
78	75	71	72	93
68	69	76	72	85
64	70	77	74	72
53	67	48	76	59
87	53	77	70	63

Source: NOAA

18. Stories in the World's Tallest Buildings The number of stories in each of a sample of the world's 30 tallest buildings follows. Construct a grouped frequency distribution and a cumulative frequency distribution with 7 classes.

88	88	110	88	80	69	102	78	70	55
79	85	80	100	60	90	77	55	75	55
54	60	75	64	105	56	71	70	65	72

Source: New York Times Almanac.

19. Ages of Declaration of Independence Signers The ages of the signers of the Declaration of Independence are shown. (Age is approximate since only the birth year appeared in the source, and one has been omitted since his birth year is unknown.) Construct a grouped frequency distribution and a cumulative frequency distribution for the data, using 7 classes.

41	54	47	40	39	35	50	37	49	42	70	32
44	52	39	50	40	30	34	69	39	45	33	42
44	63	60	27	42	34	50	42	52	38	36	45
35	43	48	46	31	27	55	63	46	33	60	62
35	46	45	34	53	50	50					

Source: The Universal Almanac.

20. Salaries of Governors Here are the salaries (in dollars) of the governors of 25 randomly selected states. Construct a grouped frequency distribution with 6 classes.

112,895	117,312	140,533	110,000	115,331
95,000	177,500	120,303	139,590	150,000
173,987	130,000	133,821	144,269	142,542
150,000	145,885	105,000	93,600	166,891
130,273	70,000	113,834	117,817	137,092

Source: World Almanac.

21. Charity Donations A random sample of 30 large companies in the United States shows the amount,

in millions of dollars, that each company donated to charity for a specific year. Construct a frequency distribution for the data, using 9 classes.

26	25	19	31	14
48	35	43	25	46
17	21	57	58	34
41	12	27	15	53
16	63	82	23	52
56	75	19	26	88

22. Unclaimed Expired Prizes The number of unclaimed expired prizes (in millions of dollars) for lottery tickets bought in a sample of states is shown. Construct a frequency distribution for the data, using 5 classes.

28.5	51.7	19	5
2	1.2	14	14.6
0.8	11.6	3.5	30.1
1.7	1.3	13	14

23. Scores in the Rose Bowl The data show the scores of the winning teams in the Rose Bowl. Construct a frequency distribution for the data using a class width of 7.

24	20	45	21	26	38	49	32	41	38
28	34	37	34	17	38	21	20	41	38
21	38	34	46	17	22	20	22	45	20
45	24	28	23	17	17	27	14	23	18

Source: The World Almanac.

24. Consumption of Natural Gas Construct a frequency distribution for the energy consumption of natural gas (in billions of Btu) by the 50 states and the District of Columbia. Use 9 classes.

474	475	205	639	197	344	3	409	247	66
377	87	747	1166	223	248	958	406	251	3462
2391	514	371	58	224	530	317	267	769	9
188	289	76	678	331	52	214	165	255	319
34	1300	284	834	114	1082	73	62	95	393
146									

Source: Time Almanac.

25. Average Wind Speeds A sample of 40 large cities was selected, and the average of the wind speeds was computed for each city over one year. Construct a frequency distribution, using 7 classes.

12.2	9.1	11.2	9.0
10.5	8.2	8.9	12.2
9.5	10.2	7.1	11.0
6.2	7.9	8.7	8.4
8.9	8.8	7.1	10.1
8.7	10.5	10.2	10.7
7.9	8.3	8.7	8.7
10.4	7.7	12.3	10.7
7.7	7.8	11.8	10.5
9.6	9.6	8.6	10.3

Source: World Almanac and Book of Facts.

26. Percentage of People Who Completed 4 or More Years of College Listed by state are the percentages of the population who have completed 4 or more years of a college education. Construct a frequency distribution with 7 classes.

21.4	26.0	25.3	19.3	29.5	35.0	34.7	26.1	25.8	23.4
27.1	29.2	24.5	29.5	22.1	24.3	28.8	20.0	20.4	26.7
35.2	37.9	24.7	31.0	18.9	24.5	27.0	27.5	21.8	32.5
33.9	24.8	31.7	25.6	25.7	24.1	22.8	28.3	25.8	29.8
23.5	25.0	21.8	25.2	28.7	33.6	33.6	30.3	17.3	25.4

Source: New York Times Almanac.

Extending the Concepts

27. JFK Assassination A researcher conducted a survey asking people if they believed more than one person was involved in the assassination of John F. Kennedy. The results were as follows: 73% said yes, 19% said no, and 9% had no opinion. Is there anything suspicious about the results?

28. The Value of Pi The ratio of the circumference of a circle to its diameter is known as π (pi). The value of π is an irrational number, which means that the decimal part goes on forever and there is no fixed sequence of numbers that repeats. People have found the decimal part of π to over a million places. We can statistically study the number. Shown here is the value of π to 40 decimal places. Construct an ungrouped frequency distribution for the digits. Based on the distribution, do you think each digit appears equally in the number?

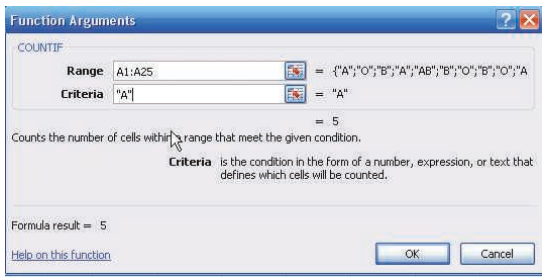
3.1415926535897932384626433832795028841971

Technology Step by Step

EXCEL Step by Step

Categorical Frequency Table (Qualitative or Discrete Data)

1. In an open workbook, select cell A1 and type in all the blood types from Example 2–1 down column A.
2. Type in the variable name **Blood Type** in cell B1.
3. Select cell B2 and type in the four different blood types down the column.
4. Type in the name **Count** in cell C1.
5. Select cell C2. From the toolbar, select the Formulas tab on the toolbar.
6. Select the Insert Function icon , then select the Statistical category in the Insert Function dialog box.
7. Select the Countif function from the function name list.
8. In the dialog box, type **A1:A25** in the **Range** box. Type in the blood type "A" in quotes in the **Criteria** box. The count or frequency of the number of data corresponding to the blood type should appear below the input. Repeat for the remaining blood types.
9. After all the data have been counted, select cell C6 in the worksheet.
10. From the toolbar select Formulas, then AutoSum and type in C2:C5 to insert the total frequency into cell C6.



After entering data or a heading into a worksheet, you can change the width of a column to fit the input. To automatically change the width of a column to fit the data:

1. Select the column or columns that you want to change.
2. On the Home tab, in the Cells group, select Format.
3. Under Cell Size, click Autofit Column Width.

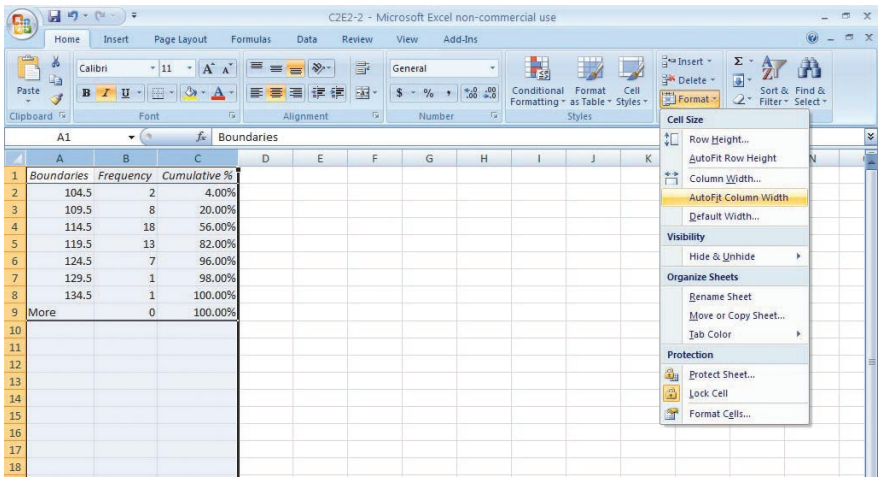
Making a Grouped Frequency Distribution (Quantitative Data)

1. Press [**Ctrl**]-**N** for a new workbook.
2. Enter the raw data from Example 2–2 in column A, one number per cell.
3. Enter the upper class boundaries in column B.
4. From the toolbar select the Data tab, then click Data Analysis.
5. In the Analysis Tools, select Histogram and click [OK].
6. In the Histogram dialog box, type **A1:A50** in the Input Range box and type **B1:B7** in the Bin Range box.
7. Select New Worksheet Ply, and check the Cumulative Percentage option. Click [OK].
8. You can change the label for the column containing the upper class boundaries and expand the width of the columns automatically after relabeling:

   Select the Home tab from the toolbar.

Highlight the columns that you want to change.

Select Format, then AutoFit Column Width.



Note: By leaving the Chart Output unchecked, a new worksheet will display the table only.

MINITAB

Step by Step

Make a Categorical Frequency Table (Qualitative or Discrete Data)

1. Type in all the blood types from Example 2–1 down C1 of the worksheet.
   A B B AB O O O B AB B B B O A O A O O O AB AB A O B A
2. Click above row 1 and name the column **BloodType.**
3. Select **Stat>Tables>Tally Individual Values.**
   The cursor should be blinking in the Variables dialog box. If not, click inside the dialog box.
4. Double-click C1 in the Variables list.
5. Check the boxes for the statistics: Counts, Percents, and Cumulative percents.
6. Click [OK]. The results will be displayed in the Session Window as shown.

Tally for Discrete Variables: BloodType

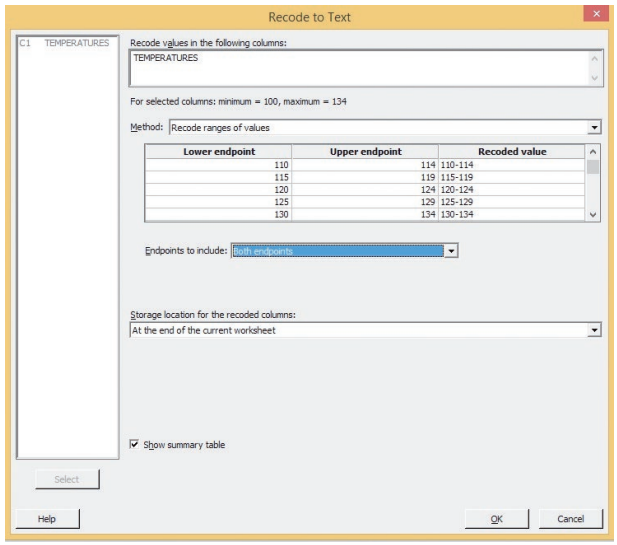
BloodType	Count	Percent	CumPct
A	5	20.00	20.00
AB	4	16.00	36.00
B	7	28.00	64.00
O	9	36.00	100.00
N=	25		

Make a Grouped Frequency Distribution (Quantitative Variable)

1. Select **File>New>Minitab Worksheet**. A new worksheet will be added to the project.
2. Type the data used in Example 2–2 into C1. Name the column **TEMPERATURES.**
3. Use the instructions in the textbook to determine the class limits of 100 to 134 in increments of 5.
   In the next step you will create a new column of data, converting the numeric variable to text categories that can be tallied.

4. Select **Data>Recode>to Text**.
   a) The cursor should be blinking in Recode values in the following columns. If not, click inside the box, then double-click C1 Temperatures in the list. Only quantitative variables will be shown in this list.
   b) Click inside the Method: box and select **Recode ranges of values.**
   c) Press [Tab] to move to the table.
   d) Type 100 in the Lower endpoint column, press [Tab], type 104 in the Upper endpoint column.
   e) Press [Tab] to move to the Recoded value column, and type the text category **100–104.**
   f) Continue to tab to each dialog box, typing the lower endpoint and upper endpoint and then the category until the last category has been entered.
   g) Click inside the Endpoints to include: box and select **Both endpoints.**

   The dialog box should look like the one shown.



5. Click [OK]. In the worksheet, a new column of data will be created in the first empty column, C2. This new variable will contain the category for each value in C1. The column C2-T contains alphanumeric data.
6. Click Stat>Tables>Tally Individual Values, then double-click Recoded TEMPERATURES in the Variables list.
   a) Check the boxes for the desired statistics, such as Counts, Percents, and Cumulative percents.
   b) Click [OK].

   The table will be displayed in the Session Window. Eighteen states have high temperatures between 110 and 114°F. Eighty-two percent of the states have record high temperatures less than or equal to 119°F.

Tally for Discrete Variables: Recoded TEMPERATURES

Recoded TEMPERATURES	Count	Percent	CumPct
100–104	2	4.00	4.00
105–109	8	16.00	20.00
110–114	18	36.00	56.00
115–119	13	26.00	82.00
120–124	7	14.00	96.00
125–129	1	2.00	98.00
130–134	1	2.00	100.00
N=	50		

7. Click File>Save Project As . . . , and type the name of the project file, **Ch2-1.** This will save the two worksheets and the Session Window.

2–2 Histograms, Frequency Polygons, and Ogives

OBJECTIVE 2

Represent data in frequency distributions graphically, using histograms, frequency polygons, and ogives.

After you have organized the data into a frequency distribution, you can present them in graphical form. The purpose of graphs in statistics is to convey the data to the viewers in pictorial form. It is easier for most people to comprehend the meaning of data presented graphically than data presented numerically in tables or frequency distributions. This is especially true if the users have little or no statistical knowledge.

Statistical graphs can be used to describe the data set or to analyze it. Graphs are also useful in getting the audience's attention in a publication or a speaking presentation. They can be used to discuss an issue, reinforce a critical point, or summarize a data set. They can also be used to discover a trend or pattern in a situation over a period of time.

The three most commonly used graphs in research are

1. The histogram.
2. The frequency polygon.
3. The cumulative frequency graph, or ogive (pronounced o-jive).

The steps for constructing the histogram, frequency polygon, and the ogive are summarized in the procedure table.

Procedure Table

Constructing a Histogram, Frequency Polygon, and Ogive

Step 1	Draw and label the x and y axes.
Step 2	On the x axis, label the class boundaries of the frequency distribution for the histogram and ogive. Label the midpoints for the frequency polygon.
Step 3	Plot the frequencies for each class, and draw the vertical bars for the histogram and the lines for the frequency polygon and ogive.

(*Note:* Remember that the lines for the frequency polygon begin and end on the x axis while the lines for the ogive begin on the x axis.)

Historical Note

Karl Pearson introduced the histogram in 1891. He used it to show time concepts of various reigns of Prime Ministers.

The Histogram

The **histogram** is a graph that displays the data by using contiguous vertical bars (unless the frequency of a class is 0) of various heights to represent the frequencies of the classes.

EXAMPLE 2–4 Record High Temperatures

Construct a histogram to represent the data shown for the record high temperatures for each of the 50 states (see Example 2–2).

Class boundaries	Frequency
99.5–104.5	2
104.5–109.5	8
109.5–114.5	18
114.5–119.5	13
119.5–124.5	7
124.5–129.5	1
129.5–134.5	1

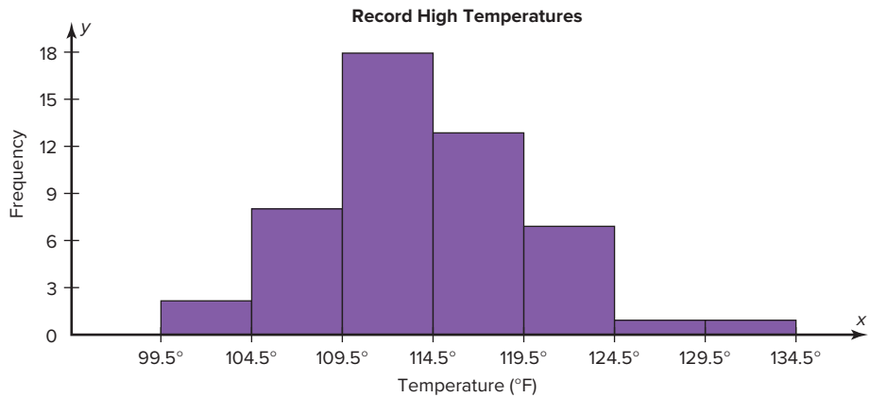
SOLUTION

Step 1 Draw and label the x and y axes. The x axis is always the horizontal axis, and the y axis is always the vertical axis.

Step 2 Represent the frequency on the y axis and the class boundaries on the x axis.

Step 3 Using the frequencies as the heights, draw vertical bars for each class. See Figure 2–1.

FIGURE 2–1 Histogram for Example 2–4



As the histogram shows, the class with the greatest number of data values (18) is 109.5–114.5, followed by 13 for 114.5–119.5. The graph also has one peak with the data clustering around it.

Historical Note

Graphs originated when ancient astronomers drew the position of the stars in the heavens. Roman surveyors also used coordinates to locate landmarks on their maps.

The development of statistical graphs can be traced to William Playfair (1759–1823), an engineer and drafter who used graphs to present economic data pictorially.

The Frequency Polygon

Another way to represent the same data set is by using a frequency polygon.

The **frequency polygon** is a graph that displays the data by using lines that connect points plotted for the frequencies at the midpoints of the classes. The frequencies are represented by the heights of the points.

Example 2–5 shows the procedure for constructing a frequency polygon. Be sure to begin and end on the x axis.

EXAMPLE 2–5 Record High Temperatures

Using the frequency distribution given in Example 2–4, construct a frequency polygon.

SOLUTION

Step 1 Find the midpoints of each class. Recall that midpoints are found by adding the upper and lower boundaries and dividing by 2:

$$\frac{99.5 + 104.5}{2} = 102 \qquad \frac{104.5 + 109.5}{2} = 107$$

and so on. The midpoints are

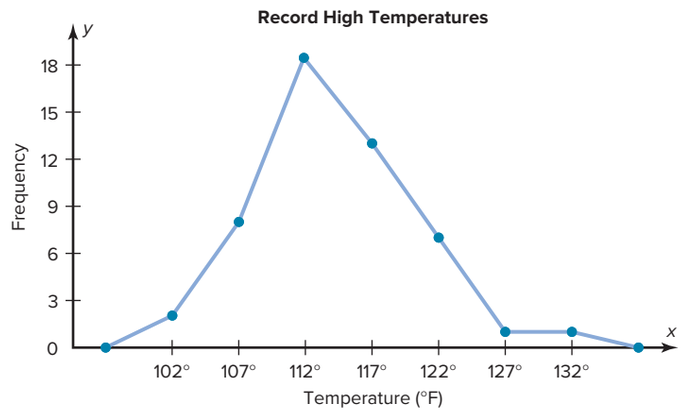
Class boundaries	Midpoints	Frequency
99.5–104.5	102	2
104.5–109.5	107	8
109.5–114.5	112	18
114.5–119.5	117	13
119.5–124.5	122	7
124.5–129.5	127	1
129.5–134.5	132	1

Step 2 Draw the *x* and *y* axes. Label the *x* axis with the midpoint of each class, and then use a suitable scale on the *y* axis for the frequencies.

Step 3 Using the midpoints for the *x* values and the frequencies as the *y* values, plot the points.

Step 4 Connect adjacent points with line segments. Draw a line back to the *x* axis at the beginning and end of the graph, at the same distance that the previous and next midpoints would be located, as shown in Figure 2–2.

FIGURE 2–2
Frequency Polygon for Example 2–5



The frequency polygon and the histogram are two different ways to represent the same data set. The choice of which one to use is left to the discretion of the researcher.

The Ogive

The third type of graph that can be used represents the cumulative frequencies for the classes. This type of graph is called the *cumulative frequency graph*, or *ogive*. The **cumulative frequency** is the sum of the frequencies accumulated up to the upper boundary of a class in the distribution.

The **ogive** is a graph that represents the cumulative frequencies for the classes in a frequency distribution.

Example 2–6 shows the procedure for constructing an ogive. Be sure to start on the *x* axis.

EXAMPLE 2–6 Record High Temperatures

Construct an ogive for the frequency distribution described in Example 2–4.

SOLUTION

Step 1 Find the cumulative frequency for each class.

	Cumulative frequency
Less than 99.5	0
Less than 104.5	2
Less than 109.5	10
Less than 114.5	28
Less than 119.5	41
Less than 124.5	48
Less than 129.5	49
Less than 134.5	50

Step 2 Draw the x and y axes. Label the x axis with the class boundaries. Use an appropriate scale for the y axis to represent the cumulative frequencies. (Depending on the numbers in the cumulative frequency columns, scales such as 0, 1, 2, 3, . . . , or 5, 10, 15, 20, . . . , or 1000, 2000, 3000, . . . can be used. Do *not* label the y axis with the numbers in the cumulative frequency column.) In this example, a scale of 0, 5, 10, 15, . . . will be used.

Step 3 Plot the cumulative frequency at each upper class boundary, as shown in Figure 2–3. Upper boundaries are used since the cumulative frequencies represent the number of data values accumulated up to the upper boundary of each class.

Step 4 Starting with the first upper class boundary, 104.5, connect adjacent points with line segments, as shown in Figure 2–4. Then extend the graph to the first lower class boundary, 99.5, on the x axis.

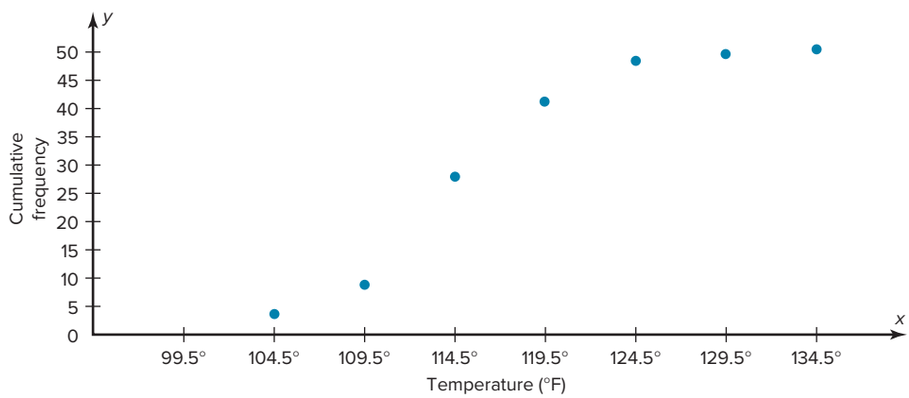


FIGURE 2–3
Plotting the Cumulative Frequency for Example 2–6

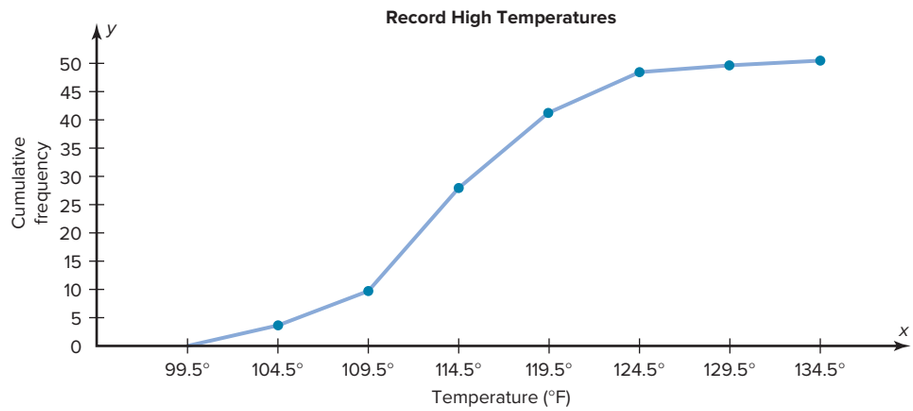
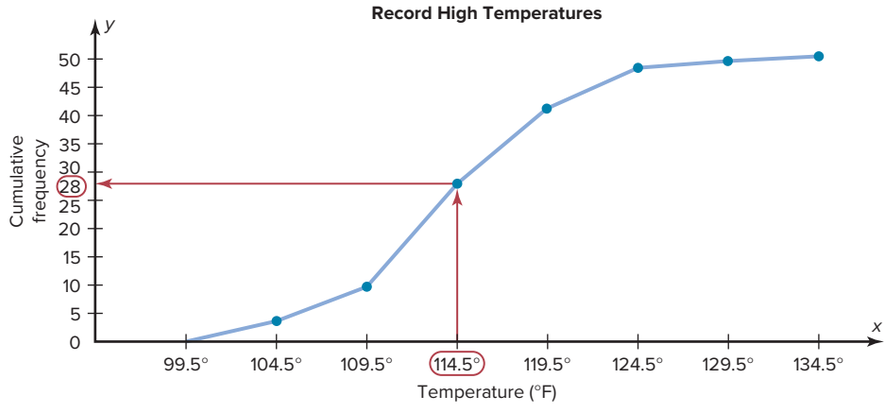


FIGURE 2–4
Ogive for Example 2–6

FIGURE 2–5
Finding a Specific Cumulative Frequency



Unusual Stats
Twenty-two percent of Americans sleep 6 hours a day or less.

Cumulative frequency graphs are used to visually represent how many values are below a certain upper class boundary. For example, to find out how many record high temperatures are less than 114.5°F, locate 114.5°F on the *x* axis, draw a vertical line up until it intersects the graph, and then draw a horizontal line at that point to the *y* axis. The *y* axis value is 28, as shown in Figure 2–5.

Relative Frequency Graphs

The histogram, the frequency polygon, and the ogive shown previously were constructed by using frequencies in terms of the raw data. These distributions can be converted to distributions using *proportions* instead of raw data as frequencies. These types of graphs are called **relative frequency graphs.**

Graphs of relative frequencies instead of frequencies are used when the proportion of data values that fall into a given class is more important than the actual number of data values that fall into that class. For example, if you wanted to compare the age distribution of adults in Philadelphia, Pennsylvania, with the age distribution of adults of Erie, Pennsylvania, you would use relative frequency distributions. The reason is that since the population of Philadelphia is 1,526,006 and the population of Erie is 101,786, the bars using the actual data values for Philadelphia would be much taller than those for the same classes for Erie.

To convert a frequency into a proportion or relative frequency, divide the frequency for each class by the total of the frequencies. The sum of the relative frequencies will always be 1. These graphs are similar to the ones that use raw data as frequencies, but the values on the *y* axis are in terms of proportions. Example 2–7 shows the three types of relative frequency graphs.

EXAMPLE 2–7 Ages of State Governors

Construct at histogram, frequency polygon, and ogive using relative frequencies for the distribution shown. This is a grouped frequency distribution using the ages (at the time of this writing) of the governors of the 50 states of the United States.

Class boundaries	Frequency
42.5–47.5	4
47.5–52.5	4
52.5–57.5	11
57.5–62.5	14
62.5–67.5	9
67.5–72.5	5
72.5–77.5	3
	Total 50

SOLUTION

Step 1 Convert each frequency to a proportion or relative frequency by dividing the frequency for each class by the total number of observations.

For the class 42.5–47.5 the relative frequency $= \frac{4}{50} = 0.08$; for the class 47.5–52.5, the relative frequency is $\frac{4}{50} = 0.08$; for the class 52.5–57.5, the relative frequency is $\frac{11}{50} = 0.22$, and so on.

Place these values in the column labeled Relative Frequency. Also, find the midpoints, as shown in Example 2-5, for each class and place them in the midpoint column

Class boundaries	Midpoints	Relative frequency
42.5–47.5	45	0.08
47.5–52.5	50	0.08
52.5–57.5	55	0.22
57.5–62.5	60	0.28
62.5–67.5	65	0.18
67.5–72.5	70	0.10
72.5–77.5	75	0.06

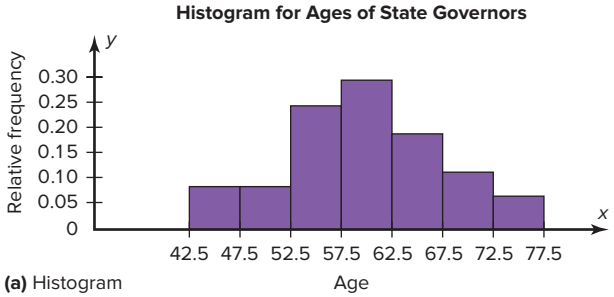
Step 2 Find the cumulative relative frequencies. To do this, add the frequency in each class to the total frequency of the preceding class. In this case, 0.00 + 0.08 = 0.08, 0.08 + 0.08 = 0.16, 0.16 + 0.22 = 0.38, 0.28 + 0.38 = 0.66, etc. Place these values in a column labeled Cumulative relative frequency.

An alternative method would be to change the cumulative frequencies for the classes to relative frequencies. (Divide each by the total).

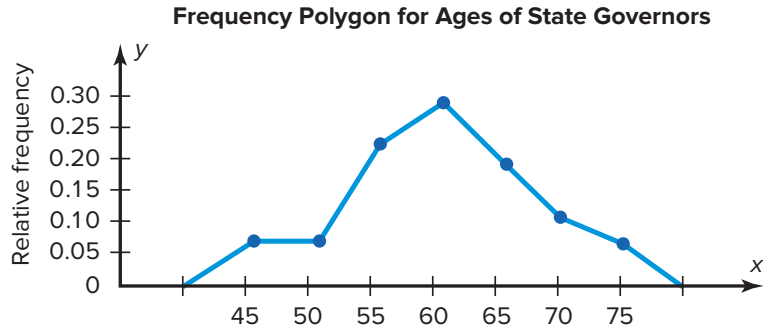
	Cumulative frequency	Cumulative relative frequency
Less than 42.5	0	0.00
Less than 47.5	4	0.08
Less than 52.5	8	0.16
Less than 57.5	19	0.38
Less than 62.5	33	0.66
Less than 67.5	42	0.84
Less than 72.5	47	0.94
Less than 77.5	50	1.00

Step 3 Draw each graph as shown in Figure 2-6. For the histogram and ogive, use the class boundaries along the *x* axis. For the frequency, use the midpoints on the *x* axis. For the scale on the *y* axis, use proportions.

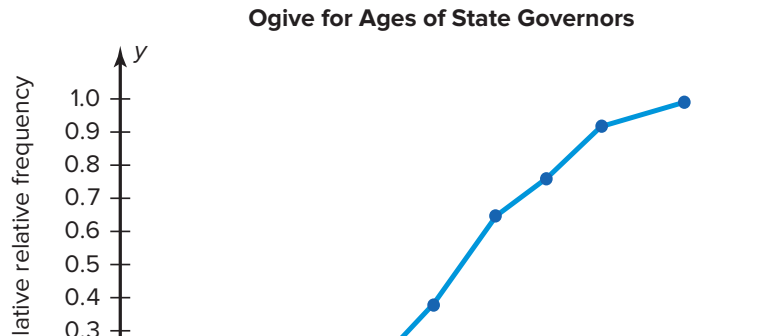
FIGURE 2–6
Graphs for Example 2–7



(a) Histogram



(b) Frequency polygon



(c) Ogive

Distribution Shapes

When one is describing data, it is important to be able to recognize the shapes of the distribution values. In later chapters, you will see that the shape of a distribution also determines the appropriate statistical methods used to analyze the data.

A distribution can have many shapes, and one method of analyzing a distribution is to draw a histogram or frequency polygon for the distribution. Several of the most common shapes are shown in Figure 2–7: *the bell-shaped or mound-shaped, the uniform-shaped, the J-shaped, the reverse J-shaped, the positively or right-skewed shape, the negatively or left-skewed shape, the bimodal-shaped, and the U-shaped.*

Distributions are most often not perfectly shaped, so it is not necessary to have an exact shape but rather to identify an overall pattern.

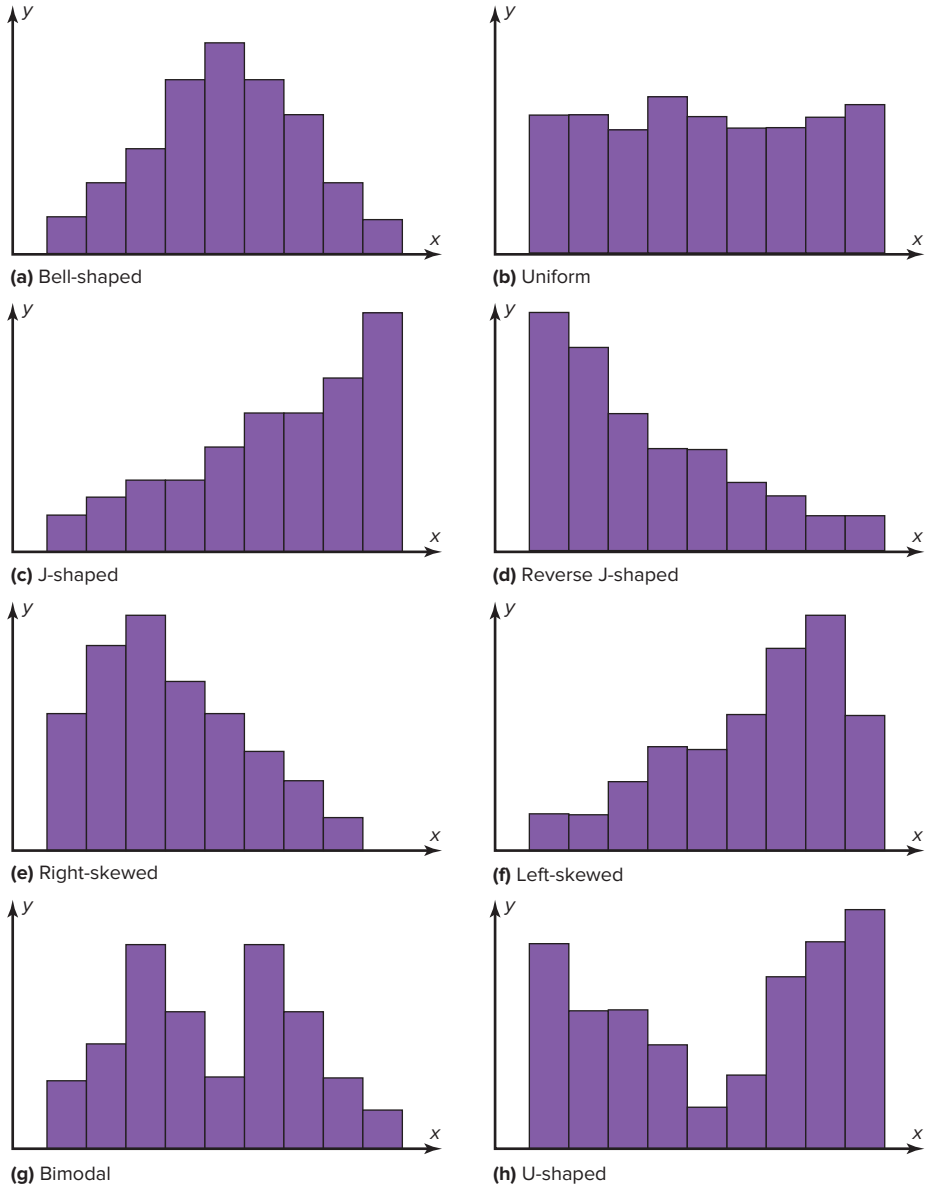
A *bell-shaped distribution* shown in Figure 2–7(a) has a single peak and tapers off at either end. It is approximately symmetric; i.e., it is roughly the same on both sides of a line running through the center.

A *uniform distribution* is basically flat or rectangular. See Figure 2–7(b).

A *J-shaped distribution* is shown in Figure 2–7(c), and it has a few data values on the left side and increases as one moves to the right. A *reverse J-shaped distribution* is the opposite of the J-shaped distribution. See Figure 2–7(d).

When the peak of a distribution is to the left and the data values taper off to the right, a distribution is said to be *positively or right-skewed.* See Figure 2–7(e). When

FIGURE 2–7
Distribution Shapes



the data values are clustered to the right and taper off to the left, a distribution is said to be *negatively or left-skewed.* See Figure 2–7(f). Skewness will be explained in detail in Chapter 3. Distributions with one peak, such as those shown in Figure 2–7(a), (e), and (f), are said to be *unimodal.* (The highest peak of a distribution indicates where the mode of the data values is. The mode is the data value that occurs more often than any other data value. Modes are explained in Chapter 3.) When a distribution has two peaks of the same height, it is said to be *bimodal.* See Figure 2–7(g). Finally, the graph shown in Figure 2–7(h) is a *U-shaped* distribution.

Distributions can have other shapes in addition to the ones shown here; however, these are some of the more common ones that you will encounter in analyzing data.

When you are analyzing histograms and frequency polygons, look at the shape of the curve. For example, does it have one peak or two peaks? Is it relatively flat, or is it U-shaped? Are the data values spread out on the graph, or are they clustered around the center? Are there data values in the extreme ends? These may be *outliers.* (See Section 3–3 for an explanation of outliers.) Are there any gaps in the histogram, or does

the frequency polygon touch the *x* axis somewhere other than at the ends? Finally, are the data clustered at one end or the other, indicating a *skewed distribution*?

For example, the histogram for the record high temperatures in Figure 2–1 shows a single peaked distribution, with the class 109.5–114.5 containing the largest number of temperatures. The distribution has no gaps, and there are fewer temperatures in the highest class than in the lowest class.

Applying the Concepts 2–2

Selling Real Estate

Assume you are a realtor in Bradenton, Florida. You have recently obtained a listing of the selling prices of the homes that have sold in that area in the last 6 months. You wish to organize those data so you will be able to provide potential buyers with useful information. Use the following data to create a histogram, frequency polygon, and cumulative frequency polygon.

142,000	127,000	99,600	162,000	89,000	93,000	99,500
73,800	135,000	119,500	67,900	156,300	104,500	108,650
123,000	91,000	205,000	110,000	156,300	104,000	133,900
179,000	112,000	147,000	321,550	87,900	88,400	180,000
159,400	205,300	144,400	163,000	96,000	81,000	131,000
114,000	119,600	93,000	123,000	187,000	96,000	80,000
231,000	189,500	177,600	83,400	77,000	132,300	166,000

1. What questions could be answered more easily by looking at the histogram rather than the listing of home prices?
2. What different questions could be answered more easily by looking at the frequency polygon rather than the listing of home prices?
3. What different questions could be answered more easily by looking at the cumulative frequency polygon rather than the listing of home prices?
4. Are there any extremely large or extremely small data values compared to the other data values?
5. Which graph displays these extremes the best?
6. Is the distribution skewed?

See page 108 for the answers.

Exercises 2–2

1. Do Students Need Summer Development? For 108 randomly selected college applicants, the following frequency distribution for entrance exam scores was obtained. Construct a histogram, frequency polygon, and ogive for the data. (The data for this exercise will be used for Exercise 13 in this section.)

Class limits	Frequency
90–98	6
99–107	22
108–116	43
117–125	28
126–134	9
	Total 108

Applicants who score above 107 need not enroll in a summer developmental program. In this group, how many students do not have to enroll in the developmental program?

2. Bear Kills The number of bears killed in 2014 for 56 counties in Pennsylvania is shown in the frequency distribution. Construct a histogram, frequency polygon, and ogive for the data. Comment on the skewness of the distribution. How many counties had 75 or fewer bears killed? (The data for this exercise will be used for Exercise 14 of this section.)

Class limits	Frequency
1–25	16
26–50	14
51–75	9
76–100	8
101–125	5
126–150	0
151–175	1
176–200	1
201–225	0
226–250	0
251–275	2
	Total 56

Source: Pennsylvania State Game Commission.

3. Pupils Per Teacher The average number of pupils per teacher in each state is shown. Construct a grouped frequency distribution with 6 classes. Draw a histogram, frequency polygon, and ogive. Analyze the distribution.

16	16	15	12	14
13	16	14	15	14
18	18	18	12	15
15	16	16	15	15
25	19	15	12	22
18	14	13	17	9
13	14	13	16	12
14	16	10	22	20
12	14	18	15	14
16	12	12	13	15

Source: U.S. Department of Education.

4. Number of College Faculty The number of faculty listed for a sample of private colleges that offer only bachelor's degrees is listed below. Use these data to construct a frequency distribution with 7 classes, a histogram, a frequency polygon, and an ogive. Discuss the shape of this distribution. What proportion of schools have 180 or more faculty?

165	221	218	206	138	135	224	204
70	210	207	154	155	82	120	116
176	162	225	214	93	389	77	135
221	161	128	310				

Source: World Almanac and Book of Facts.

5. Railroad Crossing Accidents The data show the number of railroad crossing accidents for the 50 states of the United States for a specific year. Construct a histogram, frequency polygon, and ogive for the data. Comment on the skewness of the distribution. (The data in this exercise will be used for Exercise 15 in this section.)

Class limits	Frequency
1–43	24
44–86	17
87–129	3
130–172	4
173–215	1
216–258	0
259–301	0
302–344	1
	Total 50

Source: Federal Railroad Administration.

6. NFL Salaries The salaries (in millions of dollars) for 31 NFL teams for a specific season are given in this frequency distribution.

Construct a histogram, a frequency polygon, and an ogive for the data; and comment on the shape of the distribution. (The data for this exercise will be used for Exercise 16 of this section.)

Class limits	Frequency
39.9–42.8	2
42.9–45.8	2
45.9–48.8	5
48.9–51.8	5
51.9–54.8	12
54.9–57.8	5
	Total 31

Source: NFL.com

7. Suspension Bridges Spans The following frequency distribution shows the length (in feet) of the main spans of the longest suspension bridges in the United States. Construct a histogram, frequency polygon, and ogive for the distribution. Describe the shape of the distribution.

Class limits	Frequency
1260–1734	12
1735–2209	6
2210–2684	3
2685–3159	1
3160–3634	1
3635–4109	1
4110–4584	2

Source: U.S. Department of Transportation.

8. Costs of Utilities The frequency distribution represents the cost (in cents) for the utilities of states that supply much of their own power. Construct a histogram, frequency polygon, and ogive for the data. Is the distribution skewed?

Class limits	Frequency
6–8	12
9–11	16
12–14	3
15–17	1
18–20	0
21–23	0
24–26	1
	Total 33

9. Air Pollution One of the air pollutants that is measured in selected cities is sulfur dioxide. This pollutant occurs when fossil fuels are burned. This pollutant is measured in micrograms per cubic meter ($\mu g/m^3$). The results obtained from a sample of 24 cities are shown in the frequency distributions. One sample was taken recently, and the other sample of the same cities was taken

5 years ago. Construct a histogram and compare the two distributions.

Class limits	Frequency (now)	Frequency (5 years ago)
10–14	6	5
15–19	4	4
20–24	3	2
25–29	2	3
30–34	5	6
35–39	1	2
40–44	2	1
45–49	1	1
	Total 24	Total 24

10. Making the Grade The frequency distributions shown indicate the percentages of public school students in fourth-grade reading and mathematics who performed at or above the required proficiency levels for the 50 states in the United States. Draw histograms for each, and decide if there is any difference in the performance of the students in the subjects.

Class	Reading frequency	Math frequency
17.5–22.5	7	5
22.5–27.5	6	9
27.5–32.5	14	11
32.5–37.5	19	16
37.5–42.5	3	8
42.5–47.5	1	1
	Total 50	Total 50

Source: National Center for Educational Statistics.

11. Blood Glucose Levels The frequency distribution shows the blood glucose levels (in milligrams per deciliter) for 50 patients at a medical facility. Construct a histogram, frequency polygon, and ogive for the data. Comment on the shape of the distribution. What range of glucose levels did most patients fall into?

Class limits	Frequency
60–64	2
65–69	1
70–74	5
75–79	12
80–84	18
85–89	6
90–94	5
95–99	1
	Total 50

12. Waiting Times The frequency distribution shows the waiting times (in minutes) for 50 patients at a walk-in medical facility. Construct a histogram, frequency polygon, and ogive for the data. Is the distribution skewed? How many patients waited longer than 30 minutes?

Class limits	Frequency
11–15	7
16–20	9
21–25	15
26–30	9
31–35	5
36–40	3
41–45	2
	Total 50

13. Construct a histogram, frequency polygon, and ogive, using relative frequencies for the data in Exercise 1 of this section.

14. Construct a histogram, frequency polygon, and ogive, using relative frequencies for the data in Exercise 2 of this section.

15. Construct a histogram, frequency polygon, and ogive, using relative frequencies for the data in Exercise 5 of this section.

16. Construct a histogram, frequency polygon, and ogive, using relative frequencies for the data in Exercise 6 of this section.

17. Home Runs The data show the most number of home runs hit by a batter in the American League over the last 30 seasons. Construct a frequency distribution using 5 classes. Draw a histogram, a frequency polygon, and an ogive for the date, using relative frequencies. Describe the shape of the histogram.

40	43	40
53	47	46
44	57	43
43	52	44
54	47	51
39	48	36
37	56	42
54	56	49
54	52	40
48	50	40

Source: World Almanac and Book of Facts.

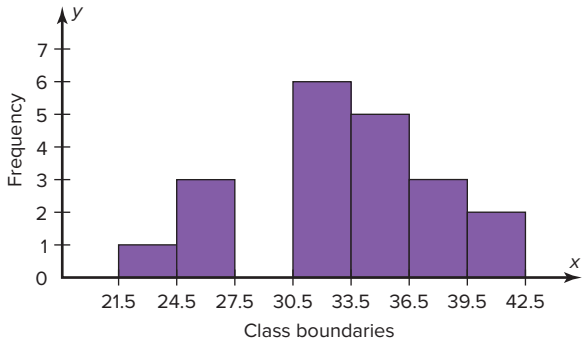
18. Protein Grams in Fast Food The amount of protein (in grams) for a variety of fast-food sandwiches is reported here. Construct a frequency distribution, using 6 classes. Draw a histogram, a frequency polygon, and an ogive for the data, using relative frequencies. Describe the shape of the histogram.

23	30	20	27	44	26	35	20	29	29
25	15	18	27	19	22	12	26	34	15
27	35	26	43	35	14	24	12	23	31
40	35	38	57	22	42	24	21	27	33

Source: The Doctor's Pocket Calorie, Fat, and Carbohydrate Counter.

Extending the Concepts

19. Using the histogram shown here, do the following.



a. Construct a frequency distribution; include class limits, class frequencies, midpoints, and cumulative frequencies.

b. Construct a frequency polygon.

c. Construct an ogive.

20. Using the results from Exercise 19, answer these questions.

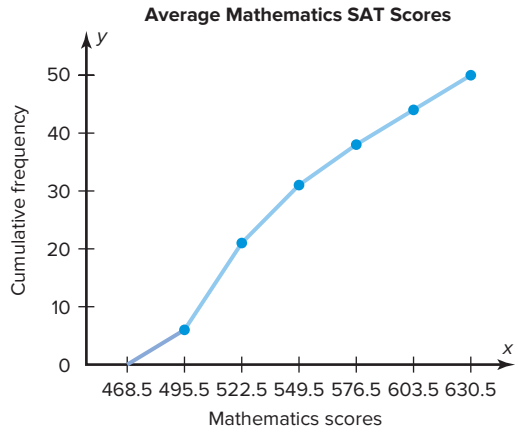
a. How many values are in the class 27.5–30.5?

b. How many values fall between 24.5 and 36.5?

c. How many values are below 33.5?

d. How many values are above 30.5?

21. Math SAT Scores Shown is an ogive depicting the cumulative frequency of the average mathematics SAT scores by state. Use it to construct a histogram and a frequency polygon.



Technology

TI-84 Plus

Step by Step

Step by Step

Constructing a Histogram

To display the graphs on the screen, enter the appropriate values in the calculator, using the WINDOW menu. The default values are $X_{min} = -10$, $X_{max} = 10$, $Y_{min} = -10$, and $Y_{max} = 10$.

The X_{scl} changes the distance between the tick marks on the x axis and can be used to change the class width for the histogram.

To change the values in the WINDOW:

1. Press **WINDOW.**
2. Move the cursor to the value that needs to be changed. Then type in the desired value and press **ENTER.**
3. Continue until all values are appropriate.
4. Press **[2nd] [QUIT]** to leave the WINDOW menu.

Input

```
WINDOW
 Xmin=100
 Xmax=135
 Xscl=5
 Ymin=-5
 Ymax=20
 Yscl=5
 Xres=1
```

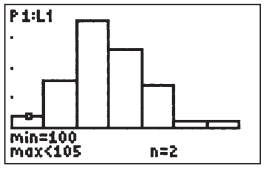
To plot the histogram from raw data:

1. Enter the data in L_1.
2. Make sure WINDOW values are appropriate for the histogram.
3. Press **[2nd] [STAT PLOT] ENTER.**
4. Press **ENTER** to turn the plot 1 on, if necessary.
5. Move cursor to the Histogram symbol and press **ENTER,** if necessary. The histogram is the third option.
6. Make sure Xlist is L_1.
7. Make sure Freq is 1.
8. Press **GRAPH** to display the histogram.
9. To obtain the frequency (number of data values in each class), press the **TRACE** key, followed by ◄ or ► keys.

Input

```
Plot1 Plot2 Plot3
On Off
Type:
Xlist:L1
Freq:1
```

Output



Example TI2–1

Plot a histogram for the following data from Example 2–2.

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

Press **TRACE** and use the arrow keys to determine the number of values in each group.

To graph a histogram from grouped data:

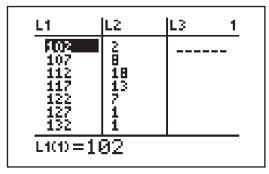
1. Enter the midpoints into L_1.
2. Enter the frequencies into L_2.
3. Make sure WINDOW values are appropriate for the histogram.
4. Press **[2nd] [STAT PLOT] ENTER.**
5. Press **ENTER** to turn the plot on, if necessary.
6. Move cursor to the histogram symbol, and press **ENTER,** if necessary.
7. Make sure Xlist is L_1.
8. Make sure Freq is L_2.
9. Press **GRAPH** to display the histogram.

Example TI2–2

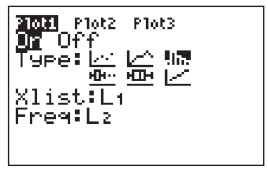
Plot a histogram for the data from Examples 2–4 and 2–5.

Class boundaries	Midpoints	Frequency
99.5–104.5	102	2
104.5–109.5	107	8
109.5–114.5	112	18
114.5–119.5	117	13
119.5–124.5	122	7
124.5–129.5	127	1
129.5–134.5	132	1

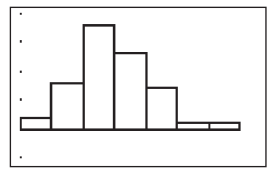
Input



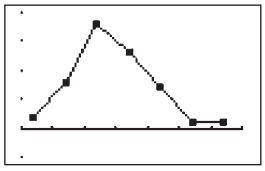
Input



Output

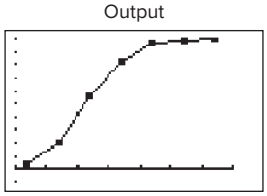


Output



To graph a frequency polygon from grouped data, follow the same steps as for the histogram except change the graph type from histogram (third graph) to a line graph (second graph).

Output



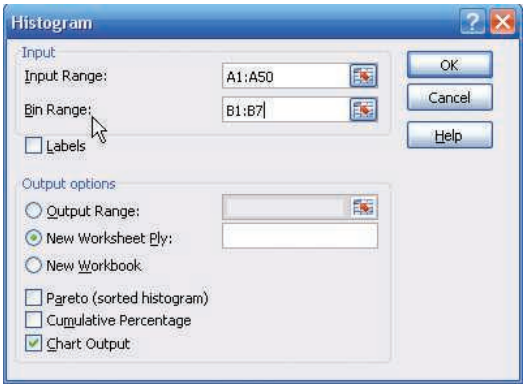
To graph an ogive from grouped data, modify the procedure for the histogram as follows:

1. Enter the upper class boundaries into L_1.
2. Enter the cumulative frequencies into L_2.
3. Change the graph type from histogram (third graph) to line (second graph).
4. Change the Y_{max} from the WINDOW menu to the sample size.

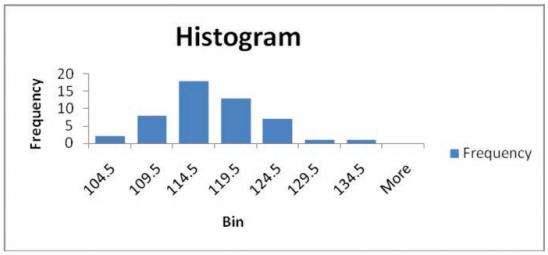
EXCEL
Step by Step

Constructing a Histogram

1. Press **[Ctrl]-N** for a new workbook.
2. Enter the data from Example 2–2 in column A, one number per cell.
3. Enter the upper boundaries into column B.
4. From the toolbar, select the Data tab, then select Data Analysis.
5. In Data Analysis, select Histogram and click [OK].
6. In the Histogram dialog box, type **A1:A50** in the Input Range box and type **B1:B7** in the Bin Range box.



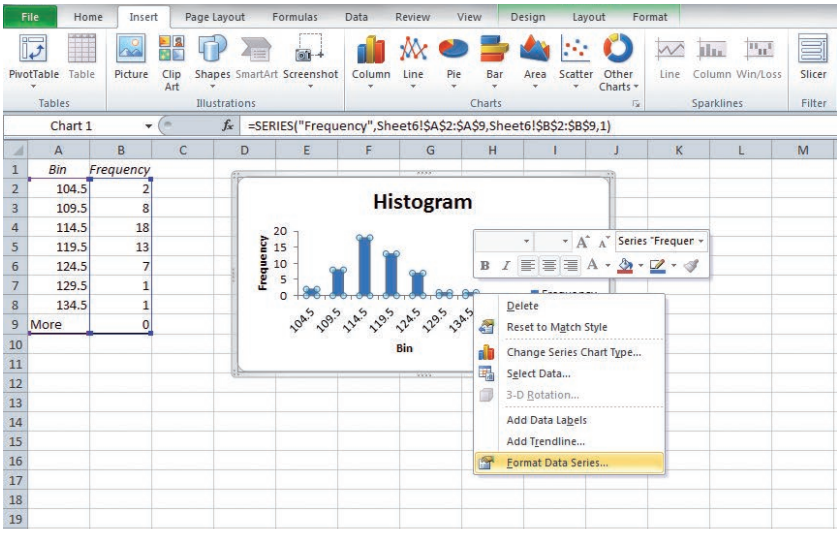
7. Select New Worksheet Ply and Chart Output. Click [OK].



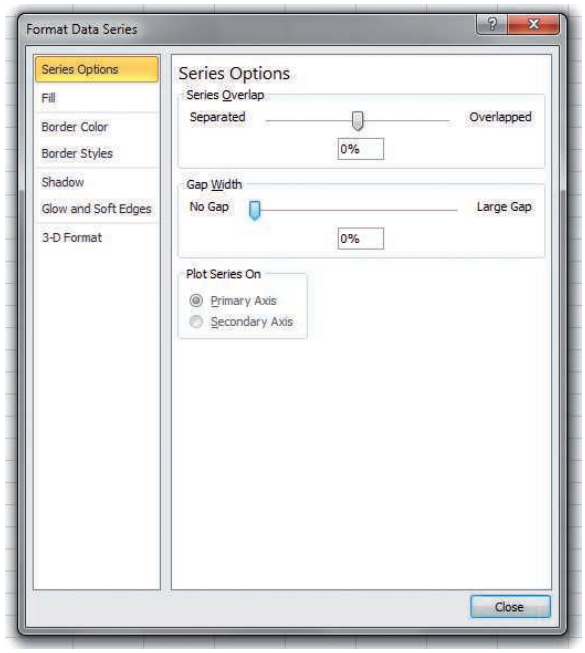
Editing the Histogram

To move the vertical bars of the histogram closer together:

1. Right-click one of the bars of the histogram, and select Format Data Series.

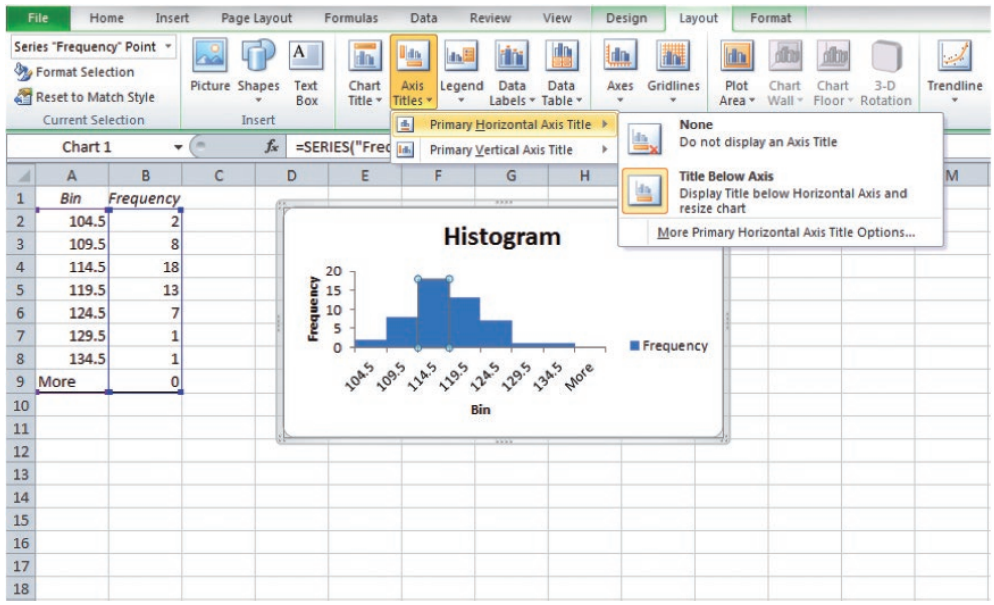


2. Move the Gap Width slider all the way to the left to change the gap width of the bars in the histogram to 0.



To change the label for the horizontal axis:

1. Left-click the mouse over any part of the histogram.
2. Select the Chart Tools tab from the toolbar.
3. Select the Layout tab, Axis Titles and Primary Horizontal Axis Title.



Once the Axis Titles text box is selected, you can type in the name of the variable represented on the horizontal axis.

Constructing a Frequency Polygon

1. Press **[CTRL]-N** for a new notebook.
2. Enter the midpoints of the data from Example 2–2 into column A and the frequencies into column B, including labels.

Note: Classes with frequency 0 have been added at the beginning and the end to "anchor" the frequency polygon to the horizontal axis.

	A	B
1	Midpoints	Frequencies
2	97	0
3	102	2
4	107	8
5	112	18
6	117	13
7	122	7
8	127	1
9	132	1
10	137	0
11		

3. Press and hold the left mouse button, and drag over the Frequencies (including the label) from column B.
4. Select the Insert tab from the toolbar and the Line Chart option.
5. Select the 2-D line chart type.