



sturti/Getty Images

1

Looking at Data— Distributions

Introduction

Statistics is the science of learning from data. Data are numerical or qualitative descriptions of objects that we want to study. In this chapter, we will master the art of examining data.

The methods detailed in this chapter will help answer questions such as

- What are the most important factors that influence the choice of a credit card?
- How can we describe characteristics of cryptocurrencies that have raised more than \$100 million in their Initial Coin Offering (ICO)?
- How can we describe the distribution of IQ scores for a group of sixth-graders?

We begin in Section 1.1 with some basic ideas about data. We will learn about the different types of data and how data sets are organized.

Section 1.2 starts our process of learning from data by looking at graphs. These visual displays give us a picture of the overall patterns in a set of data. We have excellent software tools that help us make these graphs. However, it takes a little experience and a lot of judgment to study the graphs carefully and to explain what they tell us about our data.

1.1 Data

1.2 Displaying Distributions with Graphs

1.3 Describing Distributions with Numbers

1.4 Density Curves and Normal Distributions

Section 1.3 continues our process of learning from data by computing numerical summaries. These sets of numbers describe key characteristics of the patterns that we see in our graphical summaries.

The final section in this chapter helps us make the transition from data summaries to statistical models such as the well-known Normal distributions that are used to draw conclusions and to make predictions.

1.1 Data

When you complete this section, you will be able to:

- Give examples of cases in a data set.
- Identify the variables in a data set and when a variable can be used as a label.
- Identify the values of a variable and classify variables as categorical or quantitative.
- Describe the key characteristics of a set of data.

A statistical analysis starts with a set of data. We construct a set of data by first deciding what *cases* we want to study. For each case, we record information about characteristics that we call *variables*.

Cases, variables, values, and labels

Cases are the objects described by a set of data.

A **variable** is a characteristic of a case.

Different cases can have different **values** of the variables.

A **label** is a special variable used in some data sets to uniquely identify different cases.

The following example illustrates the use of these terms in describing a set of data.

EXAMPLE 1.1



Cryptocurrencies. Cryptocurrencies are decentralized currencies that function without a central bank. The first cryptocurrency, Bitcoin, was created by a person or persons using the name Satoshi Nakamoto and was first released as open-source software in 2009. Since then, many cryptocurrencies have raised money through Initial Coin Offerings (ICOs).¹ **FIGURE 1.1** gives information for the nine ICOs that have raised more than \$100 million; these are the cases.

	A	B	C	D
1	ID	Name	Location	Amount
2	1	Filecoin	North America	257
3	2	Tezos	Europe	236
4	3	EOS	North America	200
5	4	Paragon	North America	183
6	5	The DAO	Stateless/Unknown	168
7	6	Bancor	Middle East	153
8	7	Polkadot	Europe	121
9	8	QASH	Asia	112
10	9	Status	Europe	109

FIGURE 1.1 Cryptocurrency data, Example 1.1.

Data for each case are listed in a different row. The names of the variables (ID, Name, Location, and Amount) appear in the first row. Values of the variables are given as columns. Location gives the region of the world where the ICO originated. Note that the value North America is associated with three ICOs, while Asia is associated with only one. The variable Amount is the amount raised by the ICO, expressed in millions of U.S. dollars. The values range from \$109M to \$257M. ID, a label variable, has the ICOs numbered from 1 to 9. Name is the name of the ICO. Since the names are all different, this variable could also be used as a label variable.

Some variables, such as Location, simply place ICOs into categories. Other variables, like Amount, take on numerical values that we can use to do arithmetic. It makes sense to give an average of the ICO amounts, but it does not make sense to give an “average” location. These ideas lead us to distinguish between two types of variables.

Categorical and quantitative variables

A **categorical variable** places a case into one of several groups or categories.

A **quantitative variable** takes numerical values for which arithmetic operations such as adding and averaging make sense.

EXAMPLE 1.2



Categorical and quantitative variables for cryptocurrencies. The ICO file has four variables: ID, Name, Location, and Amount. ID, Name, and Location are categorical variables. Amount is a quantitative variable. You should choose the label for your cases carefully. In our cryptocurrency example, Location would not be a good choice for a label because, for example, Filecoin, EOS, and Paragon are all located in North America. In contrast, the variable ID has different values, 1, 3, and 4, for these cases and is the label variable that we use for this data set. Although ID has values 1 to 9, it is not a quantitative variable; the average of these values is not a meaningful quantity.

- observation

We sometimes use the term **observation** to describe the data for a particular case. So, in our Cryptocurrency example, we have nine observations. Each observation consists of four pieces of data entered into a row of Figure 1.1.
- spreadsheet


In this and the following two chapters, we focus on relatively simple situations where the structure of the data is straightforward. In other situations, some judgment may be needed to define the characteristics of the data. For example, suppose we study the average daily temperatures in the counties of a particular state for a year. For one study, we could define the cases as the counties in the state and the daily average temperatures as individual variables. For a different study, we could use the combination of the county and the day as the case. The choice will often depend on the particular analysis being performed.

The display in Figure 1.1 is from an Excel **spreadsheet**. Spreadsheets are very useful for doing the kind of simple computations that you will do in Check-in question 1.2. You can type in a formula and have the same computation performed for each row.



Note that the names we have chosen for the variables in our spreadsheet do not include spaces. Someone else might have chosen the name ICO Proceeds for the amount of the ICO rather than Amount. *In some statistical software packages, spaces are not allowed in variable names.* If you are creating spreadsheets for eventual use with statistical software that has this constraint, you need to avoid spaces in variable names. Another convention is to use an underscore (_) where you would normally use a space. For our data set, we could have used ICO_Proceeds instead of Amount.

CHECK-IN

- 1.1 **Read the spreadsheet.** Refer to Figure 1.1. Give the Location and the Amount of the ICO named Polkadot.
- 1.2 **Convert the dollars to euros.** Refer to Example 1.1. Add another column to the ICO spreadsheet that gives the value of the ICO in millions of euros. Assume that 1 dollar equals 0.81 euro. Explain how you computed the entries in this column. Does the new column contain values for a categorical variable or for a quantitative variable? Explain your answer.  ICO

unit of measurement

Another important part of the description of any quantitative variable is its **unit of measurement**. For Amount in Example 1.1, the unit of measurement is millions of dollars. In other settings, the unit of measurement may not be as obvious. For example, if we were measuring heights of children, we might choose to use either inches or centimeters. The units of measurement are an important part of the description of a quantitative variable.

Key characteristics of a data set

In practice, any set of data is accompanied by background information that helps us understand the data. When you plan a statistical study or explore data from someone else’s work, ask yourself the following questions:

- 1. **Who?** What cases do the data describe? **How many** cases does the data set contain?
- 2. **What?** How many variables do the data contain? What are the exact definitions of these variables? What are the units of measurement for each quantitative variable?
- 3. **Why? What purpose** do the data have? Do we hope to answer some specific questions? Do we want to draw conclusions about cases other than the ones we actually have data for? Are the variables that are recorded suitable for the intended purpose?

CHECK-IN

- 1.3 **Who, what, and why questions for the cryptocurrency data.** What cases do the data in Figure 1.1 describe? How many cases are there? How many variables are there? What are their definitions and units of measurement? What purpose do you think the data could have?

EXAMPLE 1.3

Caiaimage/Chris Ryan/Getty Images

Statistics class data: key characteristics. Suppose that you are a teaching assistant for a statistics class and one of your jobs is to keep track of the grades for students in two sections of the course. *Who:* The cases are the 37 students in the class. *What:* There are weekly homework assignments, two exams during the semester, and a final exam. Each of these components is given a numerical score, and the components are added to get a total score that can range from 0 to 1000. Cutoffs of 900, 800, 700, etc., are used to assign letter grades A, B, C, etc.

The spreadsheet for this course will have seven variables:

- An identifier for each student (ID).
- The number of points earned for homework.
- The number of points earned for the first exam.
- The number of points earned for the second exam.
- The number of points earned for the final exam.
- The total number of points earned.
- The letter grade earned.

The student identifier is a label, and the letter grade earned is a categorical variable. The units for all the other variables are points. Because we can do arithmetic with their values, these variables are quantitative variables. There are no units for the label and grade. *Why:* The data are used to compute the letter grades earned by the students in the course.

In our example of statistics class data, the possible values for the grade variable are A, B, C, D, and F. When computing grade point averages, many colleges and universities translate these letter grades into numbers using $A = 4$, $B = 3$, $C = 2$, $D = 1$, $F = 0$. The transformed variable with numeric values is considered to be quantitative because we can average the numerical values across different courses to obtain a grade point average.

Sometimes, experts argue about numerical scales such as this. They ask whether the difference between an A and a B should be considered the same as the difference between a D and an F. Similarly, many questionnaires ask people to respond on a 1 to 5 scale, with 1 representing strongly agree, 2 representing agree, etc. Again we could ask whether the five possible values for this scale are equally spaced in some sense. From a practical point of view, the averages that can be computed when we convert categorical scales such as these to numerical values frequently provide a very useful way to summarize data. Nonetheless, *always be careful when converting categorical data to quantitative data.*

**EXAMPLE 1.4**

Statistics class data for a different purpose. Suppose that the data for the students in the introductory statistics class were also to be used to study relationships between student characteristics and success in the course. Here, we have decided to focus on the total points earned and the grade as the outcomes of interest. Other variables of interest would have been included—for example, gender, whether the student has taken a statistics course previously, and student classification as first, second, third, or fourth year. The label ID is a categorical variable, the total points earned is a quantitative variable, and the remaining variables are all categorical.

CHECK-IN

1.4 Apartment rentals. A data set lists apartments available for students to rent. Information provided includes the monthly rent per person, whether cable is included free of charge, whether or not pets are allowed, the number of bedrooms, the number of bathrooms, and the distance to the campus. Describe the cases in the data set, give the number of variables, and specify whether each variable is categorical or quantitative.

instrument

Often, the variables in a statistical study are easy to understand: height in centimeters, study time in minutes, and so on. But each area of work also has its own special variables. A psychologist uses the Minnesota Multiphasic Personality Inventory (MMPI), and a physical fitness expert measures “VO2 max” (the volume of oxygen consumed per minute while exercising at maximum capacity). Each of these variables is measured with a special **instrument**. VO2 max is measured by exercising while breathing into a mouthpiece connected to an apparatus that measures oxygen consumed. Scores on the MMPI are based on the responses to a long questionnaire, which is also called an instrument.

Part of mastering your field of work is learning what variables are important and how they are best measured. Because details of particular measurements usually require knowledge of the particular field of study, we will say little about them.



rate

Be sure that each variable really does measure what you want it to. A poor choice of variables can lead to misleading conclusions. Often, for example, the **rate** at which something occurs is a more meaningful measure than a simple count of occurrences. Here is an example.

EXAMPLE 1.5

Comparing colleges based on graduates. Think about comparing colleges based on the numbers of graduates. This view tells you something about the relative sizes of different colleges. However, if you are interested in how well colleges succeed at graduating students whom they admit, it would be better to use a rate. For example, you can find data on the Internet on the six-year graduation rates of different colleges. These rates are computed by examining the progress of first-year students who enroll in a given year. Suppose that at College A there were 1000 first-year students in a particular year, and 800 graduated within six years. The graduation rate is

$$\frac{800}{1000} = 0.80$$

or 80%. College B has 2000 students who entered in the same year, and 1200 graduated within six years. The graduation rate is

$$\frac{1200}{2000} = 0.60$$

or 60%. How do we compare these two colleges? College B has more graduates, but College A has a better graduation rate.

adjusting one variable to
create another

In Example 1.5, when we computed the graduation rate, we used the total number of students to adjust the number of graduates. We constructed a new variable by dividing the number of graduates by the total number of first-year students. Computing a rate is just one of several ways of **adjusting one variable to create another**. We often divide one variable by another to compute a more meaningful variable to study. Example 1.16 (page 18) is another type of adjustment.

CHECK-IN

1.5 How should you express the change? Between the first exam and the second exam in your statistics course, you increased the amount of time that you spent working exercises. How would you express the effect of the increased time on your grade? Give reasons for your answer. (Answers and reasons will vary.)

1.6 Which variable would you choose? Refer to Example 1.5 on colleges and their graduates.

(a) Give a setting in which you would prefer to evaluate the colleges based on the numbers of graduates. Give a reason for your choice.

(b) Give a setting in which you would prefer to evaluate the colleges based on the graduation rates. Give a reason for your choice.



Check-in questions 1.5 and 1.6 illustrate an important point about presenting the results of your statistical calculations. *Always consider how to best communicate your results to a general audience.* For example, the numbers produced by your calculator or by statistical software frequently contain more digits than are needed. Be sure that you do not include extra information generated by software that will distract from a clear explanation of what you have found.

Section 1.1 SUMMARY

- A data set contains information on a number of **cases**. Cases may be customers, companies, subjects in a study, units in an experiment, or other objects.
- For each case, the data give values for one or more **variables**. A variable describes some characteristic of a case, such as a person's height, gender, or salary. Variables can have different **values** for different cases.
- A **label** is a special variable used to identify cases in a data set.
- Some variables are **categorical**, and others are **quantitative**. A categorical variable places each individual into a category, such as male or female. A quantitative variable has numerical values that measure some characteristic of each case, such as height in centimeters or annual salary in dollars.
- The **key characteristics** of a data set answer the questions Who?, What?, and Why?
- Converting a count to a **rate** is an example of **adjusting one variable to create another**.

Now that you have completed this section, you will be able to:

- Give examples of cases in a data set. *Review Example 1.1 (page 2) and try Exercise 1.1.*
- Identify the variables in a data set and demonstrate how a label can be used as a variable in a data set. *Review Example 1.1 (page 2) and try Exercise 1.1.*
- Identify the values of a variable and classify variables as categorical or quantitative. *Review Example 1.1 (page 2) and try Exercise 1.1.*
- Describe the key characteristics of a set of data. *Review Example 1.3 (page 5) and try Exercise 1.1.*

Section 1.1 EXERCISES

1.1 Student organizations. A university website gives information about 30 student organizations. You have created a data set that summarizes this information. The variables in your data set are the name of the organization, whether the majority of the members are undergraduates or graduate students, the email

address of the primary advisor, the day of the week when meetings are typically held, and the number of members in a recent year.

- What are the cases?
- Identify the variables and their possible values.

- (c) Classify each variable as categorical or quantitative.
- (d) Was a label used? Explain your answer.
- (e) Summarize the key characteristics of your data set.

1.2 Coffee ratings. A website ranks 50 different varieties of coffee. The data include the following variables: name of the coffee, price for a 12-ounce serving, overall rating (0 to 100), roast (light, medium, or dark), flavor, aroma, and body ratings (0 to 10).

- (a) What are the cases?
- (b) Identify the variables and their possible values.
- (c) Classify each variable as categorical or quantitative.
- (d) Was a label used? Explain your answer.
- (e) Summarize the key characteristics of your data set.

1.3 A survey of graduates. A college surveys its graduates who have earned a bachelor's degree each year. For a recent year, the data include the responses of 1255 employed students who responded to the request for information about their job. The variables collected include an ID numbered 1 to 1255 that identified each respondent, the starting salary after graduation, the industry of employment (selected from a list of 20 possibilities), and the state of the employment if in the United States or the country of employment if not.

- (a) What are the cases?
- (b) Identify the variables and their possible values.
- (c) Classify each variable as categorical or quantitative.
- (d) Was a label used? Explain your answer.
- (e) Summarize the key characteristics of your data set.

1.4 An experiment on haptic feedback. A group of technology students is interested in whether haptic feedback (forces and vibrations applied through a joystick) is helpful in navigating a simulated game environment they created. To investigate this, they randomly assign 20 students to each of three joystick controller types and record the time it takes to complete a navigation mission. The joystick types are (1) a standard video game joystick, (2) a game joystick with force feedback, and (3) a game joystick with vibration feedback. The data collected included an ID variable that uniquely identifies each student, which of the three types of joystick was used, the time taken to complete the navigation mission, the age of the student, and the student's satisfaction with the navigation, rated on a scale of 1 to 5 with 5 being the highest satisfaction.

- (a) What are the cases?
- (b) Identify the variables and their possible values.
- (c) Classify each variable as categorical or quantitative.
- (d) Was a label used? Explain your answer.
- (e) Summarize the key characteristics of your data set.

1.5 Employee application data. The human resources (HR) department keeps records on all employees in a company. Here is the information HR keeps in one of its data files: employee identification number, last name, first name, middle initial, department, number of years with the company, salary, education (coded as high school, some college, or college degree), and age.

- (a) What are the cases for this data set?
- (b) Do you think that the variable last name can be treated as a label? Explain your answer.
- (c) Describe each type of information as a label, a quantitative variable, or a categorical variable.
- (d) Set up a spreadsheet that could be used to record the data. Give appropriate column headings and five sample cases.

1.6 How would you rank cities? Various organizations rank cities and produce lists of the 10 or the 100 best cities, based on various measures. Create a list of criteria that you would use to rank cities. Include at least eight variables and give reasons for your choices. Say whether each variable is quantitative or categorical.

1.7 How would you rate colleges? Popular magazines rank colleges and universities on their "academic quality" in serving undergraduate students. Describe five variables that you would like to see measured for each college if you were choosing where to study. Give reasons for each of your choices.

1.8 Attending college in your state or in another state. The U.S. Census Bureau collects a large amount of information concerning higher education.² For example, the bureau provides a table that includes the following variables: state, number of students from the state who attend college, and number of students who attend college in their home state.

- (a) What are the cases for this set of data?
- (b) Is there a label variable? If yes, what is it?
- (c) Identify each variable as categorical or quantitative.
- (d) Explain how you might use each of the quantitative variables to explain something about the states.
- (e) Consider a variable computed as the number of students in each state who attend college in the state divided by the total number of students from the state who attend college. Explain how you would use this variable to explain something about the states.

1.9 Alcohol-impaired driving fatalities. A report on drunk-driving fatalities in the United States gives the number of alcohol-impaired driving fatalities for each year from 1982 to 2017.³ Discuss at least three different ways that these numbers could be converted to rates. Give the advantages and disadvantages of each.

1.2 Displaying Distributions with Graphs

When you complete this section, you will be able to:

- Use a bar graph and a pie chart to describe the distribution of a categorical variable.
- Use a stemplot and a histogram to describe the distribution of a quantitative variable.
- Examine the distribution of a quantitative variable using the overall pattern of the data and deviations from that pattern and identify the shape, center, and spread of its distribution.
- Choose an appropriate graphical summary for a given set of data.
- Identify and describe any outliers in the distribution of a quantitative variable.
- Use a time plot to describe the trend of a quantitative variable that is measured over time.

distribution For each variable, the cases generally will have different values. The **distribution** of a variable describes how the values of a variable vary from case to case. We can use graphical and numerical descriptions for a distribution. In this section, we start with graphical summaries; we consider numerical summaries in the following section.

exploratory data analysis Statistical tools and ideas help us examine data to describe their main features. This examination is called **exploratory data analysis**. Like an explorer crossing unknown lands, we want first to simply describe what we see. Here are two basic strategies that help us organize our exploration of a set of data:

- Begin by examining each variable by itself. Then move on to study the relationships among the variables.
- Begin with a graph or graphs. Then add numerical summaries of specific aspects of the data.

We will follow these principles in organizing our learning. This chapter presents methods for describing a single variable. We will study relationships among several variables in Chapter 2. Within each chapter, we will begin with graphical displays and add numerical summaries for a more complete description.

predictive analytics When we perform an exploratory data analysis, our focus is on a careful description of the values of the variables in our data set. In many applications, particularly in business and economics, our real interest is in using our descriptions to predict something in the future. We use the term **predictive analytics** to describe data used in this way.

For example, if *Trader Joe's* wanted to use data to decide where to open a new store, the company might analyze data from its current stores with a focus on characteristics of stores that are very successful. If managers can find a new location with characteristics that are similar, they would *predict* that a new store in that location will be successful.

Categorical variables: Bar graphs and pie charts

distribution of a categorical variable
count
percent
proportion

The values of a categorical variable are names for the categories, such as “yes” and “no.” The **distribution of a categorical variable** lists the categories and gives either the **count** or the **percent** of cases that fall in each category. An alternative to the percent is the **proportion**, the count divided by the sum of the counts. Note that the percent is simply the proportion times 100.

EXAMPLE 1.6



sturti/Getty Images

Choosing a credit card. In a study, 1659 U.S. adults were asked about their reasons for choosing a credit card. Here are the most important reasons that they gave for making their choice.⁴

Reason	Count (<i>n</i>)
Cash back	680
Rewards	448
Interest rate	200
Easy to get	149
Brand	133
Other	49
Total	1659

Reason is the categorical variable in this example, and the values are the six different reasons given for the choice.



Note that the last value of the variable Reason is Other, which includes all other most important reasons not reported here. For data sets that have a large number of values for a categorical variable, we often create a category such as this that combines categories with relatively small counts or percents. *Careful judgment is needed when doing this. You don't want to cover up some important piece of information contained in the data by combining data in this way.*

EXAMPLE 1.7



Reasons as percents. When we look at the reasons for choosing a credit card, we see that cash back is the clear winner. As shown in the data set, 680 adults reported cash back as their most important reason for choosing a credit card. To interpret this number, we need to know that the total number of adults surveyed was 1659. When we say that cash back is the winner, we can describe this win by saying that 41% (680 divided by 1659) of the adults reported cash back as their most important reason. Here is a table of the percents for the different reasons:

Reason	Percent (%)
Cash back	41
Rewards	27
Interest rate	12
Easy to get	9
Brand	8
Other	3
Total	100

The use of graphical methods allows us to see this information and other characteristics of the data easily. We now examine two types of graphs, *bar graphs* and *pie charts*.

EXAMPLE 1.8

bar graph



Bar graph for the credit card choices data. FIGURE 1.2 displays the credit card choices data using a **bar graph**. The heights of the six bars show the percents of adults who reported each of the reasons for choosing a credit card as their most important factor.

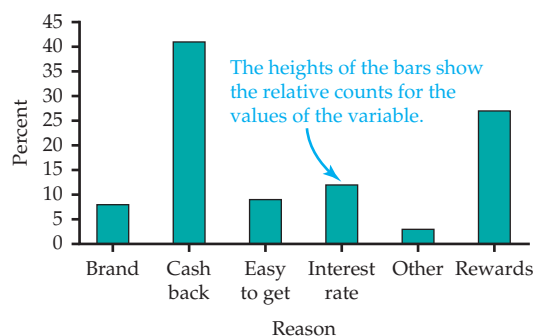


FIGURE 1.2 Bar graph for the credit card choices data, Example 1.8.

The categories in a bar graph (or pie chart) can be put in any order. In Figure 1.2, we ordered the favorite choices alphabetically. You should always consider the best way to order the values of the categorical variable in a bar graph. Choose an ordering that will be useful to you. If you have difficulty deciding, ask a friend if your choice communicates the message that you expect it to convey. You could also use counts in place of percents. A pie chart, as we'll see next, naturally uses percents.

EXAMPLE 1.9

pie chart



Pie chart for the credit card choices data. The **pie chart** in FIGURE 1.3 helps us see which part of the whole each group forms. Here it is very easy to see that cash back is the most important reason for about 41% of the adults.

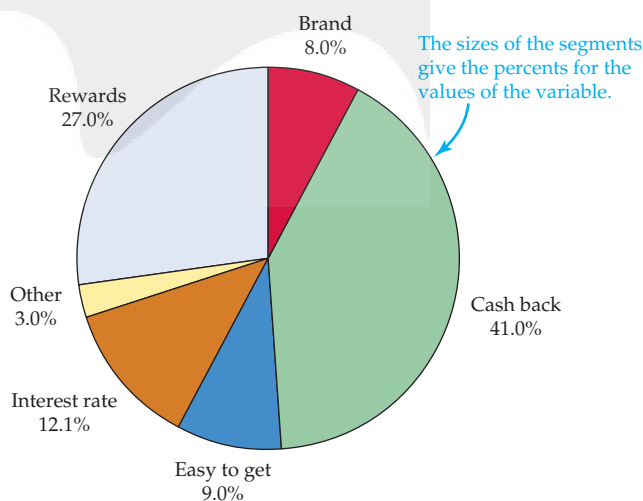


FIGURE 1.3 Pie chart for the credit card choices data, Example 1.9.

CHECK-IN

1.7 Compare the bar graph with the pie chart. Refer to the bar graph in Figure 1.2 and the pie chart in Figure 1.3 for the credit card choices data. Which graphical display does a better job of describing the data? Give reasons for your answer.

Quantitative variables: Stemplots and histograms

A *stemplot* (also called a stem-and-leaf plot) gives a quick picture of the shape of a distribution while including the actual numerical values in the graph. Stemplots work best for small numbers of observations that are all greater than 0.

Stemplot

To make a **stemplot**,

1. Separate each observation into a **stem** consisting of all but the final (rightmost) digit and a **leaf**, the final digit. Stems may have as many digits as needed, but each leaf contains only a single digit.
2. Write the stems in a vertical column with the smallest at the top, and draw a vertical line to the right of this column.
3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.

EXAMPLE 1.10



Soluble corn fiber and calcium. Soluble corn fiber (SCF) has been promoted for various health benefits. One study examined the effect of SCF on the absorption of calcium of adolescent boys and girls. Calcium absorption is expressed as a percent of calcium in the diet. Here are the data for the condition where subjects consumed 12 grams per day (g/d) of SCF.⁵

50	43	43	44	50	44	35	49	54	76	31	48
61	70	62	47	42	45	43	59	53	53	73	

To make a stemplot of these data, use the first digits as stems and the second digits as leaves. **FIGURE 1.4** shows the steps in making the plot. We use the first digit of each value as the stem. Figure 1.4(a) shows the stems that have values 3, 4, 5, 6, and 7. The first entry in our data set is 50. This appears in Figure 1.4(b) on the 5 stem with a leaf of 0. Similarly, the second value, 43, appears in the 4 stem with a leaf of 3. The stemplot is completed in Figure 1.4(c), where the leaves in each row are ordered from smallest to largest.

The middle of the distribution appears to be in the 40s, and the data are more stretched out toward high values than low values. (The highest value is 76, while the lowest is 31.) In the plot, we do not see any extreme values that lie far from the remaining data.

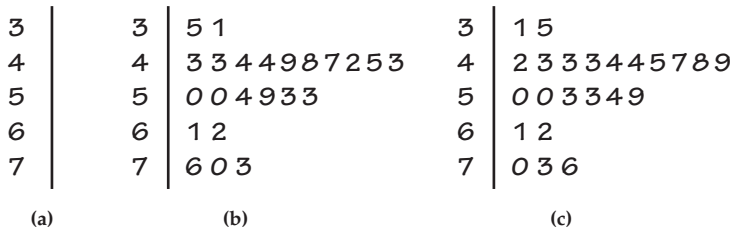



FIGURE 1.4 Making a stemplot of the data in Example 1.10. (a) Write the stems. (b) Go through the data and write each leaf on the proper stem. For example, the values on the 3 stem are 35 and 31 in the order given in the display for the example. (c) Arrange the leaves on each stem in order out from the stem. The 3 stem now has leaves 1 and 5.

CHECK-IN

1.8 Make a stemplot. Here are the scores on the first exam in an introductory statistics course for 28 students in one section of the course:  **STAT**

73	92	82	75	98	94	57	80	90	92	80	87	91	65
70	85	83	61	70	90	75	75	59	68	85	78	80	94

Use these data to make a stemplot. Then use the stemplot to describe the distribution of the first-exam scores for this course.

back-to-back stemplot When you wish to compare two related distributions, a **back-to-back stemplot** with common stems is useful. The leaves on each side are ordered out from the common stem.

EXAMPLE 1.11

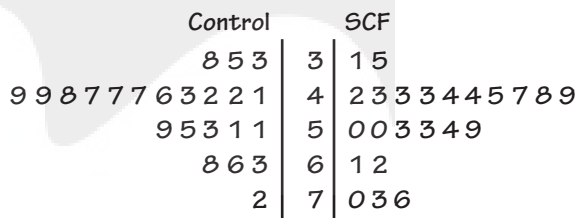


Soluble corn fiber and calcium. Refer to Example 1.10, which gives the data for subjects consuming 12 g/d of SCF. Here are the data for subjects under control conditions (0 g/d of SCF):

42	33	41	49	42	47	48	47	53	72	47	63
68	59	35	46	43	55	38	49	51	51	66	

FIGURE 1.5 gives the back-to-back stemplot for the SCF and control conditions. The values on the left give absorption for the control condition, while the values on the right give absorption when SCF was consumed. The values for SCF appear to be somewhat higher than the controls.

FIGURE 1.5 A back-to-back stemplot to compare the distributions of calcium absorption under control and SCF conditions, Example 1.11.



splitting stems Two modifications of the basic stemplot can be helpful in different situations. You can double the number of stems in a plot by **splitting stems**: separating each stem into two, one with leaves 0 to 4 and the other with leaves 5 through 9. When the observed values have many digits, you can simplify the plot by **trimming** the numbers, removing the last digit or digits before making a stemplot. If you are using software, you can also round the numbers, which is what was done for the data given in Example 1.11.

You must use your judgment in deciding whether to split stems and whether to trim or round, though statistical software will often make these choices for you. Remember that the purpose of a stemplot is to display the shape of a distribution. If there are many stems with no leaves or only one leaf, trimming will reduce the number of stems. Let's take a look at the effect of splitting the stems for our SCF data.

EXAMPLE 1.12



Stemplot with split stems for SCF. FIGURE 1.6 presents the data from Example 1.11 in a stemplot with split stems.

Control	SCF
3	3 1
8 5	3 5
3 2 2 1	4 2 3 3 3 4 4
9 9 8 7 7 7 6	4 5 7 8 9
3 1 1	5 0 0 3 3 4
9 5	5 9
3	6 1 2
8 6	6
2	7 0 3
	7 6

FIGURE 1.6 A back-to-back stemplot with split stems to compare the distributions of calcium absorption under control and SCF conditions, Example 1.12.

CHECK-IN

- 1.9 Which stemplot do you prefer? Look carefully at the stemplots for the SCF data in Figures 1.5 and 1.6. Which do you prefer? Give reasons for your answer.
- 1.10 Why should you keep the space? Suppose that you had a data set similar to the one given in Example 1.11, but in which the control values of 66 and 68 were both changed to 64.
- (a) Make a stemplot of these data using split stems.
- (b) Should you use one stem or two stems for the 60s? Give a reason for your answer. (Hint: How would your choice reveal or conceal a potentially important characteristic of the data?)

Histograms

Stemplots display the actual values of the observations. This feature makes stemplots awkward for large data sets. Moreover, the picture presented by a stemplot divides the observations into groups (stems) determined by our base 10 number system rather than by judgment.

histogram

Histograms do not have these limitations. A **histogram** breaks the range of values of a variable into classes and displays only the count or percent of the observations that fall into each class. You can choose any convenient number of classes, but you should choose classes of equal width.

Making a histogram by hand requires more work than a stemplot. In addition, histograms do not display the actual values observed. For these reasons, we prefer stemplots for small data sets.

The construction of a histogram is best shown by example. Most statistical software packages will make a histogram for you.

EXAMPLE 1.13



Distribution of IQ scores. You have probably heard that the distribution of scores on IQ tests is supposed to be roughly “bell-shaped.” Let’s look at some actual IQ scores. TABLE 1.1 displays the IQ scores of 60 fifth-grade students chosen at random from one school.

TABLE 1.1 IQ test scores for 60 randomly chosen fifth-grade students

145	139	126	122	125	130	96	110	118	118
101	142	134	124	112	109	134	113	81	113
123	94	100	136	109	131	117	110	127	124
106	124	115	133	116	102	127	117	109	137
117	90	103	114	139	101	122	105	97	89
102	108	110	128	114	112	114	102	82	101

To construct a histogram, we proceed as follows:

1. Divide the range of the data into classes of equal width. The classes should be defined so that each score is in exactly one class. Let's use

$$75 \leq \text{IQ score} < 85$$

$$85 \leq \text{IQ score} < 95$$

⋮

$$145 \leq \text{IQ score} < 155$$

Note that a student with IQ 84 would fall into the first class, but IQ 85 is in the second.

2. Count the number of individuals in each class. Each count is called a **frequency**, and a table of frequencies for all classes is a **frequency table**.

frequency
frequency table

Class	Count	Class	Count
$75 \leq \text{IQ score} < 85$	2	$115 \leq \text{IQ score} < 125$	13
$85 \leq \text{IQ score} < 95$	3	$125 \leq \text{IQ score} < 135$	10
$95 \leq \text{IQ score} < 105$	10	$135 \leq \text{IQ score} < 145$	5
$105 \leq \text{IQ score} < 115$	16	$145 \leq \text{IQ score} < 155$	1

3. Draw the histogram. First, on the horizontal axis mark the scale for the variable whose distribution you are displaying—in this case, the IQ score. The scale runs from 75 to 155 because that is the span of the classes we chose. The vertical axis contains the scale of counts. Each bar represents a class. The base of the bar covers the class, and the bar height is the class count. There is no horizontal space between the bars unless a class is empty, so that its bar has height zero. FIGURE 1.7 is our histogram. It does look roughly “bell-shaped.”

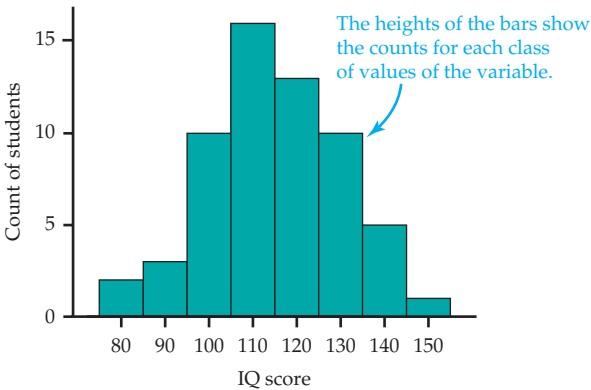



FIGURE 1.7 Histogram of the IQ scores of 60 fifth-grade students, Example 1.13.

Large sets of data are often reported in the form of frequency tables when it is not practical to publish the individual observations. In addition to the frequency (count) for each class, we may be interested in the fraction or percent of the observations that fall in each class. A histogram of percents looks just like a frequency histogram such as Figure 1.7. Simply relabel the vertical scale to read in percents. Use histograms of percents for comparing several distributions that have different numbers of observations.

CHECK-IN


1.11 Make a histogram. Refer to the first-exam scores from Check-in question 1.8 (page 13). Use these data to make a histogram with classes 50 to 59, 60 to 69, etc. Compare the histogram with the stemplot as a way of describing this distribution. Which do you prefer for these data?  **STAT**


Our eyes respond to the *area* of the bars in a histogram. Because the classes are all the same width, area is determined by height, and all classes are fairly represented. There is no one right choice of the classes in a histogram. Too few classes will give a “skyscraper” graph, with all the values in a few classes with tall bars. Too many will produce a “pancake” graph, with most classes having one or no observations. Neither choice will give a good picture of the shape of the distribution. You must use your judgment in choosing classes to display the shape. Statistical software will choose the classes for you. The software’s choice is often a good one, but you can change it if you want.



You should be aware that the appearance of a histogram can change when you change the classes. The histogram function in the *One-Variable Statistical Calculator* applet on the text website allows you to change the number of classes so that it is easy to see how the choice of classes affects the histogram.

CHECK-IN

1.12 Change the classes in the histogram. Refer to the first-exam scores from Check-in question 1.8 (page 13) and the histogram that you produced in Check-in question 1.11. Now make a histogram for these data using classes 40 to 59, 60 to 79, and 80 to 100. Compare this histogram with the one that you produced in Check-in question 1.11. Which do you prefer? Give a reason for your answer.  **STAT**

1.13 Use smaller classes. Repeat the previous Check-in question using classes 55 to 59, 60 to 64, 65 to 69, etc. Of the three histograms, which do you prefer? Give reasons for your answer.  **STAT**

Although histograms resemble bar graphs, their details and uses are distinct. A histogram shows the distribution of counts or percents among the values of a single quantitative variable. The classes define a range of values, and the heights of the bars represent counts of values within the given range. A bar graph, on the other hand, compares the counts or percents of different values for a single categorical variable. The horizontal axis of a bar graph need not have any measurement scale but may simply identify the values of the categorical variable.



Draw bar graphs with blank space between the bars to separate the items being compared. Draw histograms with no space to indicate that all values of the variable are covered. *Some spreadsheet programs, which are not primarily intended for statistics, will draw histograms as if they were bar graphs, with space between the bars.* Often, you can tell the software to eliminate the space to produce a proper histogram.

Examining distributions

Making a statistical graph is not an end in itself. The purpose of the graph is to help us understand the data. After you make a graph, always ask, “What do I see?” Once you have displayed a distribution, you can see its important features as follows.

Examining a distribution

In any graph of data, look for the **overall pattern** and for striking **deviations** from that pattern.

You can describe the overall pattern of a distribution by its **shape**, **center**, and **spread**.

An important kind of deviation is an **outlier**, an individual value that falls outside the overall pattern.

In Section 1.3, we will learn how to describe center and spread numerically. For now, we can describe the center of a distribution by its *midpoint*, the value with roughly half the observations taking smaller values and half taking larger values. We can describe the spread of a distribution by giving the *smallest and largest values*. Stemplots and histograms display the shape of a distribution in the same way. Just imagine a stemplot turned on its side so that the larger values lie to the right.

tail The extreme values of a distribution are in a **tail** of the distribution. The high values are in the upper, or right, tail, and the low values are in the lower, or left, tail. Some things to look for in describing shape are

- Does the distribution have one or several major peaks, each called a **mode**?

mode
unimodal A distribution with one major peak is called **unimodal**. A distribution with two peaks is called **bimodal**, and a distribution with three peaks is called **trimodal**.

- Is it approximately symmetric, or is it skewed in one direction? A distribution is **symmetric** if the patterns of values smaller and larger than its midpoint are mirror images of each other. It is **skewed** to the right if the right tail (larger values) is much longer than the left tail (smaller values).

Some variables commonly have distributions with predictable shapes. Many biological measurements on specimens from the same species and sex—lengths of bird bills, heights of young women—have symmetric distributions. Money amounts, on the other hand, usually have right-skewed distributions. There are many moderately priced houses, for example, but the few very expensive mansions give the distribution of house prices a strong right-skew. For small data sets, it is sometimes difficult to see a clear type of shape.

EXAMPLE 1.14


Examine the histogram of IQ scores. What does the histogram of IQ scores (Figure 1.7, page 15) tell us?

Shape: The distribution is roughly symmetric, with a single peak in the center. We don’t expect real data to be perfectly symmetric, so in judging symmetry, we are satisfied if the two sides of the histogram are roughly similar in shape and extent.

Center: You can see from the histogram that the midpoint is not far from 110. Looking at the actual data shows that the midpoint is 114.

Spread: The histogram has a spread from 75 to 155. Looking at the actual data shows that the spread is from 81 to 145. There are no outliers or other strong deviations from the symmetric, unimodal pattern.

CHECK-IN

1.14 Describe the first-exam scores. Refer to the first-exam scores from Check-in question 1.8 (page 13). Use your favorite graphical display to describe the shape, the center, and the spread of these data. Are there any outliers?  STAT

Dealing with outliers



You can spot outliers by looking for observations that stand apart (either high or low) from the overall pattern of a histogram or stemplot. *Identifying outliers is a matter for judgment. Look for points that are clearly apart from the body of the data, not just the most extreme observations in a distribution. You should search for an explanation for any outlier.* Sometimes outliers point to errors made in recording the data. In other cases, the outlying observation may be caused by equipment failure or other unusual circumstances that would require corrective action. Outliers can be one of the most important characteristics of a data set.

EXAMPLE 1.15



College students. How does the number of undergraduate college students vary by state? **FIGURE 1.8** is a histogram of the numbers of undergraduate students in each of the states.⁶ Notice that 52% of the states are included in the first bar of the histogram. These states have fewer than 250,000 undergraduates. The next bar includes another 34% of the states. These have between 250,000 and 500,000 students. The bar at the far right of the histogram corresponds to the state of California, which has 2,415,337 undergraduates. California certainly stands apart from the other states for this variable. It is an outlier.

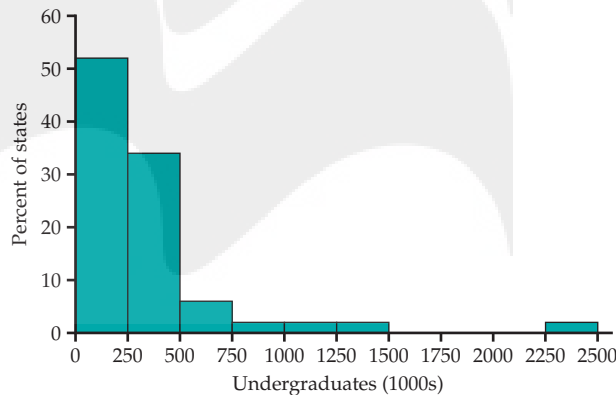


FIGURE 1.8 The distribution of the numbers of undergraduate college students for the 50 states, Example 1.15.

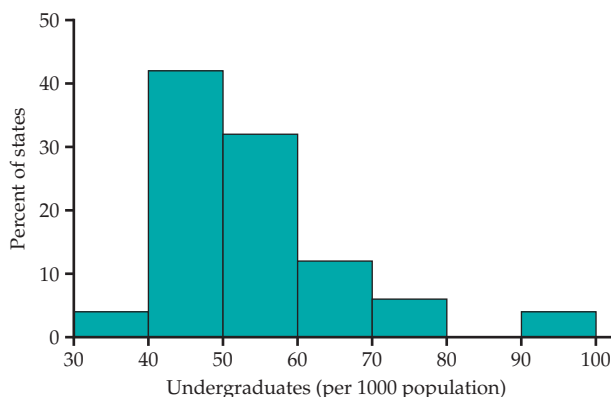
The state of California is an outlier in Example 1.15 because it has a very large number of undergraduate students. California has the largest population of all the states, so we might expect it to have a large number of college students. Let's look at these data in a different way.

EXAMPLE 1.16



College students per 1000. To account for the fact that there is large variation in the populations of the states, for each state we divide the number of undergraduate students by the population and then multiply by 1000. This gives the undergraduate college enrollment expressed as the number of students per 1000 people in each state. **FIGURE 1.9** gives a histogram of the distribution. California has 62 undergraduate students per 1000 people. This is one of the higher values in the distribution, but it is clearly not an outlier.

FIGURE 1.9 The distribution of the numbers of undergraduate college students per 1000 people in each of the 50 states, Example 1.16.



CHECK-IN

1.15 Four states with large populations. There are four states with populations greater than 15 million.  COLLEGE

- (a) Examine the data file and report the names of these four states.
- (b) Find these states in the distribution of number of undergraduate students per 1000 people. To what extent do these four states influence the distribution of number of undergraduate students per 1000 people?

In Example 1.15, we looked at the distribution of the number of undergraduate students, while in Example 1.16, we adjusted these data by expressing the counts as number per 1000 people in each state. Which way is correct? The answer depends upon why you are examining the data.

If you are interested in marketing a product to undergraduate students, the unadjusted numbers would be of interest because you want to reach the most people. On the other hand, if you are interested in comparing states with respect to how well they provide opportunities for higher education to their residents, the population-adjusted values would be more suitable. *Always think about why you are doing a statistical analysis, and this will guide you in choosing an appropriate analytic strategy.*

Here is an example with a different kind of outlier.

EXAMPLE 1.17

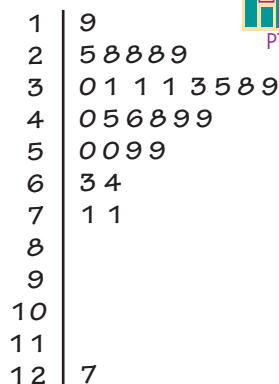


FIGURE 1.10 Stemplot of the values of PTH, Example 1.17.

Healthy bones and PTH. Bones are constantly being built up (bone formation) and torn down (bone resorption). Young people who are growing have more formation than resorption. When we age, resorption increases to the point where it exceeds formation. (The same phenomenon occurs when astronauts travel in space.) The result is osteoporosis, a disease associated with fragile bones that are more likely to break. The underlying mechanisms that control these processes are complex and involve a variety of substances. One of these is parathyroid hormone (PTH). Here are the values of PTH measured on a sample of 29 boys and girls aged 12 to 15 years:⁷

39	59	30	48	71	31	25	31	71	50	38	63	49	45	31
33	28	40	127	49	59	50	64	28	46	35	28	19	29	

The data are measured in picograms per milliliter (pg/ml) of blood. The original data were recorded with one digit after the decimal point. They have been rounded to simplify our presentation here. **FIGURE 1.10** gives a stemplot of the data.

The observation 127 clearly stands out from the rest of the distribution. A PTH measurement on this individual taken on a different day was similar to the rest of the values in the data set. We conclude that this outlier was caused by a laboratory error or a recording error, and we are confident in discarding it for any additional analysis.

Time plots



Whenever data are collected over time, it is a good idea to plot the observations in time order. *Displays of the distribution of a variable that ignore time order, such as stemplots and histograms, can be misleading when there is systematic change over time.*

Time plot

A **time plot** of a variable plots each observation against the time at which it was measured. Always put time on the horizontal scale of your plot and the variable you are measuring on the vertical scale.

EXAMPLE 1.18



Daily high temperatures in West Lafayette, Indiana. In the Northern Hemisphere, temperatures are generally lower during the winter months and higher during the summer months. **FIGURE 1.11** is a plot of the daily high temperatures in West Lafayette, Indiana, for a recent year.⁸ The temperatures are measured in degrees Fahrenheit, and the days are numbered from 1 to 365, corresponding to January 1 and December 31, respectively. Starting with day 1, the pattern increases, then levels off around day 150 to 250, and then decreases for the remaining days.

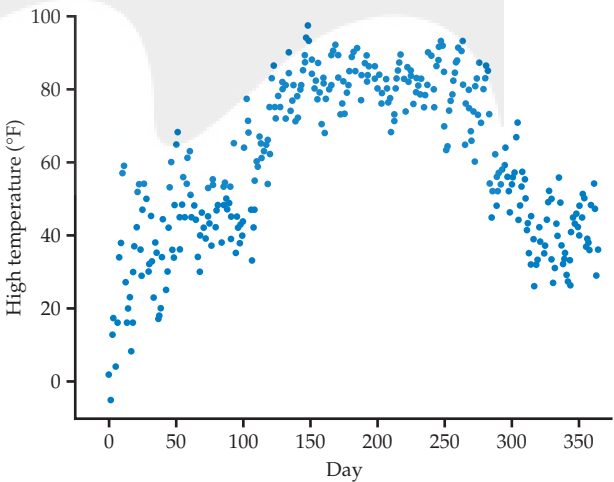


FIGURE 1.11 Plot of high temperature versus day of the year in West Lafayette, Indiana for a recent year, Example 1.18.

The general pattern we see in Figure 1.11 is expected, but the plot gives us additional information about the temperatures. The summer high temperatures are around 80°F, and January is the coldest month, somewhat colder than December. Although there is a clear overall pattern, there is considerable variation around it, with a range of about 20°F.

Plots of variables measured over time can reveal important facts that need to be taken into account in drawing conclusions from data. The changes in

daily high temperature throughout the year are associated with the numbers of hours of daylight. Cells in the skin make vitamin D in response to sunlight. As a result, blood serum levels of vitamin D tend to be lower in winter months, particularly for people who live in northern areas. An analysis of serum vitamin D levels that does not account for the time of year can be very misleading.

Section 1.2 SUMMARY

- **Exploratory data analysis** uses graphs and numerical summaries to describe the variables in a data set and the relations among them.
- The **distribution** of a variable tells us what values it takes and how often it takes these values.
- To describe a distribution, begin with a graph. **Bar graphs** and **pie charts** display the distribution of a categorical variable. **Stemplots** and **histograms** display the distributions of a quantitative variable.
- When examining any graph, look for an overall pattern and for clear **deviations** from that pattern.
- **Shape, center, and spread** describe the overall pattern of a distribution. Some distributions have simple shapes, such as **symmetric** or **skewed**. Not all distributions have a simple overall shape, especially when there are few observations.
- **Outliers** are observations that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them.
- When observations on a variable are taken over time, make a **time plot** that graphs time horizontally and the values of the variable vertically. A time plot can reveal interesting patterns in a set of data.

Now that you have completed this section, you will be able to:

- Use a bar graph and a pie chart to describe the distribution of a categorical variable. Review Examples 1.8 (page 11) and 1.9 (page 11) and try Exercise 1.21.
- Use a stemplot and a histogram to describe the distribution of a quantitative variable. Review Examples 1.10 (page 12) and 1.13 (page 14) and try Exercise 1.23.
- Examine the distribution of a quantitative variable using the overall pattern of the data and deviations from that pattern and identify the shape, center, and spread of its distribution. Review Examples 1.13 (page 14) and 1.14 (page 17) and try Exercise 1.15.
- Choose an appropriate graphical summary for a given set of data. Review Examples 1.8 (page 11), 1.9 (page 11), 1.10 (page 12), and 1.13 (page 14) and try Exercise 1.11.
- Identify and describe any outliers in the distribution of a quantitative variable. Review Example 1.15 (page 18) and try Exercise 1.15.
- Use a time plot to describe the trend of a quantitative variable that is measured over time. Review Example 1.18 (page 20) and try Exercise 1.27.

Section 1.2 EXERCISES

1.10 What's wrong? Explain what is wrong with each of the following:

- A stemplot can be used to display the distribution of a categorical variable.
- A symmetric distribution can be skewed to the right.
- Always discard outliers before doing an analysis of a set of data.

1.11 Which graphical display should you use? For each of the following scenarios, decide which graphical display (pie chart, bar graph, stemplot, or histogram) you would use to describe the distribution of the

variable. Give a reason for your choice and, if there is an alternative choice that would also be reasonable, explain why your choice was better than the alternative.

- The number of minutes you spent sleeping on each of the seven days in the past week.
- The grades on the first exam in a statistics course for the 120 students enrolled in the course.
- The favorite color of each student in the statistics course.
- The number of students in the graduating high school class for each high school in Iowa.

1.12 Frequent users of social media. A recent survey by the Pew Research Center asked social media users about how often they visited various sites. Pew defined a frequent user to be someone who visited a site several times a day. Here are the percents of users who are frequent users for several popular sites:⁹

Social media	Frequent users (%)
Facebook	51
Snapchat	46
Instagram	42
YouTube	32
Twitter	25

Use a bar graph to describe the percents of frequent users of these sites and write a short summary of the data based on your graph.

1.13 Pie chart for frequent users of social media.

Refer to the previous exercise.

- (a) Use a pie chart to describe the percents of frequent users of these sites and write a short summary of the data based on your chart.
- (b) Compare this pie chart with the bar graph that you produced for the previous exercise. Which do you prefer? Give reasons for your answer.

1.14 Facebook users by country. The following table gives the numbers of active Facebook users by country for the top 11 countries based on the number of users in July 2019.

Country	Facebook users (in millions)
India	270
United States	190
Indonesia	130
Brazil	120
Mexico	82
Philippines	68
Vietnam	58
Thailand	46
Egypt	38
Turkey	37
United Kingdom	37

- (a) Use a bar graph to describe the numbers of users in these countries.
- (b) Describe the major features of your graph in a short paragraph.

1.15 Potassium from potatoes. The 2015 Dietary Guidelines for Americans¹¹ notes that the average potassium (K) intake for U.S. adults is about half of the recommended amount. A major source of potassium is potatoes. Nutrients in the diet can have different absorption depending on the source. One study looked at absorption of potassium, measured in milligrams (mg),

from different sources. Participants ate a controlled diet for five days, and the amount of potassium absorbed was measured. Data for a diet that included 40 milliequivalents (mEq) of potassium were collected from 27 adult subjects.¹²

- (a) Make a stemplot of the data.
- (b) Describe the pattern of the distribution.
- (c) Are there any outliers? If yes, describe them and explain why you have declared them to be outliers.
- (d) Describe the shape, center, and spread of the distribution.

1.16 Potassium from a supplement. Refer to the previous exercise. Data were also recorded for 29 subjects who received a potassium salt supplement with 40 mEq of potassium. Answer the questions in the previous exercise for the supplemented subjects.

1.17 Energy consumption. The U.S. Energy Information Administration reports data summaries of various energy statistics. Let's look at the total amount of energy consumed, in quadrillions of British thermal units (Btu), for each month in a recent year. Here are the data:

Month	Energy (quadrillion Btu)	Month	Energy (quadrillion Btu)
January	9.58	July	8.23
February	8.46	August	8.21
March	8.56	September	7.64
April	7.56	October	7.78
May	7.66	November	8.19
June	7.79	December	8.82


- (a) Look at the table and describe how the energy consumption varies from month to month.
- (b) Make a time plot of the data and describe the patterns.
- (c) Suppose you wanted to communicate information about the month-to-month variation in energy consumption. Which would be more effective, the table of the data or the graph? Give reasons for your answer.

1.18 Energy consumption in a different year. Refer to the previous exercise. Here are the data for the previous year:

Month	Energy (quadrillion Btu)	Month	Energy (quadrillion Btu)
January	8.99	July	8.27
February	8.02	August	8.17
March	8.38	September	7.64
April	7.52	October	7.72
May	7.62	November	8.14
June	7.72	December	9.08

(a) Analyze these data using the questions in the previous exercise as a guide.


(b) Compare the patterns across the two years. Describe any similarities and differences.


1.19 Least favorite colors. What is your least favorite color? One survey produced the following summary of responses to that question: brown, 23%; green, 4%; gray, 12%; orange, 30%; other, 1%; purple, 13%; white, 4%; yellow, 13%.¹⁴  LFAVCOL

(a) Make a bar graph of the percents with the colors ordered alphabetically as they are given in this exercise.

(b) Make a second bar graph with the colors ordered by the percents, largest to smallest. This type of bar graph is called a **Pareto chart**.

(c) Write a short paragraph comparing these two ways to display the data graphically. Which do you prefer? Give a reason for your preference.

1.20 Cheap colors. Refer to the previous exercise. The same study also asked people about what colors they associate with the words cheap/inexpensive. Here are the results: brown, 13%; green, 6%; gray, 8%; orange, 26%; other, 3%; purple, 4%; red, 9%; white, 9%; yellow, 22%. Answer the questions from the previous exercise for these data.  CHPCOL


1.21 Garbage. The formal name for garbage is “municipal solid waste.” In the United States, approximately 250 million tons of garbage are generated in a year. Here is a breakdown of the materials that made up American municipal solid waste in a recent year.¹⁵  GARBAGE

Material	Weight (million tons)	Percent of total
Food scraps	39.7	15.1
Glass	11.5	4.4
Metals	24.0	9.1
Paper, paperboard	68.0	25.9
Plastics	34.5	13.1
Rubber, leather	8.5	3.2
Textiles	16.0	6.1
Wood	16.3	6.2
Yard trimmings	34.7	13.3
Other	9.2	3.6
Total	262.4	100.0

(a) Make a bar graph of the percents. The graph gives a clearer picture of the main contributors to garbage if you order the bars from tallest to shortest, as in a Pareto chart (see Exercise 1.19).

(b) Also make a pie chart of the percents.

(c) Compare the two graphs. Which do you prefer? Give reasons for your answer.


1.22 Vehicle colors. Vehicle colors differ among regions of the world. Here are data on the most popular colors for vehicles in North America:¹⁶  VCOLOR

Color	Percent
White	24
Black	19
Silver	16
Gray	15
Red	10
Blue	7
Brown	5
Other	4

(a) Describe these data with a bar graph.

(b) Describe these data with a pie chart.

(c) Which graphical summary do you prefer? Give reasons for your answer.

1.23 Grades and self-concept. TABLE 1.2 presents data on 78 seventh-grade students in a rural midwestern school.¹⁷ The researcher was interested in the relationship between the students’ “self-concept” and their academic performance. The data we give here include each student’s grade point average (GPA), score on a standard IQ test, and sex, taken from school records. Sex is coded as F for female and M for male. The students are identified only by an observation number. The missing observation numbers show that some students dropped out of the study. The final variable is each student’s score on the Piers-Harris Children’s Self-Concept Scale, a psychological test administered by the researcher.  SEVENGR

(a) How many variables does this data set contain? Which are categorical variables, and which are quantitative variables?


(b) Make a histogram of the distribution of GPA.

(c) Make a stemplot of the distribution of GPA.

(d) Do you prefer the histogram or the stemplot? Explain your choice.

(e) Describe the shape, center, and spread of the GPA distribution. Identify any suspected outliers from the overall pattern.

(f) Make a back-to-back stemplot of the rounded GPAs for female and male students. Write a brief comparison of the two distributions.

1.24 Describe the IQ scores. Make a graph of the distribution of IQ scores for the seventh-grade students in Table 1.2. Describe the shape, center, and spread of the distribution, as well as any outliers. IQ scores are usually said to be centered at 100. Is the midpoint for these students close to 100, clearly above, or clearly below?  SEVENGR

1.25 Sketch a skewed distribution. Sketch a histogram for a distribution that is skewed to the left. Suppose that you and your friends emptied your pockets of coins and recorded the year marked on each coin. The distribution of dates would be skewed to the left. Explain why.


1.26 Describe the self-concept scores. Based on a suitable graph, briefly describe the distribution of self-concept scores for the students in Table 1.2. Be sure to identify any suspected outliers.  SEVENGR

TABLE 1.2 Educational data for 78 seventh-grade students

Obs	GPA	IQ	Gender	Self	Obs	GPA	IQ	Gender	Self
001	7.940	111	M	67	043	10.760	123	M	64
002	8.292	107	M	43	044	9.763	124	M	58
003	4.643	100	M	52	045	9.410	126	M	70
004	7.470	107	M	66	046	9.167	116	M	72
005	8.882	114	F	58	047	9.348	127	M	70
006	7.585	115	M	51	048	8.167	119	M	47
007	7.650	111	M	71	050	3.647	97	M	52
008	2.412	97	M	51	051	3.408	86	F	46
009	6.000	100	F	49	052	3.936	102	M	66
010	8.833	112	M	51	053	7.167	110	M	67
011	7.470	104	F	35	054	7.647	120	M	63
012	5.528	89	F	54	055	0.530	103	M	53
013	7.167	104	M	54	056	6.173	115	M	67
014	7.571	102	F	64	057	7.295	93	M	61
015	4.700	91	F	56	058	7.295	72	F	54
016	8.167	114	F	69	059	8.938	111	F	60
017	7.822	114	F	55	060	7.882	103	F	60
018	7.598	103	F	65	061	8.353	123	M	63
019	4.000	106	M	40	062	5.062	79	M	30
020	6.231	105	F	66	063	8.175	119	M	54
021	7.643	113	M	55	064	8.235	110	M	66
022	1.760	109	M	20	065	7.588	110	M	44
024	6.419	108	F	56	068	7.647	107	M	49
026	9.648	113	M	68	069	5.237	74	F	44
027	10.700	130	F	69	071	7.825	105	M	67
028	10.580	128	M	70	072	7.333	112	F	64
029	9.429	128	M	80	074	9.167	105	M	73
030	8.000	118	M	53	076	7.996	110	M	59
031	9.585	113	M	65	077	8.714	107	F	37
032	9.571	120	F	67	078	7.833	103	F	63
033	8.998	132	F	62	079	4.885	77	M	36
034	8.333	111	F	39	080	7.998	98	F	64
035	8.175	124	M	71	083	3.820	90	M	42
036	8.000	127	M	59	084	5.936	96	F	28
037	9.333	128	F	60	085	9.000	112	F	60
038	9.500	136	M	64	086	9.500	112	F	70
039	9.167	106	M	71	087	6.057	114	M	51
040	10.140	118	F	72	088	6.057	93	F	21
041	9.999	119	F	54	089	6.938	106	M	56

1.27 The Boston Marathon. Women were allowed to enter the Boston Marathon in 1972. TABLE 1.3 gives the times (in minutes, rounded to the nearest minute) for the winning women from 1972 to 2019.¹⁸


Make a graph that shows change over time. What overall pattern do you see? Have times stopped improving in recent years? If so, when did improvement end?  **MARATH**

TABLE 1.3 Boston Marathon winning times for women

Year	Time	Year	Time	Year	Time	Year	Time
1972	190	1984	149	1996	147	2008	145
1973	186	1985	154	1997	146	2009	152
1974	167	1986	145	1998	143	2010	146
1975	162	1987	146	1999	143	2011	142
1976	167	1988	145	2000	146	2012	151
1977	168	1989	144	2001	144	2013	146
1978	165	1990	145	2002	141	2014	140
1979	155	1991	144	2003	145	2015	145
1980	154	1992	144	2004	144	2016	149
1981	147	1993	145	2005	145	2017	142
1982	150	1994	142	2006	143	2018	160
1983	143	1995	145	2007	149	2019	144

1.3 Describing Distributions with Numbers

When you complete this section, you will be able to:

- Describe the center of a distribution by using the mean or the median.
- Describe the spread of a distribution by using the interquartile range (*IQR*) or the standard deviation.
- Describe a distribution by using the five-number summary.
- Describe a distribution or compare data sets measured on the same variable by using boxplots.
- Identify outliers by using the $1.5 \times IQR$ rule.
- Choose measures of center and spread for a particular set of data.
- Compute the effects of a linear transformation on the mean, the median, the standard deviation, and the *IQR*.

We can begin our data exploration with graphs, but numerical summaries make our analysis more specific. For categorical variables, numerical summaries are the counts or percents that we use to construct pie charts or bar graphs. In this section, we focus on numerical summaries for quantitative variables. A brief description of the distribution of a quantitative variable should include its *shape* and numbers describing its *center* and *spread*. We describe the shape of a distribution based on inspection of a histogram or a stemplot. Now we will learn specific ways to use numbers to measure the center and spread of a distribution. We can calculate these numerical measures for any quantitative variable. But to interpret measures of center and spread, and to choose among the several measures we will examine, you must think about the shape of the distribution and the meaning of the data. The numbers, like graphs, are aids to understanding, not “the answer” in themselves.

EXAMPLE 1.19

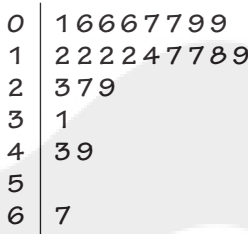


The distribution of business start times. An entrepreneur faces many bureaucratic and legal hurdles when starting a new business. The World Bank collects information about starting businesses throughout the world. It has determined the time, in days, to complete all the procedures required to start a business.¹⁹ Data for 187 countries are included in the data set, TTS. For this example, we examine data, rounded to integers, for a sample of 24 of these countries. Here are the data:

19	17	43	7	12	27	67	49	6	6	29	12
12	9	17	23	1	12	14	18	6	7	9	31

The stemplot in **FIGURE 1.12** shows us the *shape*, *center*, and *spread* of the business start times. The stems are tens of days, and the leaves are days. The distribution is skewed to the right, with a very long tail of high values. All but seven of the times are less than 20 days. The center appears to be about 12 days, and the values range from 1 day to 67 days.

FIGURE 1.12 Stemplot for the sample of 24 business start times, Example 1.19.



Measuring center: The mean

Numerical description of a distribution begins with a measure of its center or average. The two common measures of center are the *mean* and the *median*. The mean is the “average value,” and the median is the “middle value.” These are two different ideas for “center,” and the two measures behave differently. We need precise recipes for the mean and the median.

The mean \bar{x}

To find the **mean** \bar{x} of a set of observations, add their values and divide by the number of observations. If the n observations are x_1, x_2, \dots, x_n , their mean is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

or, in more compact notation,

$$\bar{x} = \frac{1}{n} \sum x_i$$

The \sum (capital Greek sigma) in the formula for the mean is short for “add them all up.” The bar over the x indicates the mean of all the x -values. Pronounce the mean \bar{x} as “x-bar.” This notation is so common that writers who are discussing data use \bar{x} , \bar{y} , etc., without additional explanation. The subscripts on the observations x_i are a way of keeping the n observations separate.

EXAMPLE 1.20



Mean time to start a business. The mean time to start a business is


$$\begin{aligned}\bar{x} &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\ &= \frac{19 + 17 + \cdots + 31}{24} \\ &= \frac{453}{24} = 18.875\end{aligned}$$


The mean time to start a business for the 24 countries in our data set is 19 days. Note that we have rounded the answer. Our goal in using the mean to describe the center of a distribution is not to demonstrate that we can compute with great accuracy. The additional digits do not provide any additional useful information. In fact, they distract our attention from the important digits that are meaningful. Do you think it would be better to report the mean as 18.9 days?

The value of the mean will not necessarily be equal to the value of one of the observations in the data set. Our example of time to start a business illustrates this fact.

In practice, you can key the data into your calculator and hit the Mean key. You don't have to actually add and divide. But you should know that this is what the calculator is doing.

CHECK-IN

1.16 Include the outlier. For Example 1.19, a random sample of 24 countries was selected from a data set that included 187 countries. The South American country Venezuela, where the start time is 230 days, was not included in the random sample. Consider the effect of adding Venezuela to the original set. Show that the mean for the new sample of 25 countries has increased to 27 days. (This is a rounded number. You should report the mean with two digits after the decimal to show that you have performed this calculation.) 

1.17 Find the mean. Here are the scores on the first exam in an introductory statistics course for 10 students: 

75	87	94	85	74	98	93	52	80	91
----	----	----	----	----	----	----	----	----	----

Find the mean first-exam score for these students.



resistant measure

Check-in question 1.16 illustrates an important weakness of the mean as a measure of center: *the mean is sensitive to the influence of a few extreme observations*. These may be outliers, but a skewed distribution that has no outliers will also pull the mean toward its long tail. Because the mean cannot resist the influence of extreme observations, we say that it is not a **resistant measure** of center.

robust measure

A measure that is resistant does more than limit the influence of outliers. Its value does not respond strongly to changes in a few observations, no matter how large those changes may be. The mean fails this requirement because we can make the mean as large as we wish by making a large enough increase in just one observation. A resistant measure is sometimes called a **robust measure**.

Measuring center: The median

We used the midpoint of a distribution as an informal measure of center in Section 1.2. The *median* is the formal version of the midpoint, with a specific rule for calculation.

The median M

The **median** M is the midpoint of a distribution. Half the observations are smaller than the median, and the other half are larger than the median. Here is a rule for finding the median:

1. Arrange all observations in order of size, from smallest to largest.
2. If the number of observations n is odd, the median M is the center observation in the ordered list. Find the location of the median by counting $(n + 1) / 2$ observations up from the bottom of the list.
3. If the number of observations n is even, the median M is the mean of the two center observations in the ordered list. The location of the median is again $(n + 1) / 2$ from the bottom of the list.



Note that the formula $(n + 1) / 2$ does not give the median, just the location of the median in the ordered list. Medians require little arithmetic, so they are easy to find by hand for small sets of data. Arranging even a moderate number of observations in order is tedious, however, so finding the median by hand for larger sets of data is unpleasant. Even simple calculators have an \bar{x} button, but you will need computer software or a graphing calculator to automate finding the median.

EXAMPLE 1.21



Median time to start a business. To find the median time to start a business for our 24 countries, we first arrange the data in order from smallest to largest:

1	6	6	6	7	7	9	9	12	12	12	12
14	17	17	18	19	23	27	29	31	43	49	67

The count of observations $n = 24$ is even. The median, then, is the average of the two center observations in the ordered list. To find the location of the center observations, we first compute



$$\text{location of } M = \frac{n + 1}{2} = \frac{25}{2} = 12.5$$

Therefore, the center observations are the 12th and 13th observations in the ordered list. The median is

$$M = \frac{12 + 14}{2} = 13$$

Note that you can use the stemplot in Figure 1.12 (page 26) directly to compute the median. In the stemplot, the cases are already ordered, and you simply need to count from the top or the bottom to the desired location.

CHECK-IN

- 1.18 Where is the median?** Suppose that the sample size is 25. Find the location of the median.
- 1.19 Include the outlier.** Include Venezuela, where the start time is 230 days, in the data set, and show that the median is 14 days. Write out the ordered list and circle the outlier. Describe the effect of the outlier on the median for this set of data.  TTS25
- 1.20 Find the median.** Here are the scores on the first exam in an introductory statistics course for 10 students:  STAT

75 87 94 85 74 98 93 52 80 91

Find the median first-exam score for these students.

Comparing the mean and the median

Check-in questions 1.16 (page 27) and 1.19 (above) illustrate an important difference between the mean and the median. Venezuela is an outlier. It pulls the mean time to start a business up from 19 days to 27 days. The median increased slightly, from 13 days to 14 days.

The median is more *resistant* than the mean. If the largest start time in the data set were 1200 days, the median for all 25 countries would still be 14 days. The largest observation just counts as one observation above the center, no matter how far above the center it lies. The mean uses the actual value of each observation, and so a single large observation will pull the mean upward.



A good way to compare the responses of the mean and median to extreme observations is to use an interactive applet that allows you to place points on a line and then drag them with your computer's mouse. Exercises 1.53 and 1.54 use the *Mean and Median* applet on the website for this text to compare the mean and the median.

The median and mean are the most common measures of the center of a distribution. For a symmetric distribution, they are close together. If a distribution is exactly symmetric, the mean and median are exactly the same. In a skewed distribution, the mean is farther out in the long tail than is the median.

The endowment for a college or university is money set aside and invested. The income from the endowment is usually used to support various programs. The distribution of the sizes of the endowments of colleges and universities is strongly skewed to the right. Most institutions have modest endowments, but a few are very wealthy. The median endowment of colleges and universities in a recent year was \$142 million—but the mean endowment was \$771 million.²⁰ The few wealthy institutions pull the mean up but do not affect the median. *Don't confuse the "average" value of a variable (the mean) with its "typical" value, which we might describe by the median.*



We can now give a better answer to the question of how to deal with outliers in data. First, look at the data to identify outliers and investigate their causes. You can then correct outliers if they are wrongly recorded, delete them for good reason, or otherwise give them individual attention. The outliers in a data set can be the most important feature of the distribution.

The outlier in Example 1.17 (page 19) can be dropped from the data once we discover that it is an error. If you have no clear reason to drop outliers, you may want to use resistant measures in your analysis so that outliers have little influence over your conclusions. The choice is often a matter for judgment.

Measuring spread: The quartiles

A measure of center alone can be misleading. Two countries with the same median family income are very different if one has extremes of wealth and poverty and the other has little variation among families. A drug manufactured with the correct mean concentration of active ingredient is dangerous if some batches are much too high and others much too low.

We are interested in the *spread* or *variability* of incomes and drug potencies as well as their centers. The simplest useful numerical description of a distribution consists of both a measure of center and a measure of spread.

quartile

The median divides the data in two; half of the observations are above the median, and half are below the median. The upper **quartile** is the median of the upper half of the data. Similarly, the lower quartile is the median of the lower half of the data. With the median, the quartiles divide the data into four equal parts; 25% of the data are in each part.

percentile

We can do a similar calculation for any percent. The p th **percentile** of a distribution is the value that has $p\%$ of the observations fall at or below it. We could call the median the 50th percentile. To calculate a percentile, arrange the observations in increasing order and count up the required percent from the bottom of the list.

Our definition of percentiles is a bit inexact because there is not always a value with exactly $p\%$ of the data at or below it. We will be content to take the nearest observation for most percentiles, but the quartiles are important enough to require an exact rule.

The quartiles Q_1 and Q_3

To calculate the quartiles:

1. Arrange the observations in increasing order and locate the median M in the ordered list of observations.
2. The **first quartile** Q_1 is the median of the observations whose positions in the ordered list are to the left of the location of the overall median.
3. The **third quartile** Q_3 is the median of the observations whose positions in the ordered list are to the right of the location of the overall median.

Here is an example that shows how the rules for quartiles work for even numbers of observations.

EXAMPLE 1.22



Finding the quartiles. Here is the ordered list of the times to start a business in our sample of 24 countries:


1	6	6	6	7	7	9	9	12	12	12	12
14	17	17	18	19	23	27	29	31	43	49	67

The count of observations $n = 24$ is even, so the median is at position $(24 + 1) / 2 = 12.5$, that is, between the 12th and the 13th observation in the ordered list. There are 12 cases above this position and 12 below it. The first quartile is the median of the first 12 observations, and the third quartile is the median of the last 12 observations. Check that $Q_1 = 8$ and $Q_3 = 25$.



Notice that the quartiles are resistant. For example, Q_3 would have the same value if the highest start time were 670 days rather than 67 days.
Be careful when several observations take the same numerical value. Write down all the observations and apply the rules just as if they all had distinct values.

CHECK-IN

1.21 Find the quartiles. Here are the scores on the first exam in an introductory statistics course for 10 students:  STAT

75	87	94	85	74	98	93	52	80	91
----	----	----	----	----	----	----	----	----	----

Find the quartiles for these first-exam scores.



There are several rules for calculating quartiles, which often give slightly different values. The differences are generally small. For describing data, just report the values that your software gives.

The five-number summary and boxplots

In Section 1.2, we used the smallest and largest observations to indicate the spread of a distribution. These single observations tell us little about the distribution as a whole, but they give information about the tails of the distribution that is missing if we know only Q_1 , M , and Q_3 . To get a quick summary of both center and spread, use all five numbers.

The five-number summary

The **five-number summary** of a set of observations consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is

Minimum	Q_1	M	Q_3	Maximum
---------	-------	-----	-------	---------

EXAMPLE 1.23




The five-number summary for the PTH data. Let's find the five-number summary for the PTH scores in Example 1.17. Here is the ordered list of PTH values for our sample of 29 children, arranged specifically for this summary:

19	25	28	28	28	29	30
31	31	31	33	35	38	39
40						
45	46	48	49	49	50	50
59	59	63	64	71	71	127

The sample size is 29, so the median is the located at position $(29 + 1) / 2 = 15$. This corresponds to the value 40. The data display shows the median on a separate line, with the smaller 14 observations above it and the larger 14 observations below it. The quartiles are the medians of the first 14 observations and the last 14 observations. Verify that these values are $Q_1 = (30 + 31) / 2 = 30.5$ and $Q_3 = (50 + 59) / 2 = 54.5$. The minimum and maximum values are 19 and 127, respectively. The five-number summary is 19, 30.5, 40, 54.5, 127.

CHECK-IN

1.22 Find the five-number summary. Here are the scores on the first exam in an introductory statistics course for 10 students:  **STAT**

75 87 94 85 74 98 93 52 80 91

Find the five-number summary for these first-exam scores.

The five-number summary leads to another visual representation of a distribution, the *boxplot*.

Boxplot

A **boxplot** is a graph of the five-number summary:

- A central box spans the quartiles Q_1 and Q_3 .
- A line in the box marks the median M .
- Lines extend from the box out to the smallest and largest observations.

whiskers
box-and-whisker plots

The lines extending to the smallest and largest observations are sometimes called **whiskers**, and boxplots are sometimes called **box-and-whisker plots**. Software provides many varieties of boxplots, some of which use different choices for the placement of the whiskers.

When you look at a boxplot, first locate the median, which marks the center of the distribution. Then look at the spread. The quartiles show the spread of the middle half of the data, and the extremes (the smallest and largest observations) show the spread of the entire data set.

EXAMPLE 1.24



IQ scores. In Example 1.13 (page 14), we used a histogram to examine the distribution of a sample of 60 IQ scores. A boxplot for these data is given in **FIGURE 1.13**. Note that the mean is marked with a + and appears very close to the median. The two quartiles are each approximately the same distance from the median, and the two whiskers are approximately the same distance from the corresponding quartiles. All these characteristics are consistent with a symmetric distribution, as illustrated by the histogram in Figure 1.7.

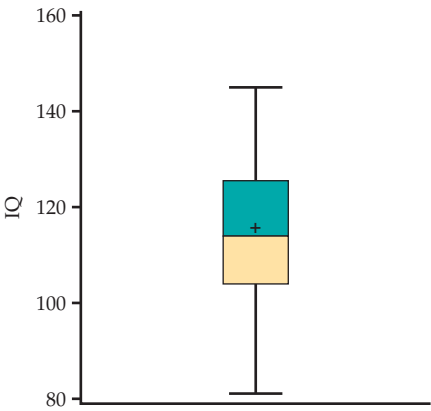



FIGURE 1.13 Boxplot for sample of 60 IQ scores, Example 1.24.

CHECK-IN

1.23 Make a boxplot. Here are the scores on the first exam in an introductory statistics course for 10 students:  STAT

75 87 94 85 74 98 93 52 80 91

Make a boxplot for these first-exam scores.

The $1.5 \times IQR$ rule for suspected outliers

If we look at the PTH data in Example 1.17 (page 19), we can spot a clear outlier; a PTH value of 127, which is almost twice as high as the next highest value. How can we describe the spread of this distribution? The smallest and largest observations are extremes that do not describe the spread of the majority of the data. The distance between the quartiles (the range of the center half of the data) is a more resistant measure of spread than the range. This distance is called the *interquartile range*.

The interquartile range *IQR*

The **interquartile range *IQR*** is the distance between the first and third quartiles,

$$IQR = Q_3 - Q_1$$

EXAMPLE 1.25

IQR for the PTH data. In Example 1.23 (page 31), we found that the five-number summary for the PTH data is 19, 30.5, 40, 54.5, 127. Therefore, we calculate

$$\begin{aligned} IQR &= Q_3 - Q_1 \\ &= 54.5 - 30.5 \\ &= 24 \end{aligned}$$

The quartiles and the *IQR* are not affected by changes in either tail of the distribution. They are resistant, therefore, because changes in a few data points have no further effect once these points move outside the quartiles.



However, *no single numerical measure of spread, such as *IQR*, is very useful for describing skewed distributions.* The two sides of a skewed distribution have different spreads, so one number can't summarize them. We can often detect skewness from the five-number summary by comparing how far the first quartile and the minimum are from the median (left tail) with how far the third quartile and the maximum are from the median (right tail). The interquartile range is mainly used as the basis for a rule of thumb for identifying suspected outliers.

The $1.5 \times IQR$ rule for outliers

Call an observation a suspected outlier if it falls more than $1.5 \times IQR$ below the first quartile or above the third quartile. This is called the **$1.5 \times IQR$ rule**.

EXAMPLE 1.26




Suspected outliers for the PTH data. For the PTH data, we have

$$1.5 \times IQR = 1.5 \times 24 = 36$$

The first quartile is 30.5, and the third quartile is 54.5, so any values below $30.5 - 36 = -5.5$ or above $54.5 + 36 = 90.5$ are flagged as possible outliers. There are no low outliers, but the value 127 is flagged as a possible high outlier.

CHECK-IN

1.24 Use the IQR rule for outliers. Here are the scores on the first exam in an introductory statistics course for 10 students: 

75 87 94 85 74 98 93 52 80 91

Find the interquartile range and use the $1.5 \times IQR$ rule to check for outliers. How low would the lowest score need to be for it to be an outlier according to this rule?

modified boxplot

Two variations on the basic boxplot can be very useful. The first, called a **modified boxplot**, uses the $1.5 \times IQR$ rule. The lines extending from the box to the whiskers are modified. If there are observations identified as outliers by the $1.5 \times IQR$ they are plotted individually and the whiskers terminate at $1.5 \times IQR$ beyond the quartile.

side-by-side boxplots

The other variation is to use two or more boxplots in the same graph to compare groups measured on the same variable. These are called **side-by-side boxplots**. The following example illustrates these two variations.

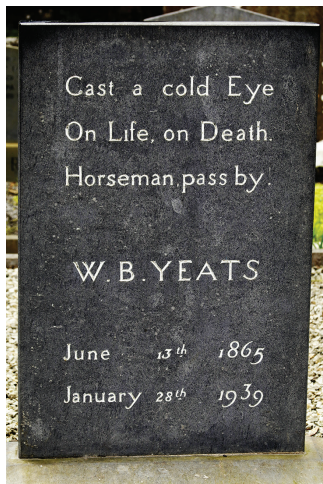
EXAMPLE 1.27



Do poets die young? According to William Butler Yeats, “She is the Gaelic muse, for she gives inspiration to those she persecutes. The Gaelic poets die young, for she is restless, and will not let them remain long on earth.” One study designed to investigate this issue examined the age at death for writers from different cultures and genders.²¹

Three categories of writers examined were novelists, poets, and nonfiction writers. We examine the ages at death for female writers in these categories from North America. **FIGURE 1.14** shows modified side-by-side boxplots for the three categories of writing.

Displaying the boxplots for the three categories of writing lets us compare the three distributions. We see that nonfiction writers tend to live the longest, followed by novelists. The poets do appear to die young! There is one outlier among the nonfiction writers, which is plotted individually along with the value of its label (110). This writer died at the age of 40, young for a non-fiction writer, but not for a novelist or a poet!



David OBrien/Shutterstock

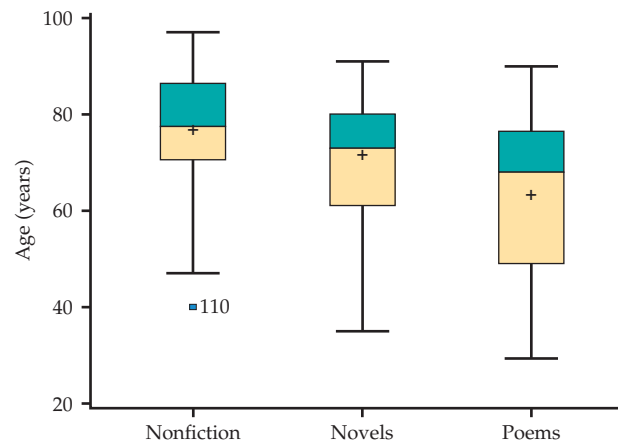


FIGURE 1.14 Modified side-by-side boxplots for the data on writers' age at death, Example 1.27.



Measuring spread: The standard deviation

The five-number summary is not the most common numerical description of a distribution. That distinction belongs to the combination of the mean to measure center and the *standard deviation* to measure spread, or variability. The standard deviation measures spread by looking at how far the observations are from their mean.

The standard deviation s

The **variance** s^2 of a set of observations is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of n observations x_1, x_2, \dots, x_n is

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1}$$

or, in more compact notation,

$$s^2 = \frac{1}{n - 1} \sum (x_i - \bar{x})^2$$

The **standard deviation** s is the square root of the variance s^2 :

$$s = \sqrt{\frac{1}{n - 1} \sum (x_i - \bar{x})^2}$$

The idea behind the variance and the standard deviation as measures of spread is as follows: The deviations $x_i - \bar{x}$ display the spread of the values x_i about their mean \bar{x} . Some of these deviations will be positive and some negative because some of the observations fall on each side of the mean. In fact, the sum of the deviations of the observations from their mean will always be zero. Squaring the deviations makes the negative deviations positive, so that observations far from the mean in either direction have large positive squared deviations. The variance is the average squared deviation. Therefore, s^2 and s will be large if the observations are widely spread about their mean and small if the observations are all close to the mean.

EXAMPLE 1.28



Metabolic rate. A person’s metabolic rate is the rate at which the body consumes energy. Metabolic rate is important in studies of weight gain, dieting, and exercise. Here are the metabolic rates of seven men who took part in a study of dieting. (The units are calories per 24 hours. These are the same calories used to describe the energy content of foods.)

1792	1666	1362	1614	1460	1867	1439
------	------	------	------	------	------	------

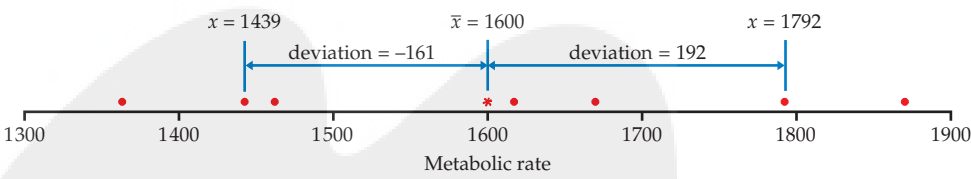
Use software to verify that

$$\bar{x} = 1600 \text{ calories} \quad s = 189.24 \text{ calories}$$

FIGURE 1.15 plots these data as dots on the calorie scale, with their mean marked by an asterisk (*). The arrows mark two of the deviations from the mean. If you were calculating s by hand, you would find the first deviation as

$$x_1 - \bar{x} = 1792 - 1600 = 192$$

FIGURE 1.15 Metabolic rates for seven men, with the mean (*) and the deviations of two observations from the mean, Example 1.28.



Exercise 1.52 (page 45) asks you to calculate the seven deviations from Example 1.28, square them, and find s^2 and s directly from the deviations. Working one or two short examples by hand helps you understand how the standard deviation is obtained. In practice, you will use software to find s .

CHECK-IN

1.25 Find the variance and the standard deviation. Here are the scores on the first exam in an introductory statistics course for 10 students:

75	87	94	85	74	98	93	52	80	91
----	----	----	----	----	----	----	----	----	----

Find the variance and the standard deviation for these first-exam scores.

The idea of the variance is straightforward: it is the average of the squares of the deviations of the observations from their mean. The details we have just presented, however, raise some questions.

Why do we square the deviations?

- First, the sum of the squared deviations of any set of observations from their mean is the smallest that the sum of squared deviations from any number can possibly be. This is not true of the unsquared distances. So squared deviations point to the mean as center in a way that other distances do not.
- Second, the standard deviation turns out to be the natural measure of spread for a particularly important class of symmetric unimodal distributions, the *Normal distributions*. We will meet the Normal distributions in the next section.

Why do we emphasize the standard deviation rather than the variance?

- One reason is that s , not s^2 , is the natural measure of spread for Normal distributions, which are introduced in the next section.
- There is also a more general reason to prefer s to s^2 . Because the variance involves squaring the deviations, it does not have the same unit of measurement as the original observations. The variance of the metabolic rates, for example, is measured in squared calories. Taking the square root gives us a description of the spread of the distribution in the original measurement units.

Why do we average by dividing by $n - 1$ rather than n in calculating the variance?

degrees of freedom

- Because the sum of the deviations is always zero, the last deviation can be found once we know the other $n - 1$. So we are not averaging n unrelated numbers. Only $n - 1$ of the squared deviations can vary freely, and we average by dividing the total by $n - 1$.
- The number $n - 1$ is called the **degrees of freedom** of the variance or standard deviation. Many calculators offer a choice between dividing by n and dividing by $n - 1$, so be sure to use $n - 1$.

Properties of the standard deviation

Here are the basic properties of the standard deviation s as a measure of spread.

Properties of the standard deviation

- s measures spread about the mean and should be used only when the mean is chosen as the measure of center.
- $s = 0$ only when there is *no spread*. This happens only when all observations have the same value. Otherwise, $s > 0$. As the observations become more spread out about their mean, s gets larger.
- s , like the mean \bar{x} , is not resistant. A few outliers can make s very large.


CHECK-IN

1.26 A standard deviation of zero. Construct a data set with four cases that has a variable with $s = 0$.



The use of squared deviations renders s even more sensitive than \bar{x} to a few extreme observations. For example, when we add Venezuela to our sample of 24 countries for the analysis of the time to start a business, we increase the standard deviation from 15.7 to 44.9! Distributions with outliers and strongly skewed distributions have standard deviations that do not give much helpful information about such distributions.

CHECK-IN

1.27 Effect of an outlier on the IQR. Find the IQR for the time to start a business with and without Venezuela. What do you conclude about the sensitivity of this measure of spread to the inclusion of an outlier?  TTS24, TTS25

Choosing measures of center and spread

How do we choose between the five-number summary and \bar{x} and s to describe the center and spread of a distribution? Because the two sides of a strongly skewed distribution have different spreads, no single number such as s describes the spread well. The five-number summary, with its two quartiles and two extremes, does a better job.

Choosing a numerical summary

The five-number summary is usually better than the mean and standard deviation for describing a skewed distribution or a distribution with strong outliers. Use \bar{x} and s for reasonably symmetric distributions that are free of outliers.



Remember that a graph gives the best overall picture of a distribution. *Numerical measures of center and spread report specific facts about a distribution, but they do not describe its shape.* Numerical summaries do not disclose the presence of multiple modes or gaps, for example. Always plot your data.

EXAMPLE 1.29



Results from software. We prefer to examine the numerical summaries and graphical summaries together. FIGURE 1.16 gives a boxplot, a histogram, and numerical summaries for the time to start a business data from Example 1.19 (page 26) using Minitab. Similar displays are given for SPSS in FIGURE 1.17 and for JMP in FIGURE 1.18. Examine and compare the outputs carefully. Notice that they give different numbers of significant digits for some of these numerical summaries. There are also variations in how they make the boxplots and how they define classes for the histograms.

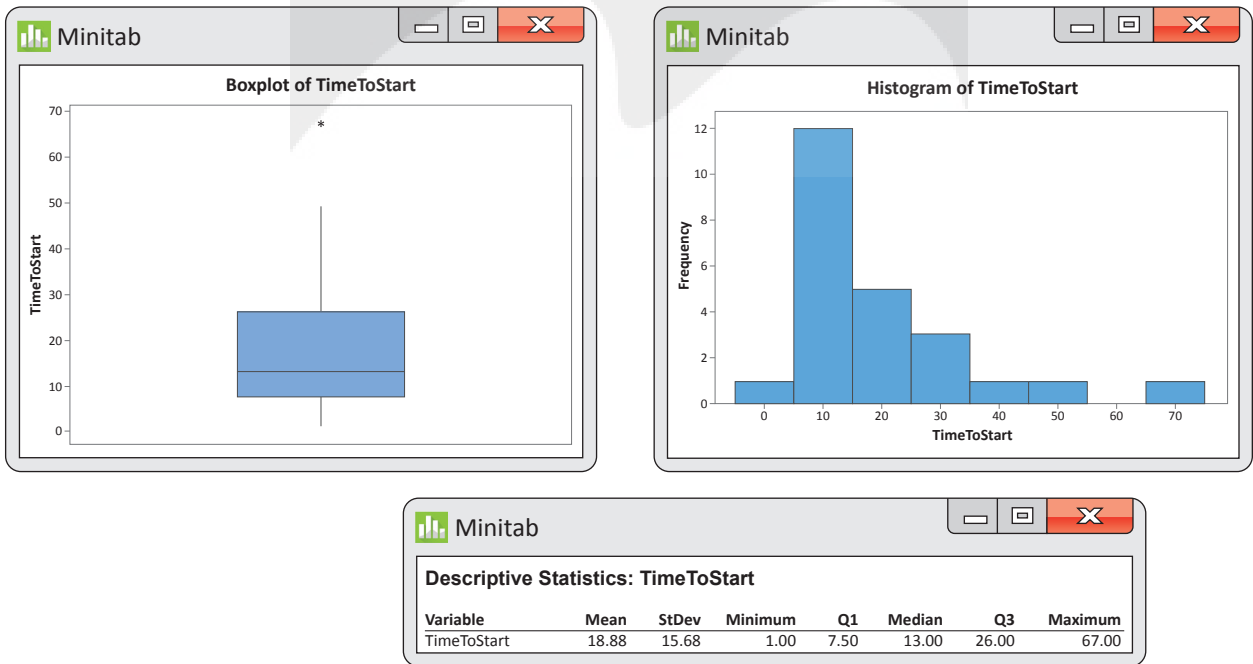


FIGURE 1.16 Graphical and numerical summaries from Minitab: boxplot, histogram, and numerical summaries for the time to start a business, Example 1.29.

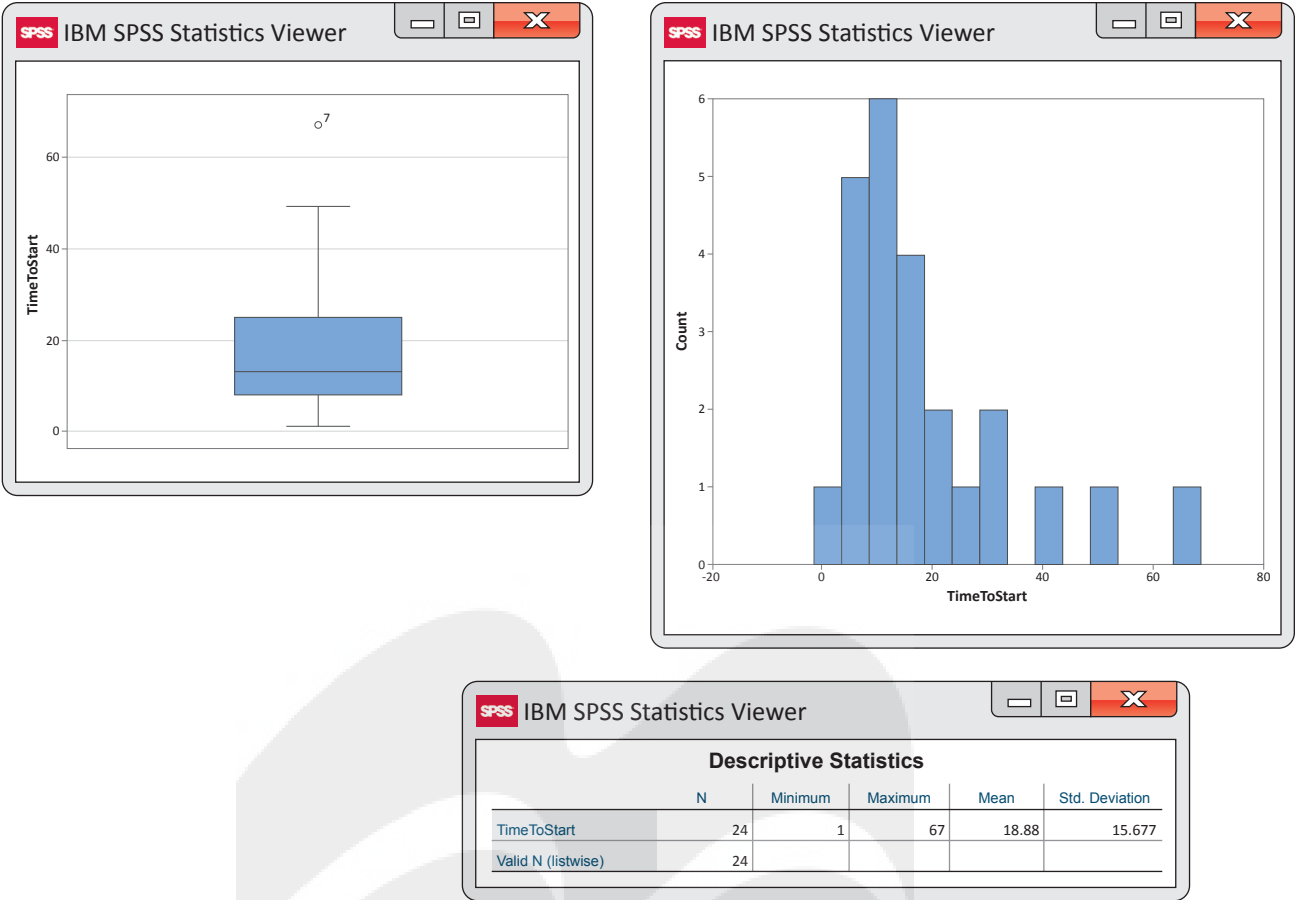


FIGURE 1.17 Graphical and numerical summaries from SPSS: boxplot, histogram, and numerical summaries for the time to start a business, Example 1.29.

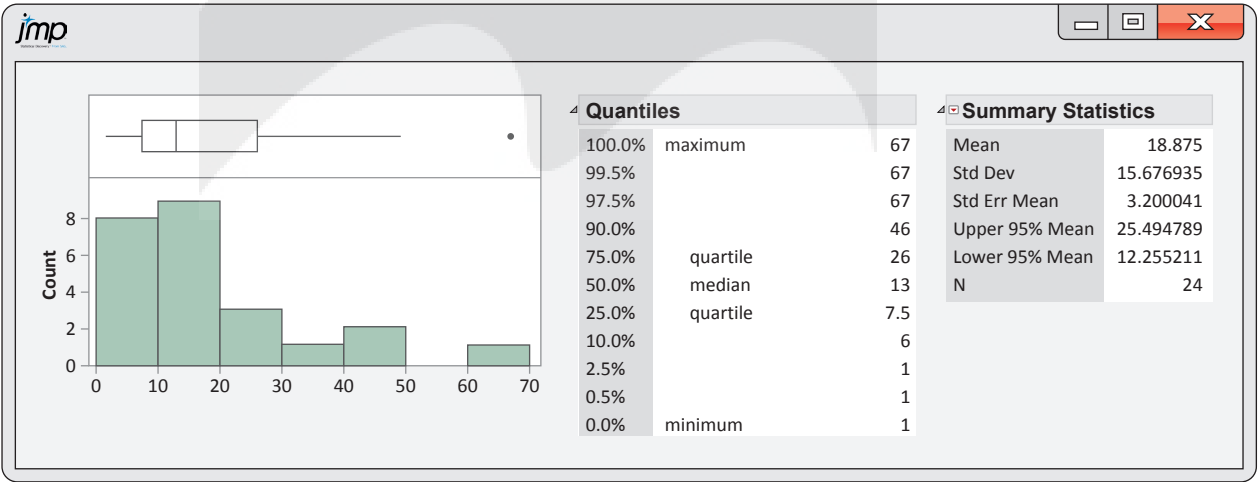


FIGURE 1.18 Graphical and numerical summaries from JMP for the time to start a business, Example 1.29.

Changing the unit of measurement

The same variable can be recorded in different units of measurement. Americans commonly record distances in miles and temperatures in degrees Fahrenheit, while the rest of the world measures distances in kilometers and temperatures in degrees Celsius. Fortunately, it is easy to convert numerical

descriptions of a distribution from one unit of measurement to another. This is true because a change in the measurement unit is a *linear transformation* of the measurements.

Linear transformations

A **linear transformation** changes the original variable x into the new variable x_{new} given by an equation of the form

$$x_{\text{new}} = a + bx$$

Adding the constant a shifts all values of x upward or downward by the same amount. In particular, such a shift changes the origin (zero point) of the variable. Multiplying by the positive constant b changes the size of the unit of measurement.

EXAMPLE 1.30



Marcocphoto/Getty Images

Change the units.

- (a) A temperature x measured in degrees Fahrenheit must be reexpressed in degrees Celsius to be easily understood by the rest of the world. The transformation is

$$x_{\text{new}} = \frac{5}{9}(x - 32) = -\frac{160}{9} + \frac{5}{9}x$$

Thus, the high of 95°F on a hot American summer day translates into 35°C. In this case,

$$a = -\frac{160}{9} \quad \text{and} \quad b = \frac{5}{9}$$

This linear transformation changes both the unit size and the origin of the measurements. The origin in the Celsius scale (0°C, the temperature at which water freezes) is 32° in the Fahrenheit scale.

- (b) If a distance x is measured in kilometers, the same distance in miles is

$$x_{\text{new}} = 0.62x$$

For example, a 10-kilometer race covers 6.2 miles. This transformation changes the units without changing the origin; a distance of 0 kilometers is the same as a distance of 0 miles.

Linear transformations do not change the shape of a distribution. If measurements on a variable x have a right-skewed distribution, any new variable x_{new} obtained by a linear transformation $x_{\text{new}} = a + bx$ (for $b > 0$) will also have a right-skewed distribution. If the distribution of x is symmetric and unimodal, the distribution of x_{new} remains symmetric and unimodal.

Although a linear transformation preserves the basic shape of a distribution, the center and spread will change. Because linear changes of measurement scale are common, we must be aware of their effect on numerical descriptive measures of center and spread. Fortunately, the changes follow a simple pattern.

EXAMPLE 1.31

Use scores to find the points. In an introductory statistics course, homework counts for 300 points out of a total of 1000 possible points for all course requirements. During the semester, there were 12 homework assignments, and each was given a grade on a scale of 0 to 100. The maximum total score for the 12 homework assignments is therefore 1200. To convert the homework scores to final grade points, we need to convert the scale of 0 to 1200 to a scale of 0 to 300. We do this by multiplying the homework scores by $300/1200$. In other words, we divide the homework scores by 4. Here are the homework scores and the corresponding final grade points for five students:

Student	1	2	3	4	5
Score	1056	1080	900	1164	1020
Points	264	270	225	291	255

These two sets of numbers measure the same performance on homework for the course. Because we obtained the points by dividing the scores by 4, the mean of the points will be the mean of the scores divided by 4. Similarly, the standard deviation of points will be the standard deviation of the scores divided by 4.

CHECK-IN

1.28 Calculate the points for a student. Use the setting of Example 1.31 to find the points for a student whose score is 960.

Here is a summary of the rules for linear transformations.

Effect of a linear transformation

To see the effect of a linear transformation on measures of center and spread, apply these rules:

- Multiplying each observation by a positive number b multiplies both measures of center (mean and median) and measures of spread (interquartile range and standard deviation) by b .
- Adding the same number a (either positive or negative) to each observation adds a to measures of center and to quartiles and other percentiles but does not change measures of spread.

In Example 1.31, when we converted from score to points, we described the transformation as dividing by 4. The multiplication part of the summary of the effect of a linear transformation applies to this case because division by 4 is the same as multiplication by 0.25. Similarly, the second part of the summary applies to subtraction as well as addition because subtraction is simply the addition of a negative number.

The measures of spread IQR and s do not change when we add the same number a to all the observations because adding a constant changes the location of the distribution but leaves the spread unaltered. You can find the effect of a linear transformation $x_{\text{new}} = a + bx$ by combining these rules. For example, if x has mean \bar{x} , the transformed variable x_{new} has mean $a + b\bar{x}$.

Section 1.3 SUMMARY

- A numerical summary of a distribution should report its **center** and its **spread** or **variability**.
- The **mean** \bar{x} and the **median** M describe the center of a distribution in different ways. The mean is the arithmetic average of the observations, and the median is their midpoint.
- When you use the median to describe the center of a distribution, describe its spread by giving the **quartiles**. The **first quartile** Q_1 has one-fourth of the observations below it, and the **third quartile** Q_3 has three-fourths of the observations below it.
- The **interquartile range** is the difference between the quartiles. It is the spread of the center half of the data. The **$1.5 \times IQR$ rule** flags observations more than $1.5 \times IQR$ beyond the quartiles as possible outliers.
- The **five-number summary**—consisting of the median, the quartiles, and the smallest and largest individual observations—provides a quick overall description of a distribution. The median describes the center, and the quartiles and extremes show the spread.
- **Boxplots** based on the five-number summary are useful for comparing several distributions. The box spans the quartiles and shows the spread of the central half of the distribution. The median is marked within the box. Lines extend from the box to the extremes and show the full spread of the data. In a **modified boxplot**, points identified by the $1.5 \times IQR$ rule are plotted individually. **Side-by-side boxplots** can be used to display boxplots for more than one group on the same graph.
- The **variance** s^2 and especially its square root, the **standard deviation** s , are common measures of spread about the mean as center. The standard deviation s is zero when there is no spread and gets larger as the spread increases.
- A **resistant measure** of any aspect of a distribution is relatively unaffected by changes in the numerical value of a small proportion of the total number of observations, no matter how large these changes are. The median and quartiles are resistant, but the mean and the standard deviation are not.
- The mean and standard deviation are good descriptions for symmetric distributions without outliers. They are most useful for the Normal distributions introduced in the next section. The five-number summary is a better exploratory description for skewed distributions.
- **Linear transformations** have the form $x_{\text{new}} = a + bx$. A linear transformation changes the origin if $a \neq 0$ and changes the size of the unit of measurement if $b > 0$. Linear transformations do not change the overall shape of a distribution. A linear transformation multiplies a measure of spread by b and changes a percentile or measure of center m into $a + bm$.
- Numerical measures of particular aspects of a distribution, such as center and spread, do not report the entire shape of most distributions. In some cases, particularly distributions with multiple peaks and gaps, these measures may not be very informative.


Now that you have completed this section, you will be able to:

- Describe the center of a distribution by using the mean or the median. *Review Examples 1.20 (page 27) and 1.21 (page 28) and try Exercise 1.29.*
- Describe the spread of a distribution by using the interquartile range (IQR) or the standard deviation. *Review Examples 1.25 (page 33) and 1.28 (page 36) and try Exercise 1.31.*
- Describe a distribution by using the five-number summary. *Review Example 1.23 (page 31) and try Exercise 1.31.*
- Describe a distribution or compare data sets measured on the same variable by using boxplots. *Review Examples 1.24 (page 32) and 1.27 (page 34) and try Exercise 1.35.*
- Identify outliers by using the $1.5 \times IQR$ rule. *Review Example 1.26 (page 34) and try Exercise 1.39.*
- Choose measures of center and spread for a particular set of data. *Review Example 1.29 (page 38) and try Exercises 1.37 and 1.41.*
- Compute the effects of a linear transformation on the mean, the median, the standard deviation, and the IQR . *Review Example 1.30 (page 40) and try Exercise 1.57.*


Section 1.3 EXERCISES

1.28 What's wrong? Explain what is wrong with each of the following:


- (a) The mean is a resistant measure of the center of a distribution.
- (b) If you multiply a variable by 10, you do not change the value of the mean.
- (c) The five number summary includes the mean and the standard deviation.

1.29 Potassium from potatoes. Refer to Exercise 1.15 (page 22), where you examined the potassium absorption of a group of 27 adults who ate a controlled diet that included 40 mEq of potassium from potatoes for five days.  KPOT40


- (a) Compute the mean for these data.
- (b) Compute the median for these data.
- (c) Which measure do you prefer for describing the center of this distribution: the mean or the median? Explain your answer. (You may include a graphical summary as part of your explanation.)

1.30 Potassium from a supplement. Refer to Exercise 1.16 (page 22), where you examined the potassium absorption of a group of 29 adults who ate a controlled diet that included 40 mEq of potassium from a supplement for five days.  KSUP40

- (a) Compute the mean for these data.
- (b) Compute the median for these data.
- (c) Which measure do you prefer for describing the center of this distribution: the mean or the median? Explain your answer. (You may include a graphical summary as part of your explanation.)


1.31 Potassium from potatoes. Refer to Exercise 1.15 (page 22), where you examined the potassium absorption of a group of 27 adults who ate a controlled diet that included 40 mEq of potassium from potatoes for five days.  KPOT40

- (a) Compute the standard deviation for these data.
- (b) Compute the quartiles for these data.
- (c) Give the five-number summary and explain the meaning of each of the five numbers.
- (d) Which numerical summary do you prefer for describing this distribution: the mean, the standard deviation, or the five-number summary? Explain your answer. (You may include a graphical summary as part of your explanation.)


1.32 Potassium from a supplement. Refer to Exercise 1.16 (page 22), where you examined the potassium absorption of a group of 29 adults who ate a controlled diet that included 40 mEq of potassium from a supplement for five days.  KSUP40

- (a) Compute the standard deviation for these data.

- (b) Compute the quartiles for these data.
- (c) Give the five-number summary and explain the meaning of each of the five numbers.
- (d) Which numerical summary do you prefer for describing this distribution: the mean, the standard deviation, or the five-number summary? Explain your answer. (You may include a graphical summary as part of your explanation.)

1.33 Potassium from potatoes. Refer to Exercise 1.15 (page 22), where you examined the potassium absorption of a group of 27 adults who ate a controlled diet that included 40 mEq of potassium from potatoes for five days. In Exercise 1.15, you used a stemplot to examine the distribution of the potassium absorption.  KPOT40


- (a) Make a histogram and use it to describe the distribution of potassium absorption.
- (b) Make a boxplot and use it to describe the distribution of potassium absorption.
- (c) Compare the stemplot, the histogram, and the boxplot as graphical summaries of this distribution. Which do you prefer? Give reasons for your answer.


1.34 Potassium from a supplement. Refer to Exercise 1.16 (page 22), where you examined the potassium absorption of a group of 29 adults who ate a controlled diet that included 40 mEq of potassium from a supplement for five days. In Exercise 1.16, you used a stemplot to examine the distribution of the potassium absorption.  KSUP40

- (a) Make a histogram and use it to describe the distribution of potassium absorption.
- (b) Make a boxplot and use it to describe the distribution of potassium absorption.
- (c) Compare the stemplot, the histogram, and the boxplot as graphical summaries of this distribution. Which do you prefer? Give reasons for your answer.

1.35 Compare the potatoes with the supplement. Refer to Exercises 1.15 and 1.16 (page 22).  KPS40

- (a) Use a back-to-back stemplot to display the data for the two sources of potassium. Compare the two distributions and write a short summary of your findings.
- (b) Use side-by-side boxplots to display the data for the two sources of potassium. Compare the two distributions and write a short summary of your findings.
- (c) Do you prefer stemplots or boxplots to compare these distributions? Give reasons for your answer.

1.36 Potassium sources. The data for potassium absorption in the previous exercise were expressed in milligrams (mg). Convert the data to grams (g) and answer the questions given in the previous exercise. There are 1000 mg in 1 g, so 3000 mg is the same as 3 g. In what ways are your answers here similar to the ones you gave in the previous exercise?  KPS40


1.37 Gosset's data on double stout sales. William Sealy Gosset worked at the Guinness Brewery in Dublin and made substantial contributions to the practice of statistics.²² In his work at the brewery, he collected and analyzed a great deal of data. Archives with Gosset's handwritten tables, graphs, and notes have been preserved at the Guinness Storehouse in Dublin.²³ In one study, Gosset examined the change in the double stout market before and after World War I (1914–1918). For various regions in England and Scotland, he calculated the ratio of sales in 1925, after the war, as a percent of sales in 1913, before the war. Here are the data:  **STOUT**

Bristol	94	Glasgow	66
Cardiff	112	Liverpool	140
English Agents	78	London	428
English O	68	Manchester	190
English P	46	Newcastle-on-Tyne	118
English R	111	Scottish	24

- (a) Compute the mean for these data.
- (b) Compute the median for these data.
- (c) Which measure do you prefer for describing the center of this distribution? Explain your answer. (You may include a graphical summary as part of your explanation.)


1.38 Measures of spread for the double stout data. Refer to the previous exercise.  **STOUT**

- (a) Compute the standard deviation for these data.
- (b) Compute the quartiles for these data.
- (c) Which measure do you prefer for describing the spread of this distribution? Explain your answer. (You may include a graphical summary as part of your explanation.)

1.39 Are there outliers in the double stout data? Refer to the previous two exercises.  **STOUT**

- (a) Find the *IQR* for these data.
- (b) Use the $1.5 \times IQR$ rule to identify and name any outliers.
- (c) Make a boxplot for these data and describe the distribution using only the information in the boxplot.
- (d) Make a modified boxplot for these data and describe the distribution using only the information in the boxplot.
- (e) Make a stemplot for these data.
- (f) Compare the boxplot, the modified boxplot, and the stemplot. Evaluate the advantages and disadvantages of each graphical summary for describing the distribution of the double stout data.

1.40 Smolts. Smolts are young salmon at a stage when their skin becomes covered with silvery scales, and they start to migrate from freshwater to the sea. The reflectance of a light shined on a smolt's skin is

a measure of the smolt's readiness for the migration. Here are the reflectances, in percents, for a sample of 50 smolts:²⁴  **SMOLTS**

57.6	54.8	63.4	57.0	54.7	42.3	63.6	55.5	33.5	63.3
58.3	42.1	56.1	47.8	56.1	55.9	38.8	49.7	42.3	45.6
69.0	50.4	53.0	38.3	60.4	49.3	42.8	44.5	46.4	44.3
58.9	42.1	47.6	47.9	69.2	46.6	68.1	42.8	45.6	47.3
59.6	37.8	53.9	43.2	51.4	64.5	43.8	42.7	50.9	43.8


- (a) Find the mean reflectance for these smolts.
- (b) Find the median reflectance for these smolts.
- (c) Do you prefer the mean or the median as a measure of center for these data? Give reasons for your preference.

1.41 Measures of spread for smolts. Refer to the previous exercise.  **SMOLTS**

- (a) Find the standard deviation of the reflectance for these smolts.
- (b) Find the quartiles of the reflectance for these smolts.
- (c) Do you prefer the standard deviation or the quartiles as a measure of spread for these data? Give reasons for your preference.

1.42 Are there outliers in the smolt data? Refer to the previous two exercises.  **SMOLTS**


- (a) Find the *IQR* for the smolt data.
- (b) Use the $1.5 \times IQR$ rule to identify any outliers.
- (c) Make a boxplot for the smolt data and describe the distribution using only the information in the boxplot.
- (d) Make a modified boxplot for these data and describe the distribution using only the information in the boxplot.
- (e) Make a stemplot for these data.
- (f) Compare the boxplot, the modified boxplot, and the stemplot. Evaluate the advantages and disadvantages of each graphical summary for describing the distribution of the smolt reflectance data.

1.43 Potatoes. A quality product is one that is consistent and has very little variability in its characteristics. Controlling variability can be more difficult with agricultural products than with products that are manufactured. The following table gives the weights, in ounces, of the 25 potatoes sold in a 10-pound bag:  **POTATO**

7.8	7.9	8.2	7.3	6.7	7.9	7.9	7.9	7.6	7.8	7.0	4.7	7.6
6.3	4.7	4.7	4.7	6.3	6.0	5.3	4.3	7.9	5.2	6.0	3.7	

- (a) Summarize the data graphically and numerically. Give reasons for the methods you chose to use in your summaries.

- (b) Do you think that your numerical summaries do an effective job of describing these data? Why or why not?
- (c) There appear to be two distinct clusters of weights for these potatoes. Divide the sample into two subsamples based on the clustering. Give the mean and standard deviation for each subsample. Do you think that this way of summarizing these data is better than a numerical summary that uses all the data as a single sample? Give a reason for your answer.

1.44 The alcohol content of beer. Brewing beer involves a variety of steps that can affect the alcohol content. A website gives the percent alcohol for 160 domestic brands of beer.²⁵ Use graphical and numerical summaries of your choice to describe the data. Give reasons for your choice.  BEER

1.45 Outliers for alcohol content of beer. Refer to the previous exercise.  BEER

- (a) Calculate the mean with and without the outliers. Do the same for the median. Explain how these values change when the outliers are excluded.
- (b) Calculate the standard deviation with and without the outliers. Do the same for the quartiles. Explain how these values change when the outliers are excluded.
- (c) Write a short paragraph summarizing what you have learned in this exercise.

1.46 Calories in beer. Refer to the previous two exercises. The data set also lists calories per 12 ounces of beverage.  BEER

- (a) Analyze the data and summarize the distribution of calories for these 160 brands of beer.
- (b) Are there any outliers? If yes, list them by name. How do these outliers compare with those you identified when analyzing the alcohol content?


1.47 Median versus mean for net worth. A report on the assets of American households says that the median net worth of U.S. families is \$97,300. The mean net worth of these families is \$692,100.²⁶ What explains the difference between these two measures of center?


1.48 Create a data set. Create a data set with five observations for which the median would change by a large amount if the largest observation were deleted.


1.49 Mean versus median. A small accounting firm pays each of its six clerks \$40,000, four junior accountants \$46,000 each, and the firm's owner \$700,000. What is the mean salary paid at this firm? How many of the employees earn less than the mean? What is the median salary?

1.50 Be careful about how you treat the zeros. In computing the median income of any group, some federal agencies omit all members of the group who had no income. Give an example to show that the reported median income of a group can go down even though the group becomes economically better off. Is this also true of the mean income?

1.51 How does the median change? The firm in Exercise 1.49 gives no raises to the clerks and junior accountants, while the owner's take increases to \$950,000. How does this change affect the mean? How does it affect the median?

1.52 Metabolic rates. Calculate the mean and standard deviation of the metabolic rates in Example 1.28 (page 36), showing each step in detail. First find the mean \bar{x} by summing the seven observations and dividing by 7. Then find each of the deviations $x_i - \bar{x}$ and their squares. Check that the deviations have sum 0. Calculate the variance as an average of the squared deviations (remember to divide by $n - 1$). Finally, obtain s as the square root of the variance.  METABOL

 **1.53 Mean and median for two observations.** The *Mean and Median* applet allows you to place observations on a line and see their mean and median visually. Place two observations on the line by clicking below it. Why does only one arrow appear?


 **1.54 Mean and median for six observations.** In the *Mean and Median* applet, place six observations on the line by clicking below it, five close together near the center of the line, and one somewhat to the right of these five.

(a) Pull the single rightmost observation out to the right. (Place the cursor on the point, hold down a mouse button, and drag the point.) How does the mean behave? How does the median behave? Explain briefly why each measure acts as it does.

(b) Now drag the rightmost point to the left as far as you can. What happens to the mean? What happens to the median as you drag this point past the other five?

1.55 Imputation. Various problems with data collection can cause some observations to be missing. Suppose a data set has 20 cases. Here are the values of the variable x for 10 of these cases:


18	9	12	15	20	23	9	12	16	21
----	---	----	----	----	----	---	----	----	----

The values for the other 10 cases are missing. One way to deal with missing data is called **imputation**. The basic idea is that missing values are replaced, or imputed, with values that are based on an analysis of the data that are not missing. For a data set with a single variable, the usual choice of a value for imputation is the mean of the values that are not missing. The mean for this data set is 16.  IMPUTE

(a) Verify that the mean is 16 and find the standard deviation for the 10 cases for which x is not missing.

(b) Create a new data set with 20 cases by setting the values for the 10 missing cases to 16. Compute the mean and standard deviation for this data set.

(c) Summarize what you have learned about the possible effects of this type of imputation on the mean and the standard deviation.

1.56 Longleaf pine trees. The Wade Tract in Thomas County, Georgia, is an old-growth forest of longleaf pine trees (*Pinus palustris*) that has survived in a relatively undisturbed state since before the settlement of the area by Europeans. A study collected data on 584 of these trees.²⁷ One of the variables measured was the diameter at breast height (DBH). This is the diameter of the tree at 4.5 feet, and the units are centimeters (cm). Only trees with DBH greater than 1.5 cm were sampled. Here are the diameters of a random sample of 40 of these trees:  PINES



10.5	13.3	26.0	18.3	52.2	9.2	26.1	17.6	40.5	31.8
47.2	11.4	2.7	69.3	44.4	16.9	35.7	5.4	44.2	2.2
4.3	7.8	38.1	2.2	11.4	51.5	4.9	39.7	32.6	51.8
43.6	2.3	44.6	31.5	40.3	22.3	43.3	37.5	29.1	27.9


- (a) Find the five-number summary for these data.
- (b) Make a boxplot.
- (c) Make a histogram.
- (d) Write a short summary of the major features of this distribution. Do you prefer the boxplot or the histogram for these data?

1.57 Weight gain. A study of diet and weight gain deliberately overfed 15 volunteers for eight weeks. The

mean increase in fat was $\bar{x} = 2.31$ kilograms, and the standard deviation was $s = 1.30$ kilograms. What are \bar{x} and s , in pounds? (A kilogram is 2.2 pounds.)

1.58 Changing units from inches to centimeters. Changing the unit of length from inches to centimeters multiplies each length by 2.54 because there are 2.54 centimeters in an inch. This change of units multiplies our usual measures of spread by 2.54. This is true of *IQR* and the standard deviation. What happens to the variance when we change units in this way?

 **1.59 A different type of mean.** The **trimmed mean** is a measure of center that is more resistant than the mean but uses more of the available information than the median. To compute the 10% trimmed mean, discard the highest 10% and the lowest 10% of the observations and compute the mean of the remaining 80%. Trimming eliminates the effect of a small number of outliers. Compute the 10% trimmed mean of the beer alcohol data in Exercise 1.44 (page 45). Then compute the 20% trimmed mean. Compare the values of these measures with the median and the ordinary untrimmed mean.  BEER

1.60 Changing units from centimeters to inches. Refer to Exercise 1.56. Change the measurements from centimeters to inches by multiplying each value by 0.39. Answer the questions from that exercise and explain the effect of the transformation on these data.  PINES

1.4 Density Curves and Normal Distributions

When you complete this section, you will be able to:

- Sketch a Normal distribution for any given mean and standard deviation.
- Apply the 68–95–99.7 rule to find proportions of observations within one, two, and three standard deviations of the mean for any Normal distribution.
- Find the z-score for any observation x .
- Compute areas under a Normal curve using software or Table A.
- Perform inverse Normal calculations to find values of a Normal variable corresponding to various areas.
- Assess the extent to which the distribution of a set of data can be approximated by a Normal distribution.

We now have a kit of graphical and numerical tools for describing distributions. What is more, we have a clear strategy for exploring data on a single quantitative variable:

1. Always plot your data: make a graph, usually a stemplot or a histogram.
2. Look for the overall pattern and for striking deviations such as outliers.
3. Calculate an appropriate numerical summary to briefly describe center and spread.

Technology has expanded the set of graphs that we can choose for Step 1. It is possible, though painful, to make histograms by hand. Using software, clever algorithms can describe a distribution in a way that is not feasible by hand, by fitting a smooth curve to the data in addition to or instead of a histogram. The curves used are called *density curves*. Before we examine density curves in detail, here is an example of what software can do.

EXAMPLE 1.32

Density curve for times to start a business. FIGURE 1.19 illustrates the use of a density curve along with a histogram to describe distributions. It shows the distribution of the times to start a business for 186 countries (see Example 1.19, page 26). The outlier, Venezuela, described in Check-in question 1.16 (page 27), has been deleted from the data set. The distribution is highly skewed to the right. Most of the data are in the first several classes, with 50 or fewer days to start a business, but there are a few countries with very large start times.

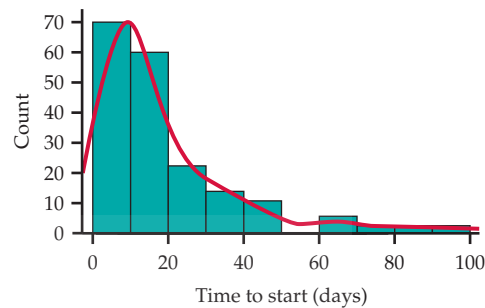


FIGURE 1.19 The distribution of 186 times to start a business, Example 1.32. Venezuela, the outlier, has been eliminated from this plot. The distribution is pictured with both a histogram and a density curve. This distribution has a single mode with a long tail.

A smooth density curve is an idealization that gives the overall pattern of the data but ignores minor irregularities. We first discuss density curves in general and then focus on a special class of density curves, the bell-shaped Normal curves.

Density curves

One way to think of a density curve is as a smooth approximation to the irregular bars of a histogram. FIGURE 1.20 shows a histogram of the scores of all 947 seventh-grade students in Gary, Indiana, on the vocabulary part of the Iowa Test of Basic Skills. Scores of many students on this national test have a very regular distribution. The histogram is symmetric, and both tails fall off quite smoothly from a single center peak. There are no large gaps or obvious outliers. The curve drawn through the tops of the histogram bars in Figure 1.20 is a good description of the overall pattern of the data.

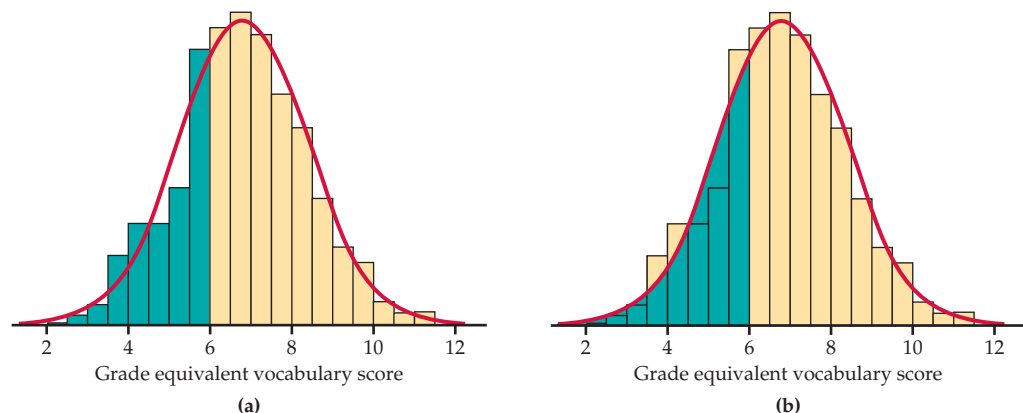


FIGURE 1.20 (a) The distribution of Iowa Test vocabulary scores for Gary, Indiana, seventh-graders, Example 1.33. The shaded bars in the histogram represent scores less than or equal to 6.0. (b) The shaded area under the Normal density curve also represents scores less than or equal to 6.0. This area is 0.293, close to the true 0.303 for the actual data.

EXAMPLE 1.33

Vocabulary scores. In a histogram, the heights of the bars represent either counts or proportions of the observations. In Figure 1.20(a), we shaded the bars that represent students with vocabulary scores 6.0 or lower. There are 287 such students, who make up the proportion $287 / 947 = 0.303$ of all Gary seventh-graders. The shaded bars in Figure 1.20(a) make up proportion 0.303 of the total area under all the bars. If we adjust the scale so that the total area of the bars is 1, the area of the shaded bars will also be 0.303.

In Figure 1.20(b), we shaded the area under the curve to the left of 6.0. If we adjust the scale so that the total area under the curve is exactly 1, areas under the curve will then represent proportions of the observations. That is, *area = proportion*. The curve is then a density curve. The shaded area under the density curve in Figure 1.20(b) represents the proportion of students with score 6.0 or lower. This area is 0.293, only 0.010 away from the histogram result. You can see that areas under the density curve give quite good approximations of areas given by the histogram.

Density curve

A **density curve** is a curve that

- Is always on or above the horizontal axis.
- Has area exactly 1 underneath it.

A density curve describes the overall pattern of a distribution. The area under the curve within any range of values is the proportion of all observations that fall in that range.

The density curve in Figure 1.20 is a *Normal curve*. Density curves, like distributions, come in many shapes. FIGURE 1.21 shows two density curves: a symmetric Normal density curve and a right-skewed curve.

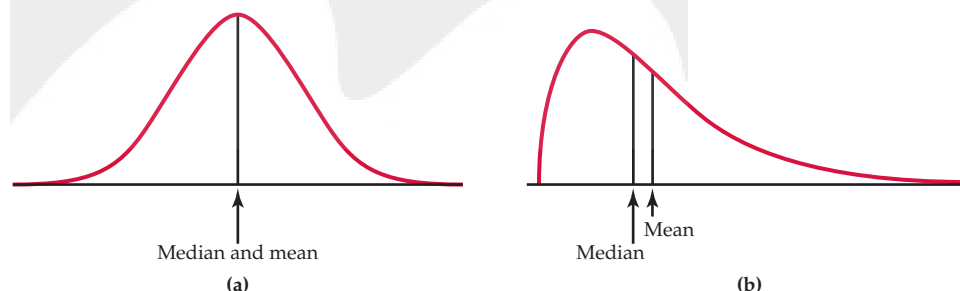


FIGURE 1.21 (a) A symmetric Normal density curve with its mean and median marked. (b) A right-skewed density curve with its mean and median marked.

We will discuss Normal density curves in detail in this section because of the important role they play in statistics. There are, however, many applications where the use of other families of density curves are essential.

A density curve of an appropriate shape is often an adequate description of the overall pattern of a distribution. Outliers, which are deviations from the overall pattern, are not described by the curve.

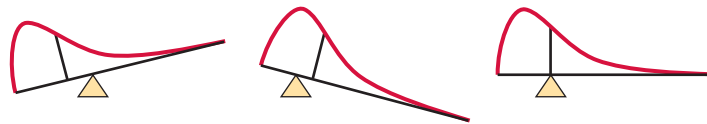
Measuring center and spread for density curves

Our measures of center and spread apply to density curves as well as to actual sets of observations, but only some of these measures are easily seen from the curve. A mode of a distribution described by a density curve is a peak point of the curve, the location where the curve is highest. Because areas under a

density curve represent proportions of the observations, the median is the point with half the total area on each side. You can roughly locate the quartiles by dividing the area under the curve into quarters as accurately as possible by eye. The *IQR* is the distance between the first and third quartiles. There are mathematical ways of calculating areas under curves. These allow us to locate the median and quartiles exactly on any density curve.

What about the mean and standard deviation? The mean of a set of observations is their arithmetic average. If we think of the observations as weights strung out along a thin rod, the mean is the point at which the rod would balance. This fact is also true of density curves. The mean is the point at which the curve would balance if it were made out of solid material. FIGURE 1.22 illustrates this interpretation of the mean.

FIGURE 1.22 The mean of a density curve is the point at which it would balance.



A symmetric curve, such as the Normal curve in Figure 1.21(a), balances at its center of symmetry. Half the area under a symmetric curve lies on either side of its center, so this is also the median.

For a right-skewed curve, such as those shown in Figures 1.21(b) and 1.22, the small area in the long right tail tips the curve more than the same area near the center. The mean (the balance point), therefore, lies to the right of the median. It is hard to locate the balance point by eye on a skewed curve. There are mathematical ways of calculating the mean for any density curve, so we are able to mark the mean as well as the median in Figure 1.21(b). The standard deviation can also be calculated mathematically, but it can't be located by eye on most density curves.

Median and mean of a density curve

The **median of a density curve** is the equal-areas point, the point that divides the area under the curve in half.

The **mean of a density curve** is the balance point, at which the curve would balance if made of solid material.

The median and mean are the same for a symmetric density curve. They both lie at the center of the curve. The mean of a skewed curve is pulled away from the median in the direction of the long tail.

A density curve is an idealized description of a distribution of data. For example, the density curve in Figure 1.20 (page 47) is exactly symmetric, but the histogram of vocabulary scores is only approximately symmetric. We therefore need to distinguish between the mean and standard deviation of the density curve and the numbers \bar{x} and s computed from the actual observations. The usual notation for the **mean** of an idealized distribution is μ (the Greek letter mu). We write the **standard deviation** of a density curve as σ (the Greek letter sigma). In Chapter 5, we refer to \bar{x} and s as statistics associated with a sample and to μ and σ as parameters associated with a population.

Normal distributions

One particularly important class of density curves has already appeared in Figures 1.20 and 1.21(a). These density curves are symmetric, unimodal, and bell-shaped. They are called **Normal curves**, and they describe **Normal distributions**. All Normal distributions have the same overall shape.

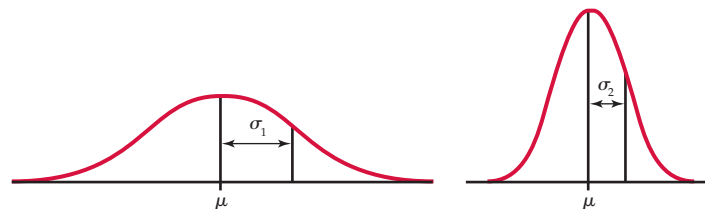
mean μ
standard deviation σ

Normal curves
Normal distributions

The exact density curve for a particular Normal distribution is specified by giving the distribution's mean μ and its standard deviation σ . The mean is located at the center of the symmetric curve and is the same as the median. Changing μ without changing σ moves the Normal curve along the horizontal axis without changing its spread.

The standard deviation σ controls the spread of a Normal curve. FIGURE 1.23 shows two Normal curves with different values of σ . The curve with the larger standard deviation is more spread out.

FIGURE 1.23 Two Normal curves, both showing the same mean μ but with differing standard deviations σ_1 and σ_2 .



The standard deviation σ is the natural measure of spread for Normal distributions. Not only do μ and σ completely determine the shape of a Normal curve, but we can locate σ by eye on the curve. Here's how. As we move out in either direction from the center μ , the curve changes from falling ever more steeply

to falling ever less steeply

The points at which this change of curvature takes place are located at distance σ on either side of the mean μ . You can feel the change as you run your finger along a Normal curve and so find the standard deviation. *Remember that μ and σ alone do not specify the shape of most distributions and that the shape of density curves in general does not reveal σ .* These are special properties of Normal distributions.

There are other symmetric bell-shaped density curves that are not Normal. The Normal density curves are specified by a particular equation. The height of the density curve at any point x is given by

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

We will not make direct use of this fact, although it is the basis of mathematical work with Normal distributions. Notice that the equation of the curve is completely determined by the mean μ and the standard deviation σ .

Why are the Normal distributions important in statistics? Here are three reasons:

1. Normal distributions are good descriptions for some distributions of real data. Distributions that are often close to Normal include scores on tests taken by many people (such as the Iowa Test of Figure 1.20, page 47), repeated careful measurements of the same quantity, and characteristics of biological populations (such as lengths of baby pythons and yields of corn).
2. Normal distributions are good approximations to the results of many kinds of chance outcomes, such as tossing a coin many times.
3. Many statistical inference procedures based on Normal distributions work well for other roughly symmetric distributions.



However, even though many sets of data follow a Normal distribution, many do not. Most income distributions, for example, are skewed to the right and so are not Normal. Non-Normal data, like nonnormal people, not only are common but are also sometimes more interesting than their Normal counterparts.

The 68–95–99.7 rule

Although there are many Normal curves, they all have common properties. Here is one of the most important.

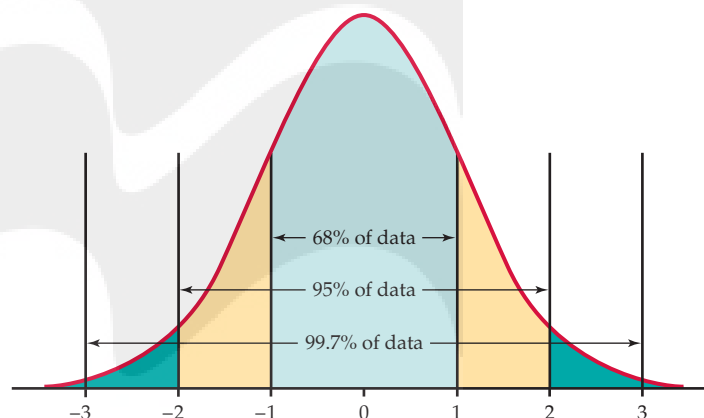
The 68–95–99.7 rule

In the Normal distribution with mean μ and standard deviation σ :

- Approximately **68%** of the observations fall within σ of the mean μ .
- Approximately **95%** of the observations fall within 2σ of μ .
- Approximately **99.7%** of the observations fall within 3σ of μ .

FIGURE 1.24 illustrates the 68–95–99.7 rule. By remembering these three numbers, you can think about Normal distributions without constantly making detailed calculations.

FIGURE 1.24 The 68–95–99.7 rule for Normal distributions.



EXAMPLE 1.34

Heights of young women. The distribution of heights of young women aged 18 to 24 is approximately Normal with mean $\mu = 64.5$ inches and standard deviation $\sigma = 2.5$ inches. FIGURE 1.25 shows what the 68–95–99.7 rule says about this distribution.

Two standard deviations equals five inches for this distribution. The 95 part of the 68–95–99.7 rule says that the middle 95% of young women are between $64.5 - 5$ and $64.5 + 5$ inches tall—that is, between 59.5 and 69.5 inches. This fact is exactly true for an exactly Normal distribution. It is approximately true for the heights of young women because the distribution of heights is approximately Normal.

The other 5% of young women have heights outside the range from 59.5 to 69.5 inches. Because the Normal distributions are symmetric, half of these women are on the tall side. So the tallest 2.5% of young women are taller than 69.5 inches.



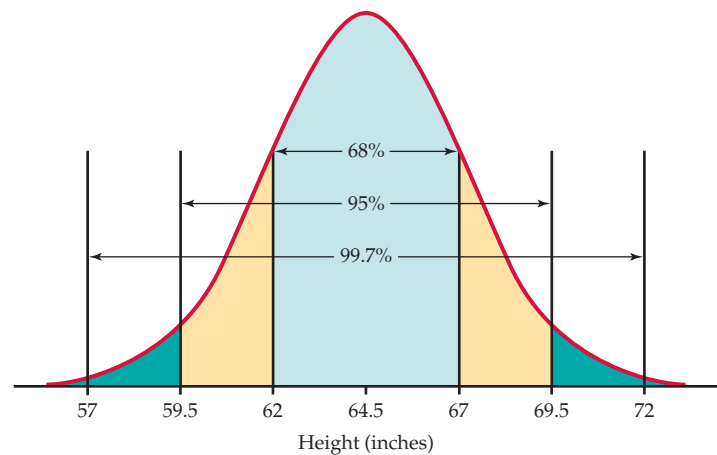


FIGURE 1.25 The 68–95–99.7 rule applied to the heights of young women, Example 1.34.

Because we will mention Normal distributions often, a short notation is helpful. We abbreviate the Normal distribution with mean μ and standard deviation σ as $N(\mu, \sigma)$. For example, the distribution of young women’s heights is $N(64.5, 2.5)$.

CHECK-IN

1.29 Test scores. Many states assess the skills of their students in various grades. One program that is available for this purpose is the National Assessment of Educational Progress (NAEP).²⁸ One of the tests provided by the NAEP assesses the mathematics skills of eighth-grade students. In a recent year, the national mean score was 282, and the standard deviation was 40. Assuming that these scores are approximately Normally distributed, $N(282, 40)$, use the 68–95–99.7 rule to give a range of scores that includes 95% of these students.

1.30 Use the 68–95–99.7 rule. Refer to the previous Check-in question. Use the 68–95–99.7 rule to give a range of scores that includes 99.7% of these students.

Standardizing observations

As the 68–95–99.7 rule suggests, all Normal distributions share many properties. In fact, all Normal distributions are the same if we measure in units of size σ about the mean μ as center. Changing to these units is called *standardizing*. To standardize a value, subtract the mean of the distribution and then divide by the standard deviation.

Standardizing and z-scores

If x is an observation from a distribution that has mean μ and standard deviation σ , the **standardized value** of x is

$$z = \frac{x - \mu}{\sigma}$$

A standardized value is often called a **z-score**.

A z -score tells us how many standard deviations the original observation falls away from the mean, and in which direction. Observations larger than the mean are positive when standardized, and observations smaller than the mean are negative.

To compare scores based on different measures, z -scores can be very useful. For example, see Exercise 1.85 (page 65), where you are asked to compare an SAT score with an ACT score.

EXAMPLE 1.35

Find some z -scores. The heights of young women are approximately Normal with $\mu = 64.5$ inches and $\sigma = 2.5$ inches. The z -score for height is

$$z = \frac{\text{height} - 64.5}{2.5}$$

A woman's standardized height is the number of standard deviations by which her height differs from the mean height of all young women. A woman 68 inches tall, for example, has z -score

$$z = \frac{68 - 64.5}{2.5} = 1.4$$

or a height that is 1.4 standard deviations above the mean. Similarly, a woman 5 feet (60 inches) tall has z -score

$$z = \frac{60 - 64.5}{2.5} = -1.8$$

or a height that is 1.8 standard deviations less than the mean.

CHECK-IN

1.31 Find the z -score. Consider the NAEP scores (see Check-in question 1.29, page 52), which we assume are approximately Normal, $N(282, 40)$. Give the z -score for a student who received a score of 300.

1.32 Find another z -score. Consider the NAEP scores, which we assume are approximately Normal, $N(282, 40)$. Give the z -score for a student who received a score of 200. Explain why your answer is negative even though all the test scores are positive.

We need a way to write variables, such as “height” in Example 1.34, that follow a theoretical distribution such as a Normal distribution. We use capital letters near the end of the alphabet for such variables. If X is the height of a young woman, we can then shorten “the height of a young woman is less than 68 inches” to “ $X < 68$.” We will use lowercase x to stand for any specific value of the variable X .

We often standardize observations from symmetric distributions to express them in a common scale. We might, for example, compare the heights of two children of different ages by calculating their z -scores. The standardized heights tell us where each child stands in the distribution for his or her age group.

Standardizing is a linear transformation that transforms the data into the standard scale of z -scores. We know that a linear transformation does not change the shape of a distribution and that the mean and standard deviation change in a simple manner. In particular, the standardized values for any distribution always have mean 0 and standard deviation 1.

If the variable we standardize has a Normal distribution, standardizing does more than give a common scale. It makes all Normal distributions into a single distribution, and this distribution is still Normal. Standardizing a variable that has any Normal distribution produces a new variable that has the *standard Normal distribution*.

The standard Normal distribution

The **standard Normal distribution** is the Normal distribution $N(0,1)$ with mean 0 and standard deviation 1.

If a variable X has any Normal distribution $N(\mu,\sigma)$ with mean μ and standard deviation σ , then the **standardized variable**

$$Z = \frac{X - \mu}{\sigma}$$

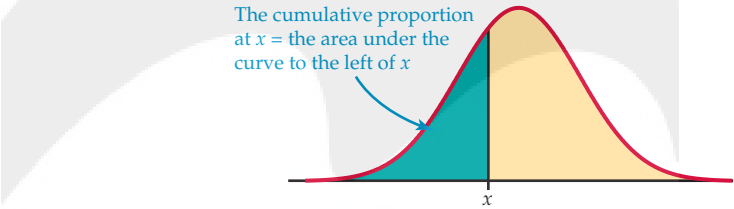
has the standard Normal distribution.

Normal distribution calculations

cumulative proportion

Areas under a Normal curve represent proportions of observations from that Normal distribution. There is no formula for areas under a Normal curve. Calculations use either software that calculates areas or a table of areas. The table and most software calculate one kind of area: **cumulative proportion**, which is the proportion of observations in a distribution that lie at or below a given value. When the distribution is given by a density curve, the cumulative proportion is the area under the curve to the left of a given value. **FIGURE 1.26** shows the idea more clearly than words do.

FIGURE 1.26 The cumulative proportion for a value x is the proportion of all observations from the distribution that are less than or equal to x . This is the area to the left of x under the Normal curve.



The key to calculating Normal proportions is to match the area you want with areas that represent cumulative proportions. Then get areas for cumulative proportions either from software or (with an extra step) from a table. The following examples show the method in pictures.

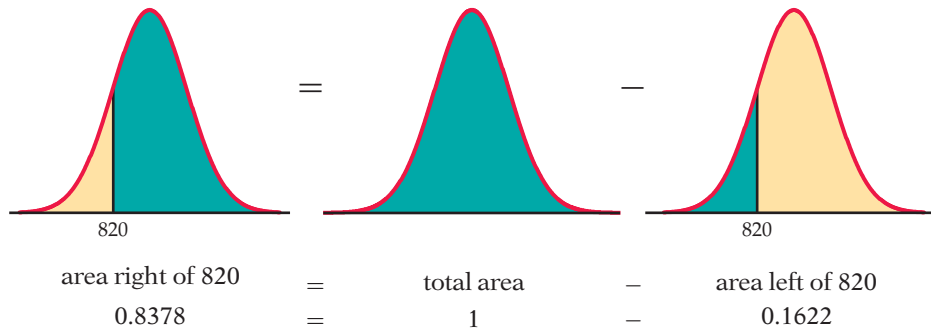
EXAMPLE 1.36

The NCAA standard for SAT scores. The National Collegiate Athletic Association (NCAA) requires Division I athletes to get a combined score of at least 820 on the SAT Mathematics and Verbal tests to compete in their first college year.²⁹ (Higher scores are required for students with poor high school grades.) The scores of the 1.4 million students who took the SATs were approximately Normal with mean 1026 and standard deviation 209. What proportion of all students had SAT scores of at least 820?

Here is the calculation in pictures: the proportion of scores above 820 is the area under the curve to the right of 820. That's the total area under the curve (which is always 1) minus the cumulative proportion up to 820. Note that we have used software for these calculations.



Mitchell Layton/Getty Images

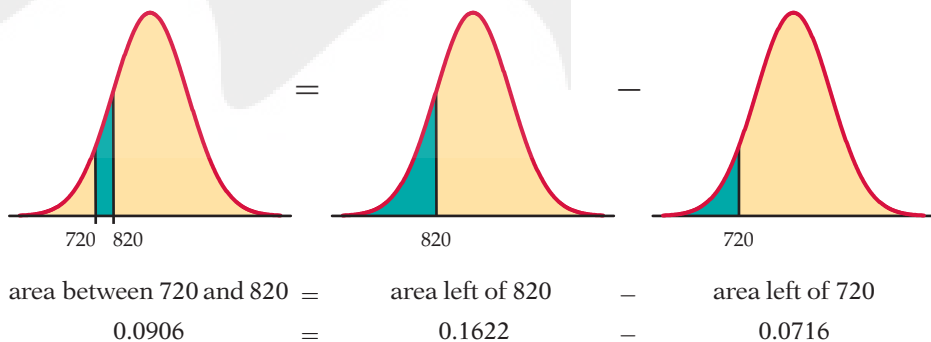


Thus, the proportion of all SAT test-takers who would be NCAA qualifiers is 0.8378, or about 84%.

There is *no* area under a smooth curve that is exactly over the point 820. Consequently, the area to the right of 820 (the proportion of scores > 820) is the same as the area at or to the right of this point (the proportion of scores ≥ 820). The actual data may contain a student who scored exactly 820 on the SAT. That the proportion of scores exactly equal to 820 is 0 for a Normal distribution is a consequence of the idealized smoothing of Normal distributions for data.

EXAMPLE 1.37

Partial qualifiers. The NCAA considers a student to be a “partial qualifier”—eligible to practice and receive an athletic scholarship, but not compete—if the combined SAT score is at least 720.³⁰ What proportion of all students who take the SAT would be partial qualifiers? That is, what proportion have scores between 720 and 820? Here are the pictures:



About 9% of all students who take the SAT have scores between 720 and 820.

How do we find the numerical values of the areas in Examples 1.36 and 1.37? If you use software, just plug in mean 1026 and standard deviation 209. Then ask for the cumulative proportions for 820 and for 720. (Your software will probably refer to these as “cumulative probabilities.” We will learn in Chapter 4 why the language of probability fits.) Sketches of the areas that you want similar to the ones in Examples 1.36 and 1.37 are very helpful in making sure that you are doing the correct calculations.



You can use the *Normal Curve* applet on the text website to find Normal proportions. The applet is more flexible than most software—it will find any Normal proportion, not just cumulative proportions. The applet is an excellent way to understand Normal curves. But, because of the limitations of web browsers, the applet is not as accurate as statistical software.

If you are not using software, you can find cumulative proportions for Normal curves from a table. That requires an extra step, as we now explain.

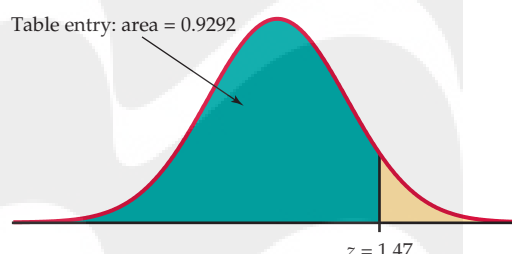
Using the standard Normal table

The extra step in finding cumulative proportions from a table is that we must first standardize to express the problem in the standard scale of z -scores. This allows us to get by with just one table, a table of *standard Normal cumulative proportions*. Table A in the back of the book gives standard Normal probabilities. The picture at the top of the table reminds us that the entries are cumulative proportions, areas under the curve to the left of a value z .

EXAMPLE 1.38

Find the proportion from z . What proportion of observations on a standard Normal variable Z take values less than 1.47? We need to find the area to the left of 1.47, locate 1.4 in the left-hand column of Table A and then locate the remaining digit 7 as .07 in the top row. The entry opposite 1.4 and under .07 is .9292. This is the cumulative proportion we seek. FIGURE 1.27 illustrates this area.

FIGURE 1.27 The area under a standard Normal curve to the left of the point $z = 1.47$ is 0.9292, Example 1.38.



Now that you see how Table A works, let's redo the NCAA Examples 1.36 and 1.37 using the table.

EXAMPLE 1.39

Find the proportion from x . What proportion of college-bound students who take the SAT have scores of at least 820? The picture that leads to the answer is exactly the same as in Example 1.36. The extra step is that we first standardize to read cumulative proportions from Table A. If x is SAT score, we want the proportion of students for which $X \geq x$, where $x = 820$.

1. *Standardize.* Subtract the mean, then divide by the standard deviation, to transform the problem about x into a problem about a standard Normal z :

$$\begin{aligned} X &\geq 820 \\ \frac{X - 1026}{209} &\geq \frac{820 - 1026}{209} \\ Z &\geq -0.99 \end{aligned}$$

2. *Use the table.* Look at the pictures in Example 1.36. From Table A, we see that the proportion of observations less than -0.99 is 0.1611 . The area to the right of -0.99 is therefore $1 - 0.1611 = 0.8389$. This is about 84%.

The area from the table in Example 1.39 (0.8389) is slightly less accurate than the area from software in Example 1.36 (0.8378) because we must round z to two places when we use Table A. The difference is rarely important in practice.

EXAMPLE 1.40

Eligibility for aid and practice. What proportion of all students who take the SAT would be eligible to receive athletic scholarships and to practice with the team but would not be eligible to compete in the eyes of the NCAA? That is, what proportion of students have SAT scores between 720 and 820? First, sketch the areas, exactly as in Example 1.37. We again use X as shorthand for an SAT score.

1. *Standardize.*

$$\begin{aligned} 720 &\leq X < 820 \\ \frac{720 - 1026}{209} &\leq \frac{X - 1026}{209} < \frac{820 - 1026}{209} \\ -1.46 &\leq Z < -0.99 \end{aligned}$$

2. *Use the table.*

$$\begin{aligned} \text{area between } -1.46 \text{ and } -0.99 &= (\text{area left of } -0.99) - (\text{area left of } -1.46) \\ &= 0.1611 - 0.0721 = 0.0890 \end{aligned}$$

As in Example 1.37, about 9% of students would be eligible to receive athletic scholarships and to practice with the team.

Sometimes we encounter a value of z more extreme than those appearing in Table A. For example, the area to the left of $z = -4$ is not given in the table. The z -values in Table A leave only area 0.0002 in each tail unaccounted for. For practical purposes, we can act as if there is zero area outside the range of Table A.

CHECK-IN

1.33 Find the proportion. Consider the NAEP scores, which are approximately Normal, $N(282, 40)$. Find the proportion of students who have scores less than 350. Find the proportion of students who have scores greater than or equal to 350. Sketch the relationship between these two calculations using pictures of Normal curves similar to the ones given in Example 1.36 (page 54).

1.34 Find another proportion. Consider the NAEP scores, which are approximately Normal, $N(282, 40)$. Find the proportion of students who have scores between 300 and 350. Use pictures of Normal curves similar to the ones given in Example 1.37 (page 55) to illustrate your calculations.

Inverse Normal calculations

Examples 1.36 to 1.40 illustrate the use of Normal distributions to find the proportion of observations in a given event, such as “SAT score between 720 and 820.” We may instead want to find the observed value corresponding to a given proportion.

Statistical software will do this directly. Without software, use Table A backward, finding the desired proportion in the body of the table and then reading the corresponding z from the left column and top row.

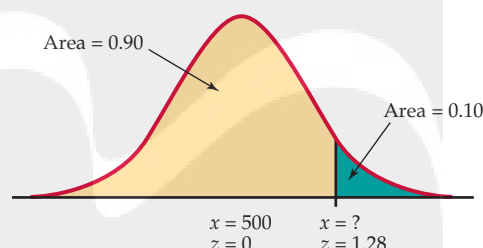
EXAMPLE 1.41

How high for the top 10%? Scores for college-bound students on the SAT Verbal test in recent years follow approximately the $N(500, 110)$ distribution.³¹ How high must a student score to place in the top 10% of all students taking the SAT?

Again, the key to the problem is to draw a picture. FIGURE 1.28 shows that we want the score x with an area of 0.10 above it. That’s the same as area below x equal to 0.90.

Statistical software has a function that will give you the x for any cumulative proportion you specify. The function often has a name such as “inverse cumulative probability.” Plug in mean 500, standard deviation 110, and cumulative proportion 0.9. The software tells you that $x = 641$ to place in the highest 10%.

FIGURE 1.28 Locating the point on a Normal curve with area 0.10 to its right, Example 1.41.



Without software, first find the standard score z with cumulative proportion 0.9, then “unstandardize” to find x . Here is the two-step process:

1. *Use the table.* Look in the body of Table A for the entry closest to 0.9. It is 0.8997. This is the entry corresponding to $z = 1.28$. So $z = 1.28$ is the standardized value with area 0.9 to its left.
2. *Unstandardize* to transform the solution from z back to the original x scale. We know that the standardized value of the unknown x is $z = 1.28$. So x itself satisfies

$$\frac{x - 500}{110} = 1.28$$

Solving this equation for x gives

$$x = 500 + (1.28)(110) = 640.8$$

This equation should make sense: it finds the x that lies 1.28 standard deviations above the mean on this particular Normal curve. That is the

“unstandardized” meaning of $z = 1.28$. The general rule for unstandardizing a z -score is

$$x = \mu + z\sigma$$

CHECK-IN

1.35 What score is needed to be in the top 20%? Consider the NAEP scores, which are approximately Normal, $N(282, 40)$. How high a score is needed to be in the top 20% of students who take this exam?

1.36 Find the score that 75% of students will exceed. Consider the NAEP scores, which are approximately Normal, $N(282, 40)$. Seventy-five percent of the students will score above x on this exam. Find x .

Normal quantile plots

The Normal distributions provide good descriptions of some distributions of real data, such as the Iowa Test vocabulary scores. The distributions of some other common variables are usually skewed and therefore distinctly non-Normal. Examples include economic variables such as personal income and gross sales of business firms, the survival times of cancer patients after treatment, and the service lifetime of mechanical or electronic components. While experience can suggest whether or not a Normal distribution is plausible in a particular case, it is risky to assume that a distribution is Normal without actually inspecting the data.

A histogram or stemplot can reveal distinctly non-Normal features of a distribution, such as outliers, pronounced skewness, or gaps and clusters. If the stemplot or histogram appears roughly symmetric and unimodal, however, we need a more sensitive way to judge the adequacy of a Normal model. The most useful tool for assessing Normality is another graph, the **Normal quantile plot**.

Here is the basic idea of a Normal quantile plot. The graphs produced by software use more sophisticated versions of this idea. It is not practical to make Normal quantile plots by hand.

1. Arrange the observed data values from smallest to largest. Record what percentile of the data each value occupies. For example, the smallest observation in a set of 20 is at the 5% point, the second smallest is at the 10% point, and so on.
2. Do Normal distribution calculations to find the values of z corresponding to these same percentiles. For example, $z = -1.645$ is the 5% point of the standard Normal distribution, and $z = -1.282$ is the 10% point. We call these values of Z **Normal scores**.
3. Plot each data point x against the corresponding Normal score. If the data distribution is close to any Normal distribution, the plotted points will lie close to a straight line.

Any Normal distribution produces a straight line on the plot because standardizing turns any Normal distribution into a standard Normal distribution. Standardizing is a linear transformation that can change the slope and intercept of the line in our plot but cannot turn a line into a curved pattern.

Normal quantile plot

Normal scores

Use of Normal quantile plots

- If the points on a **Normal quantile plot** lie close to a straight line, the plot indicates that the data are Normal.
- Systematic deviations from a straight line indicate a non-Normal distribution.
- Outliers appear as points that are far away from the overall pattern of the plot.
- An optional line can be drawn on the plot that corresponds to the Normal distribution with mean equal to the mean of the data and standard deviation equal to the standard deviation of the data.

Figures 1.29 and 1.30 are Normal quantile plots for data we have met earlier. The data x are plotted vertically against the corresponding standard Normal z -score plotted horizontally. The z -score scale generally extends from -3 to 3 because almost all of a standard Normal curve lies between these values. These figures show how Normal quantile plots behave.

EXAMPLE 1.42



IQ scores are approximately Normal. FIGURE 1.29 is a Normal quantile plot of the 60 fifth-grade IQ scores from Table 1.1 (page 15). The points lie very close to the straight line drawn on the plot. We conclude that the distribution of IQ data is approximately Normal.

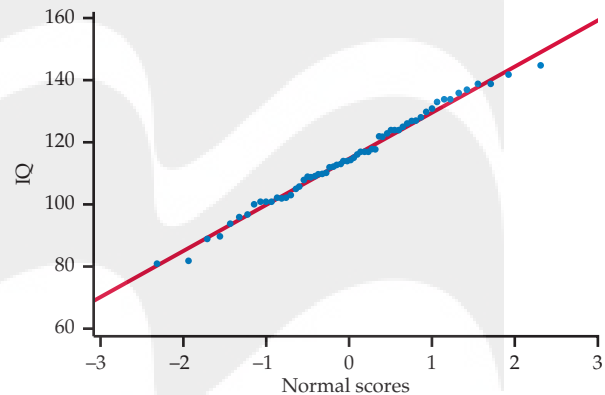


FIGURE 1.29 Normal quantile plot of IQ scores, Example 1.42. This distribution is approximately Normal.

EXAMPLE 1.43



Times to start a business are skewed. FIGURE 1.30 is a Normal quantile plot of the data on times to start a business from Example 1.19. The line drawn on the plot shows clearly that the plot of the data is curved. We conclude that these data are not Normally distributed. The shape of the curve is what we typically see with a distribution that is strongly skewed to the right.

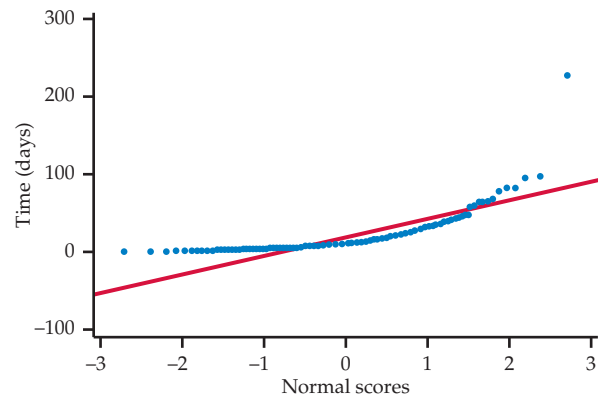


FIGURE 1.30 Normal quantile plot for the length of time required to start a business, Exercise 1.43. This distribution is highly skewed.



Real data often show some departure from the theoretical Normal model. When you examine a Normal quantile plot, look for shapes that show clear departures from Normality. Don't overreact to minor wiggles in the plot. When we discuss statistical methods that are based on the Normal model, we are interested in whether or not the data are sufficiently Normal for these procedures to work properly. We are not concerned about minor deviations from Normality. Many common methods work well as long as the data are approximately Normal and outliers are not present.

Beyond the Basics

Density estimation

density estimator

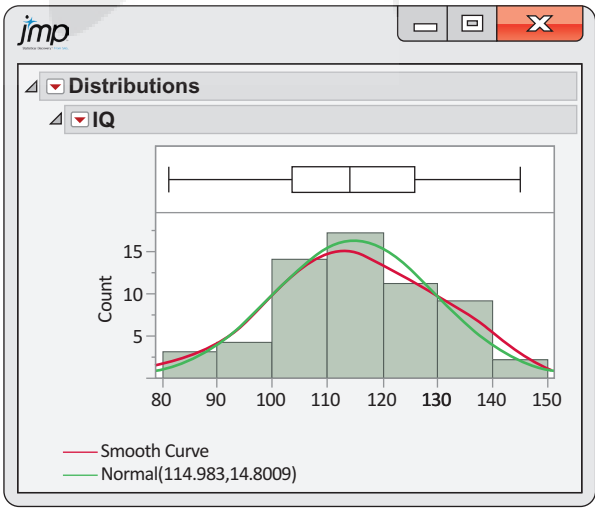
A density curve gives a compact summary of the overall shape of a distribution. Many distributions do not have the Normal shape. There are other families of density curves that are used as mathematical models for various distribution shapes. Modern software offers more flexible options. A **density estimator** does not start with any specific shape, such as the Normal shape. It looks at the data and draws a density curve that describes the overall shape of the data. Density estimators join stemplots and histograms as useful graphical tools for exploratory data analysis.

EXAMPLE 1.44



Density estimation for IQ scores. In Example 1.42 we observed that the points in the Normal quantile plot for the IQ data were very close to a straight line. This suggests that a Normal distribution is a good fit for these data. FIGURE 1.31 provides another way to look at this issue. Here we see the histogram with a density estimate, the red curve, along with the best-fitting Normal density curve, the green curve. Because the two curves are approximately the same, we are confident in any further analysis of these data based on the assumption that the data are approximately Normal.

FIGURE 1.31 Histogram of IQ scores, with a density estimate and a Normal curve, Example 1.44. The IQ scores are approximately Normal.

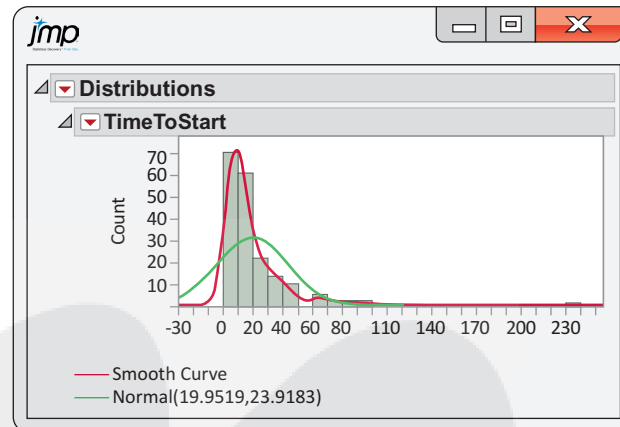


Here is another example where we see a different picture.

EXAMPLE 1.45

Density estimation for times to start a business. In Example 1.43, we examined the Normal quantile plot for the time to start a business data. FIGURE 1.32 shows the histogram for these data along with a density estimate and the best-fitting Normal distribution. The two density curves are very different, and we conclude that a Normal distribution does not give a good fit for these data. Not only are the data strongly skewed, but there is also a clear outlier. We should be very cautious about using a statistical analysis based on an assumption that the data are approximately Normal in this case.

FIGURE 1.32 Histogram of the length of time required to start a business, with a density estimate and a Normal curve, Example 1.45. The Normal distribution is not a good fit for these data.



Section 1.4 SUMMARY

- We can describe the overall pattern of a distribution by a **density curve**. A density curve has total area 1 underneath it. An area under a density curve gives the proportion of observations that fall in a range of values.
- A density curve is an idealized description of the overall pattern of a distribution that smooths out the irregularities in the actual data. We write the mean of a density curve as μ and the standard deviation of a density curve as σ to distinguish them from the mean \bar{x} and the standard deviation s of the actual data.
- The **mean** μ is the balance point of the curve. The **median** divides the area under the curve in half. The **quartiles** and the median divide the area under the curve into quarters. The **standard deviation** σ cannot be located by eye on most density curves.
- The mean and median are equal for symmetric density curves, but the mean of a skewed curve is located farther toward the long tail than is the median.
- The **Normal distributions** are described by a special family of bell-shaped, symmetric, unimodal density curves. The mean μ and standard deviation σ completely specify a Normal distribution $N(\mu, \sigma)$. The mean is the center of the curve, and σ is the distance from μ to the change-of-curvature points on either side.
- To **standardize** any observation x , subtract the mean of the distribution and then divide by the standard deviation. The resulting **z-score** $z = (x - \mu) / \sigma$ says how many standard deviations x lies from the distribution mean.
- All Normal distributions are the same when measurements are transformed to the standardized scale. In particular, all Normal distributions satisfy the **68–95–99.7 rule**, which describes what percent of observations lie within one, two, and three standard deviations of the mean.
- If x has the $N(\mu, \sigma)$ distribution, then the standardized variable $Z = (X - \mu) / \sigma$ has the **standard Normal distribution** $N(0, 1)$. Proportions for any Normal distribution can be calculated by software or from the **standard Normal table** (Table A), which gives the **cumulative proportions** of $Z < z$ for many values of x .
- The adequacy of a Normal model for describing a distribution of data is best assessed by a **Normal quantile plot**, which is available in most statistical software packages. A pattern on such a plot that deviates substantially from a straight line indicates that the data are not Normal.

Now that you have completed this section, you will be able to:

- Sketch a Normal distribution for any given mean and standard deviation. *Review Example 1.34 (page 51) and try Exercise 1.65.*
- Apply the 68–95–99.7 rule to find the proportions of observations within one, two, and three standard deviations of the mean for any Normal distribution. *Review Example 1.34 (page 51) and try Exercise 1.65.*
- Find the z-score for any observation x . *Review Example 1.35 (page 53) and try Exercise 1.67.*
- Compute areas under a Normal curve using software or Table A. *Review Example 1.36 (page 54) and try Exercise 1.79.*
- Perform inverse Normal calculations to find values of a Normal variable corresponding to any given area. *Review Example 1.41 (page 58) and try Exercise 1.81.*
- Assess the extent to which the distribution of a set of data can be approximated by a Normal distribution. *Review Examples 1.42 (page 60) and 1.43 (page 60) and try Exercise 1.105.*

Section 1.4 EXERCISES

1.61 What's wrong? Explain what is wrong with each of the following:

- Standardized values are always positive.
- Ninety-five percent of the values of a Normal distribution will be within one standard deviation of the mean.
- The standard Normal distribution has mean equal to 1 and standard deviation equal to 0.

1.62 Means and medians.

- Sketch a symmetric distribution that is *not* Normal. Mark the location of the mean and the median.
- Sketch a distribution that is skewed to the right. Mark the location of the mean and the median.

1.63 The effect of changing the standard deviation.

- Sketch a Normal curve that has mean 20 and standard deviation 2.
- On the same x axis, sketch a Normal curve that has mean 20 and standard deviation 4.
- How does the Normal curve change when the standard deviation is varied but the mean stays the same?

1.64 The effect of changing the mean.

- Sketch a Normal curve that has mean 20 and standard deviation 2.
- On the same x axis, sketch a Normal curve that has mean 30 and standard deviation 2.
- How does the Normal curve change when the mean is varied but the standard deviation stays the same?

1.65 NAEP eighth-grade geography scores. In Check-in question 1.29 (page 52) we examined the

distribution of NAEP scores for the eighth-grade mathematics skills assessment. For eighth-grade students, the average geography score is approximately Normal, with mean 261 and standard deviation 31.

- Sketch this Normal distribution.
- Make a table that includes values of the scores corresponding to plus or minus one, two, and three standard deviations from the mean. Mark these points on your sketch along with the mean.
- Apply the 68–95–99.7 rule to this distribution. Give the ranges of reading score values that are within one, two, and three standard deviations of the mean.

1.66 NAEP 12th-grade geography scores. Refer to the previous exercise. The scores for 12th-grade students on the geography assessment are approximately $N(282, 26)$. Answer the questions in the previous exercise for this assessment.

1.67 Standardize some NAEP eighth-grade geography scores. The NAEP geography assessment scores for eighth-grade students are approximately $N(261, 31)$. Find z-scores by standardizing the following scores: 200, 250, 280, 300, 320.

1.68 Compute the percentile scores. Refer to the previous exercise. When scores such as the NAEP assessment scores are reported for individual students, the actual values of the scores are not particularly meaningful. Usually, they are transformed into percentile scores. The percentile score is the proportion of students who would score less than or equal to the score for the individual student. Compute the percentile scores for the five scores in the previous exercise. State whether you used software or Table A for these computations.

1.69 Are the NAEP eighth-grade geography scores approximately Normal? In Exercise 1.65, we assumed that the NAEP U.S. geography scores for eighth-grade students are approximately Normal with the reported mean and standard deviation, $N(261, 31)$. Let's check that assumption. In addition to means and standard deviations, you can find selected percentiles for the NAEP assessments (see previous exercise). For the 8th-grade geography scores, the following percentiles are reported:

Percentile	Score
10%	220
25%	242
50%	263
75%	283
90%	300

Use these percentiles to assess whether or not the NAEP geography scores for 8th-grade students are approximately Normal. Write a short report describing your methods and conclusions.

1.70 Are the NAEP eighth-grade mathematics scores approximately Normal? Refer to the previous exercise. For the NAEP eighth-grade mathematics scores, the mean is 282, and the standard deviation is 40. Here are the reported percentiles:

Percentile	Score
10%	231
25%	255
50%	282
75%	309
90%	333

Is the $N(282, 40)$ distribution a good approximation for the NAEP mathematics scores? Write a short report describing your methods and conclusions.

1.71 Do women talk more? Conventional wisdom suggests that women are more talkative than men. One study designed to examine this stereotype collected data on the speech of 42 women and 37 men in the United States.³²

- (a) The mean number of words spoken per day by the women was 14,297, with a standard deviation of 6441. Use the 68–95–99.7 rule to describe this distribution.
- (b) Do you think that applying the rule in this situation is reasonable? Explain your answer.
- (c) The men averaged 14,060 words per day, with a standard deviation of 9065. Answer the questions in parts (a) and (b) for the men.
- (d) Do you think that the data support the conventional wisdom? Explain your answer. Note that in Section 7.2 we will learn formal statistical methods to answer this type of question.

1.72 Data from Mexico. Refer to the previous exercise. A similar study in Mexico was conducted with 31 women and 20 men. The women averaged 14,704 words per day, with a standard deviation of 6215. For men the mean was 15,022, and the standard deviation was 7864.

- (a) Answer the questions from the previous exercise for the Mexican study.
- (b) The means for both men and women are higher for the Mexican study than for the U.S. study. What conclusions can you draw from this observation?

1.73 A uniform distribution. If you ask a computer to generate “random numbers” between 0 and 1, you will get observations from a **uniform distribution**. FIGURE 1.33 graphs the density curve for a uniform distribution. Use areas under this density curve to answer the following questions.

- (a) What proportion of the observations lie below 0.75?
- (b) What proportion of the observations lie below 0.50?
- (c) What proportion of the observations lie between 0.50 and 0.75?
- (d) Why is the total area under this curve equal to 1?

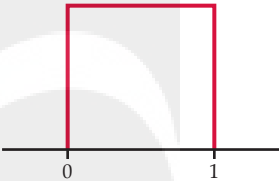


FIGURE 1.33 The density curve of a uniform distribution, Exercise 1.73.

1.74 Use a different range for the uniform distribution. Many random number generators allow users to specify the range of the random numbers to be produced. Suppose that you specify that the outcomes are to be distributed uniformly between 0 and 5. Then the density curve of the outcomes has constant height between 0 and 5 and height 0 elsewhere.

- (a) What is the height of the density curve between 0 and 5? Draw a graph of the density curve.
- (b) Use your graph from part (a) and the fact that areas under the curve are proportions of outcomes to find the proportion of outcomes that are more than 2.
- (c) Find the proportion of outcomes that lie between 2.5 and 3.0.

1.75 Find the mean, the median, and the quartiles. What are the mean and the median of the uniform distribution in Figure 1.33? What are the quartiles?

1.76 Three density curves. FIGURE 1.34 displays three density curves, each with three points marked on it. At which of these points on each curve do the mean and the median fall?

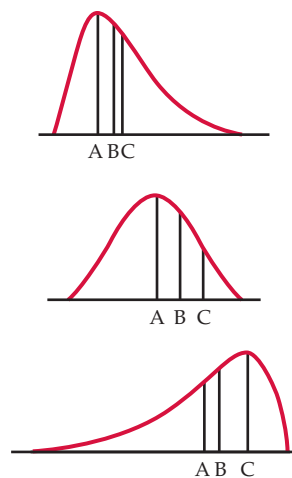


FIGURE 1.34 Three density curves, Exercise 1.76.

1.77 Use the Normal Curve applet. Use the *Normal Curve* applet for the standard Normal distribution to say how many standard deviations above and below the mean the quartiles of any Normal distribution lie.

1.78 Use the Normal Curve applet. The 68–95–99.7 rule for Normal distributions is a useful approximation. You can use the *Normal Curve* applet on the text website to see how accurate the rule is. Drag one flag across the other so that the applet shows the area under the curve between the two flags.

(a) Place the flags one standard deviation on either side of the mean. What is the area between these two values? What does the 68–95–99.7 rule say this area is?

(b) Repeat for locations two and three standard deviations on either side of the mean. Again compare the 68–95–99.7 rule with the area given by the applet.

1.79 Find some proportions. Using either software or Table A, find the proportion of observations from a standard Normal distribution that satisfies each of the following statements. In each case, sketch a standard Normal curve and shade the area under the curve that is the answer to the question.

- (a) $Z > 1.85$
- (b) $Z < 1.85$
- (c) $Z > -0.90$
- (d) $-0.90 < Z < 1.85$

1.80 Find more proportions. Using either software or Table A, find the proportion of observations from a standard Normal distribution for each of the following events. In each case, sketch a standard Normal curve and shade the area representing the proportion.

- (a) $Z \leq -1.7$
- (b) $Z \geq -1.7$
- (c) $Z > 2.1$
- (d) $-1.7 < Z < 2.1$

1.81 Find some values of z . Find the value z of a standard Normal variable Z that satisfies each of the following conditions. (If you use Table A, report the value of z that comes closest to satisfying the condition.) In each case, sketch a standard Normal curve with your value of z marked on the axis.

- (a) 68% of the observations fall below z
- (b) 75% of the observations fall above z

1.82 Find more values of z illustrate the result with a sketch. The variable Z has a standard Normal distribution.

- (a) Find the number z that has cumulative proportion 0.68.
- (b) Find the number z such that the event $Z > z$ has proportion 0.122.

1.83 Find some values of z . The Wechsler Adult Intelligence Scale (WAIS) is the most common IQ test. The scale of scores is set separately for each age group, and the scores are approximately Normal, with mean 100 and standard deviation 15. People with WAIS scores below 70 are considered developmentally disabled when, for example, applying for Social Security disability benefits. What percent of adults are developmentally disabled by this criterion?

1.84 High IQ scores. Refer to the previous exercise. The organization MENSA, which calls itself “the high-IQ society,” requires a WAIS score of 130 or higher for membership. What percent of adults would qualify for membership?

There are two major tests of readiness for college, the ACT and the SAT. ACT scores are reported on a scale from 1 to 36. The distribution of ACT scores is approximately Normal, with mean $\mu = 21.5$ and standard deviation $\sigma = 5.4$. SAT scores are reported on a scale from 400 to 1600. The distribution of SAT scores is approximately Normal, with mean $\mu = 1026$ and standard deviation $\sigma = 209$. Exercises 1.85 through 1.94 are based on this information.

1.85 Compare an SAT score with an ACT score.

Jessica scores 1240 on the SAT. Ashley scores 28 on the ACT. Assuming that both tests measure the same thing, who has the higher score? Report the z -scores for both students.

1.86 Make another comparison. Joshua scores 14 on the ACT. Anthony scores 690 on the SAT. Assuming that both tests measure the same thing, who has the higher score? Report the z -scores for both students.

1.87 Find the ACT equivalent. Jorge scores 1400 on the SAT. Assuming that both tests measure the same thing, what score on the ACT is equivalent to Jorge’s SAT score?

1.88 Find the SAT equivalent. Alyssa scores 32 on the ACT. Assuming that both tests measure the same thing, what score on the SAT is equivalent to Alyssa’s ACT score?

1.89 Find an SAT percentile. Reports on a student's ACT or SAT results usually give the percentile as well as the actual score. The percentile is just the cumulative proportion stated as a percent: the percent of all scores that were lower than or equal to this one. Renee scores 1360 on the SAT. What is her percentile?

1.90 Find an ACT percentile. Reports on a student's ACT or SAT results usually give the percentile as well as the actual score. The percentile is just the cumulative proportion stated as a percent: the percent of all scores that were lower than or equal to this one. Joshua scores 21 on the ACT. What is his percentile?

1.91 How high is the top 15%? What SAT scores make up the top 15% of all scores?

1.92 How low is the bottom 15%? What SAT scores make up the bottom 15% of all scores?

1.93 Find the ACT quintiles. The quintiles of any distribution are the values with cumulative proportions 0.20, 0.40, 0.60, and 0.80. What are the quintiles of the distribution of ACT scores?

1.94 Find the SAT quartiles. The quartiles of any distribution are the values with cumulative proportions 0.25 and 0.75. What are the quartiles of the distribution of SAT scores?

1.95 Do you have enough “good cholesterol”? High-density lipoprotein (HDL) is sometimes called the “good cholesterol” because high values are associated with a reduced risk of heart disease. According to the American Heart Association, people over the age of 20 years should have at least 40 milligrams per deciliter (mg/dl) of HDL cholesterol.³³ U.S. women aged 20 and over have a mean HDL of 55 mg/dl with a standard deviation of 15.5 mg/dl. Assume that the distribution is Normal.

- (a) What percent of women have low values of HDL (40 mg/dl or less)?
- (b) HDL levels of 60 mg/dl and higher are believed to protect people from heart disease. What percent of women have protective levels of HDL?
- (c) Women with more than 40 mg/dl but less than 60 mg/dl of HDL are in the intermediate range, neither very good or very bad. What proportion are in this category?

1.96 Men and HDL cholesterol. HDL cholesterol levels for men have a mean of 46 mg/dl, with a standard deviation of 13.6 mg/dl. Assume that the distribution is Normal. Answer the questions given in the previous exercise for the population of men.


1.97 Diagnosing osteoporosis. Osteoporosis is a condition in which the bones become brittle due to loss of minerals. To diagnose osteoporosis, an elaborate apparatus measures bone mineral density (BMD). BMD is usually reported in standardized form. The standardization is based on a population of healthy young adults. The World Health Organization (WHO) criterion for osteoporosis is a BMD 2.5 standard

deviations below the mean for young adults. BMD measurements in a population of people similar in age and sex roughly follow a Normal distribution.


- (a) What percent of healthy young adults have osteoporosis by the WHO criterion?
- (b) Women aged 70 to 79 are of course not young adults. The mean BMD in this age is about -2 on the standard scale for young adults. Suppose that the standard deviation is the same as for young adults. What percent of this older population has osteoporosis?

1.98 Deciles of Normal distributions. The **deciles** of any distribution are the 10th, 20th, . . . , 90th percentiles. The first and last deciles are the 10th and 90th percentiles, respectively.


- (a) What are the first and last deciles of the standard Normal distribution?
- (b) The weights of 9-ounce potato chip bags are approximately Normal, with mean 9.11 ounces and standard deviation 0.14 ounce. What are the first and last deciles of this distribution?

 **1.99 Quartiles for Normal distributions.** The quartiles of any distribution are the values with cumulative proportions 0.25 and 0.75.


- (a) What are the quartiles of the standard Normal distribution?
- (b) Using your numerical values from (a), write an equation that gives the quartiles of the $N(\mu, \sigma)$ distribution in terms of μ and σ .


 **1.100 IQR for Normal distributions.** Continue your work from the previous exercise. The interquartile range *IQR* is the distance between the first and third quartiles of a distribution.

- (a) What is the value of the *IQR* for the standard Normal distribution?
- (b) There is a constant c such that $IQR = c\sigma$ for any Normal distribution $N(\mu, \sigma)$. What is the value of c ?


 **1.101 Outliers for Normal distributions.** Continue your work from the previous two exercises. The percent of the observations that are suspected outliers according to the $1.5 \times IQR$ rule is the same for any Normal distribution. What is this percent?

1.102 Deciles of HDL cholesterol. The deciles of any distribution are the 10th, 20th, . . . , 90th percentiles. Refer to Exercise 1.95 where we assumed that the distribution of HDL cholesterol in U.S. women aged 20 and over is Normal with mean 55 mg/dl and standard deviation 15.5 mg/dl. Find the deciles for this distribution.

1.103 Longleaf pine trees. Exercise 1.56 (page 46) gives the diameter at breast height (DBH) for 40 longleaf pine trees from the Wade Tract in Thomas County, Georgia. Make a Normal quantile plot for these data and write a short paragraph interpreting what it describes. 

1.104 Potassium from potatoes. Refer to Exercise 1.15 (page 22), where you used a stemplot to examine the potassium absorption of a group of 27 adults who ate a controlled diet that included 40 mEq of potassium from potatoes for five days. In Exercise 1.33 (page 43), you compared the stemplot, the histogram, and the boxplot as graphical summaries of this distribution.  KPOT40

- Generate these three graphical summaries.
- Make a Normal quantile plot and interpret it.

1.105 Potassium from a supplement. Refer to Exercise 1.16 (page 22), where you used a stemplot to examine the potassium absorption of a group of 29 adults who ate a controlled diet that included 40 mEq of potassium from a supplement for five days. In Exercise 1.34 (page 43), you compared the stemplot, the histogram, and the boxplot as graphical summaries of this distribution.  KSUP40

- Generate these three graphical summaries.
- Make a Normal quantile plot and interpret it.


Chapter 1 EXERCISES


1.106 Sources of energy consumed. Energy consumed in the United States can be classified as coming from one of three sources: fossil fuels, nuclear and electric power, and renewable energy. In 2018, the energy from these three sources was 81.0, 8.4, and 11.5 quadrillion Btu, respectively. In 2008, the corresponding amounts were 83.0, 8.4, and 7.2.³⁴ Write a description of the changes from 2008 to 2018 expressed in these data. Illustrate your summary with appropriate graphical summaries. Be sure to discuss both the amounts of energy from each source as well as the percents.


1.107 Sources of renewable energy consumed. Refer to the previous exercise. Renewable energy is classified into five sources. Here are the 2008 and 2018 energy data for these sources:

Source	Amount	
	2008	2018
Hydroelectric	2.511	2.688
Geothermal	0.192	0.218
Solar	0.074	0.951
Wind	0.546	2.533
Biomass	3.851	5.132

Write a description of the changes from 2008 to 2018 expressed in these data. Illustrate your summary with appropriate graphical summaries. Be sure to discuss both the amounts of energy from each source as well as the percents.

1.108 CO₂ emissions in vehicles. Natural Resources Canada tests new vehicles each year and reports several variables related to fuel consumption for vehicles in different classes.³⁵ For 2018, it provides data for 502 vehicles that use conventional fuel. Two variables reported are carbon dioxide (CO₂) emissions and highway fuel consumption. CO₂ is measured in grams per kilometer (g/km), and highway fuel consumption measured in liters per 100 kilometers (L/km). Use graphical and numerical summaries to describe the distribution of CO₂ emissions for these vehicles. Be sure to justify your choice of summaries.  CANFREG

1.109 Highway fuel consumption. Refer to the previous exercise. Use graphical and numerical summaries to describe the distribution of highway fuel consumption for these vehicles. Be sure to justify your choice of summaries.  CANFREG

1.110 Flopping in the World Cup. Soccer players are often accused of spending an excessive amount of time dramatically falling to the ground followed by other activities, in attempts to show that a possible injury is very serious. It has been suggested that these tactics are often designed to influence the call of a referee or to take extra time off the clock. Recordings of the first 32 games of the 2014 World Cup were analyzed, and there were 302 times when the referee interrupted the match because of possible injuries. The number of injuries and the total time, in minutes, spent flopping for each of the 32 teams who participated in these matches was recorded.³⁶ Here are the data:  FLOPS

Country	Injuries	Time
Brazil	17	3.30
Chile	16	6.97
Honduras	15	7.67
Nigeria	15	6.42
Mexico	15	3.97
Costa Rica	13	3.80
USA	12	6.40
Ecuador	12	4.55
France	10	7.32
South Korea	10	4.52
Algeria	10	4.05
Iran	9	5.43
Russia	9	5.27
Ivory Coast	9	4.63
Croatia	9	4.32
Colombia	9	4.32
Uruguay	9	4.12
Greece	9	2.65
Cameroon	8	3.15
Germany	8	1.97
Spain	8	1.82

Continued

Country	Injuries	Time
Belgium	7	3.38
Japan	7	2.08
Italy	7	1.60
Switzerland	7	1.35
England	7	3.13
Argentina	6	2.80
Ghana	6	1.85
Australia	6	1.83
Portugal	4	1.82
Netherlands	4	1.65
Bosnia and Herzegovina	2	0.40


Describe these data using the methods you learned in this chapter and write a short summary about flopping in the 2014 World Cup based on your analysis.

1.111 What graph would you use? What type of graph or graphs would you plan to make in a study of each of the following issues?


- (a) What makes of cars do students drive? How old are their cars?
- (b) How many hours per week do students study? How does the number of study hours change during a semester?
- (c) Which radio stations are most popular with students?
- (d) When many students measure the concentration of the same solution for a chemistry course laboratory assignment, do their measurements follow a Normal distribution?

1.112 Canadian international trade. The government organization Statistics Canada provides data on many topics related to Canada’s population, resources, economy, society, and culture. Go to the web page statcan.gc.ca/start-debut-eng.html. Under the “Subjects” tab, choose “International trade.” Pick some data from the resources listed and use the methods that you learned in this chapter to create graphical and numerical summaries. Write a report summarizing your findings that includes supporting evidence from your analyses.

1.113 Travel and tourism in Canada. Refer to the previous exercise. Under the “Subjects” tab, choose “Travel and tourism.” Pick some data from the resources listed and use the methods that you learned in this chapter to create graphical and numerical summaries. Write a report summarizing your findings that includes supporting evidence from your analyses.


 **1.114 Leisure time for college students.** You want to measure the amount of “leisure time” that college students enjoy. Write a brief discussion of two issues:


- (a) How will you define “leisure time”?
- (b) Once you have defined leisure time, how will you measure it?

 **1.115 How much vitamin C is needed?** The Food and Nutrition Board of the Institute of Medicine, working in cooperation with scientists from Canada, have used scientific data to answer this question for a variety of vitamins and minerals.³⁷ Their methodology assumes that needs, or requirements, follow a distribution. They have produced guidelines called dietary reference intakes for different gender-by-age combinations. For vitamin C, there are three dietary reference intakes: the estimated average requirement (EAR), which is the mean of the requirement distribution; the recommended dietary allowance (RDA), which is the intake that would be sufficient for 97% to 98% of the population; and the tolerable upper level (UL), the intake that is unlikely to pose health risks. For women aged 19 to 30 years, the EAR is 60 milligrams per day (mg/d), the RDA is 75 mg/d, and the UL is 2000 mg/d.³⁸

(a) The researchers assumed that the distribution of requirements for vitamin C is Normal. The EAR gives the mean. From the definition of the RDA, let’s assume that its value is the 97.72 percentile. Use this information to determine the standard deviation of the requirement distribution.

(b) Sketch the distribution of vitamin C requirements for 19- to 30-year-old women. Mark the EAR, the RDA, and the UL on your plot.

 **1.116 How much vitamin C do men need?** Refer to the previous exercise. For men aged 19 to 30 years, the EAR is 75 milligrams per day (mg/d), the RDA is 90 mg/d, and the UL is 2000 mg/d. Answer the questions in the previous exercise for this population.


 **1.117 How much vitamin C do women consume?** To evaluate whether or not the intake of a vitamin or mineral is adequate, comparisons are made between the intake distribution and the requirement distribution. Here is some information about the distribution of vitamin C intake, in milligrams per day, for women aged 19 to 30 years:³⁹

Percentile (mg/d)									
Mean	1st	5th	19th	25th	50th	75th	90th	95th	99th
84.1	31	42	48	61	79	102	126	142	179

(a) Use the 5th, the 50th, and the 95th percentiles of this distribution to estimate the mean and standard deviation of this distribution assuming that the distribution is Normal. Explain your method for doing this.

(b) Sketch your Normal intake distribution on the same graph with a sketch of the requirement distribution that you produced in part (b) of Exercise 1.115.

(c) Do you think that many women aged 19 to 30 years are getting the amount of vitamin C that they need? Explain your answer.

 **1.118 How much vitamin C do men consume?** To evaluate whether or not the intake of a vitamin or mineral is adequate, comparisons are made between the intake distribution and the requirement distribution. Here

is some information about the distribution of vitamin C intake, in milligrams per day, for men aged 19 to 30 years:

Percentile (mg/d)									
Mean	1st	5th	19th	25th	50th	75th	90th	95th	99th
122.2	39	55	65	85	114	150	190	217	278

- (a) Use the 5th, the 50th, and the 95th percentiles of this distribution to estimate the mean and standard deviation of this distribution assuming that the distribution is Normal. Explain your method for doing this.
- (b) Sketch your Normal intake distribution on the same graph with a sketch of the requirement distribution that you produced in Exercise 1.116.
- (c) Do you think that many men aged 19 to 30 years in the United States are getting the amount of vitamin C that they need? Explain your answer.

1.119 Time spent studying. Do women study more than men? We asked the students in a large first-year college class how many minutes they studied on a typical weeknight. Here are the responses of random samples of 30 women and 30 men from the class:

Women					Men				
170	120	180	360	240	80	120	30	90	200
120	180	120	240	170	90	45	30	120	75
150	120	180	180	150	150	120	60	240	300
200	150	180	150	180	240	60	120	60	30
120	60	120	180	180	30	230	120	95	150
90	240	180	115	120	0	200	120	120	180

- (a) Examine the data. Why are you not surprised that most responses are multiples of 10 minutes? We eliminated one student who claimed to study 30,000 minutes per night. Are there any other responses that you consider suspicious?
- (b) Make a back-to-back stemplot of these data. Report the approximate midpoints of both groups. Does it appear that women study more than men (or at least claim that they do)?
- (c) Make side-by-side boxplots of these data. Compare the boxplots with the stemplot you made in part (b). Which to you prefer? Give reasons for your answer.

1.120 Spam filters. A university department installed a spam filter on its computer system. During a 21-day period, 6693 messages were tagged as spam. How much spam you get depends on what your online habits are. Here are the counts for some students and faculty in this department (with log-in IDs changed, of course):

ID	Count	ID	Count	ID	Count	ID	Count
AA	1818	BB	1358	CC	442	DD	416
EE	399	FF	389	GG	304	HH	251
II	251	JJ	178	KK	158	LL	103

All other department members received fewer than 100 spam messages. How many did the others receive in total? Make a graph and comment on what you learn from these data.

1.121 Phish. One of the most favored songs of the band Phish is “Divided Sky.” The band plays this song at many of their concerts. Frequently, after the main theme, Trey, the guitarist, pauses before playing the resolving note.⁴⁰ The data file PHISH gives the date of each concert where “Divided Sky” was played, the venue, and the length of the pause, in minutes, for 366 concerts. Analyze the data and write a report summarizing what you have found. Be sure to include graphical and numerical summaries. Include the rationale for decisions that you made in performing your analysis. For example, did you give any consideration to the relatively large number of zeros?

PUTTING IT ALL TOGETHER

1.122 Blueberries and anthocyanins. Anthocyanins are compounds that have been associated with health benefits to the heart, bones, and brain. Blueberries are a good source of many different anthocyanins. Researchers at the Piedmont Research Station of North Carolina State University have assembled a database giving the concentrations of 18 different anthocyanins for 267 varieties of blueberries.⁴¹ Four of the anthocyanins measured are delphinidin-3-arabinoside, malvidin-3-arabinoside, cyanidin-3-galactoside, and delphinidin-3-glucoside, all measured in units of mg/100g of berries. In the data file, we have simplified the names of these anthocyanins to Antho1, Antho2, Antho3, and Antho4. FIGURE 1.35 gives graphical and numeric summaries from JMP for Antho1. Use this output to write a summary of the distribution of Antho1 using the methods and ideas that you learned in this chapter.

1.123 Blueberries and anthocyanins, Antho2. Refer to the previous exercise. Generate your own output for the analysis of Antho2 and use your output to write a summary of the distribution of Antho2 using the methods and ideas that you learned in this chapter.

1.124 Blueberries and anthocyanins, Antho3. Refer to Exercise 1.122. FIGURE 1.36 gives the JMP output for Antho3. Use this output to write a summary of the distribution of Antho3 using the methods and ideas that you learned in this chapter.

1.25 Blueberries and anthocyanins, Antho4. Refer to Exercise 1.122. Generate your own output for the analysis of Antho4 and use your output to write a summary of the distribution of Antho4 using the methods and ideas that you learned in this chapter.

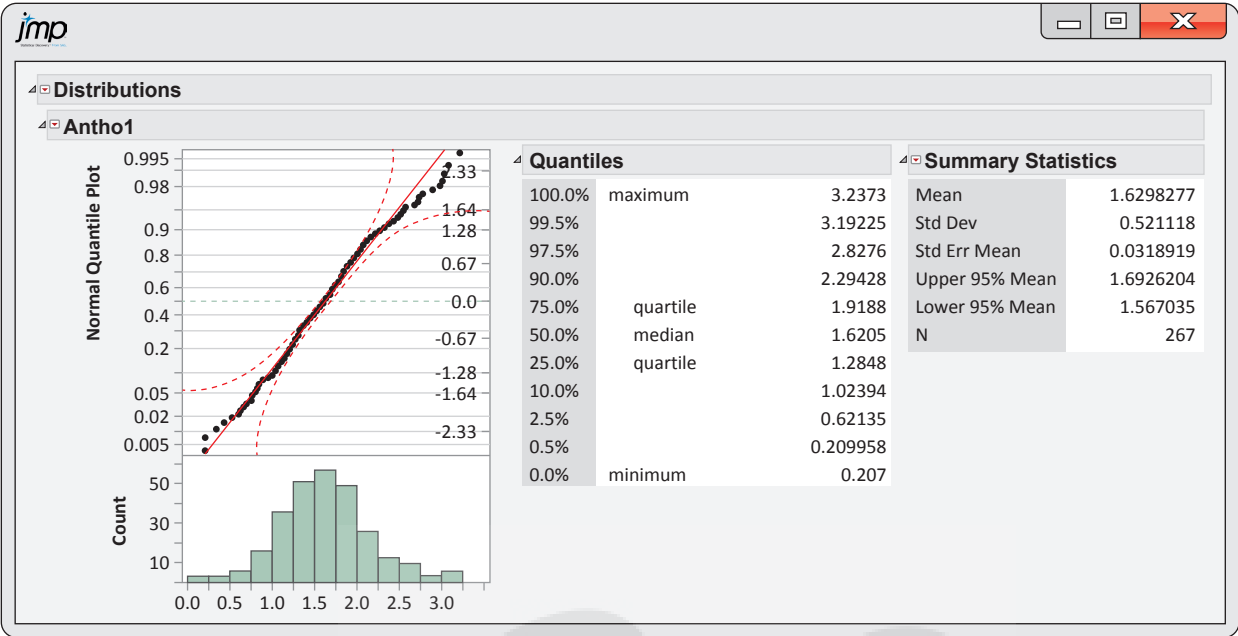


FIGURE 1.35 JMP descriptive statistics for Antho1, Exercise 1.122.

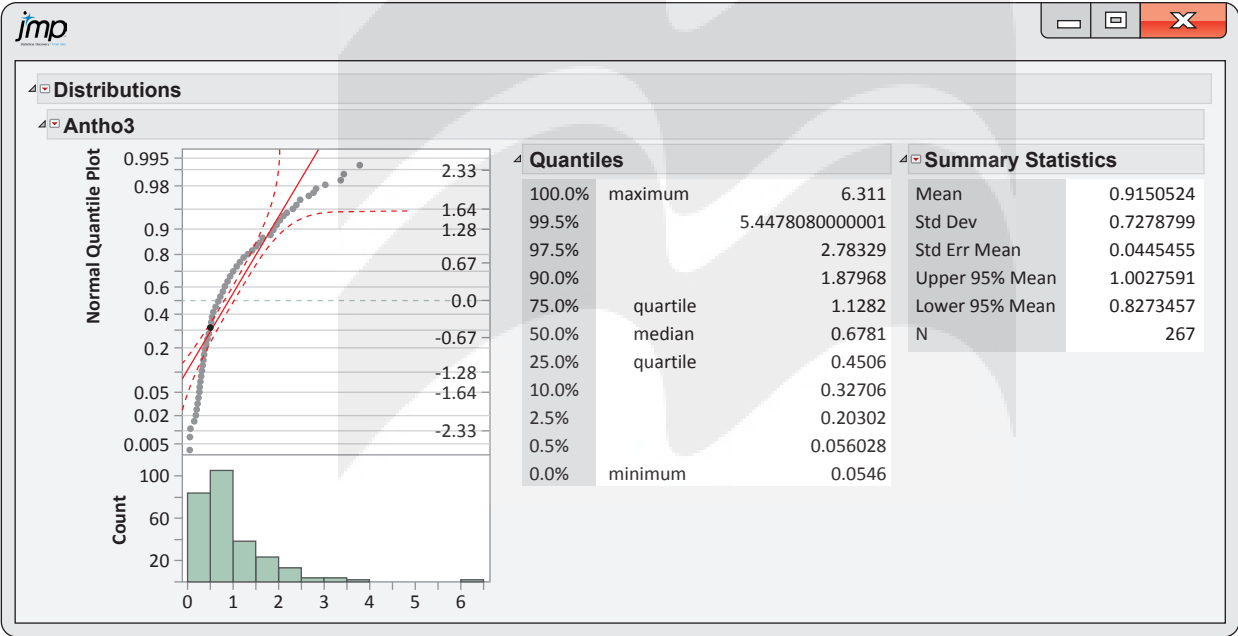


FIGURE 1.36 JMP descriptive statistics for Antho3, Exercise 1.124.