Chapter

2

Description of Samples and Populations

Objectives

In this chapter we will study how to describe data. In particular, we will

- show how frequency distributions are used to make bar charts and histograms.
- compare the mean and median as measures of center.
- demonstrate how to construct and read a variety of graphics including dotplots, boxplots, and scatterplots.
- compare several measures of variability with emphasis on the standard deviation.
- examine how transformations of variables affect distributions.
- consider the relationship between populations and samples.

2.1 Introduction

Statistics is the science of analyzing and learning from data. In this section we introduce some terminology and notation for dealing with data.

Variables

We begin with the concept of a **variable**. A variable is a characteristic of a person or a thing that can be assigned a number or a category. For example, blood type (A, B, AB, O) and age are two variables we might measure on a person.

Blood type is an example of a **categorical variable**: A categorical variable is a variable that records which of several categories a person or thing is in. Examples of categorical variables are

Blood type of a person: A, B, AB, O

Sex of a fish: male, female

Color of a flower: red, pink, white

Shape of a seed: wrinkled, smooth

For some categorical variables, the categories can be arrayed in a meaningful rank order. Such a variable is said to be **ordinal**. For example, the response of a patient to therapy might be none, partial, or complete.

Age is an example of a **numeric variable**. A numeric variable is a variable that records the amount of something. A **continuous variable** is a numeric variable that is measured on a continuous scale. Examples of continuous variables are

Weight of a baby
Cholesterol concentration in a blood specimen
Optical density of a solution

A variable such as weight is continuous because, in principle, two weights can be arbitrarily close together. Some types of numeric variables are not continuous but fall on a discrete scale, with spaces between the possible values. A **discrete variable** is a numeric variable for which we can list the possible values. For example, the number of eggs in a bird's nest is a discrete variable because only the values 0, 1, 2, 3, . . . , are possible. Other examples of discrete variables are

Number of bacteria colonies in a petri dish
Number of cancerous lymph nodes detected in a patient
Length of a DNA segment in basepairs

The distinction between continuous and discrete variables is not a rigid one. After all, physical measurements are always rounded off. We may measure the weight of a steer to the nearest kilogram, of a rat to the nearest gram, or of an insect to the nearest milligram. The scale of the actual measurements is always discrete, strictly speaking. The continuous scale can be thought of as an approximation to the actual scale of measurement.

Observational Units

When we collect a sample of n persons or things and measure one or more variables on them, we call these persons or things **observational units** or cases. The following are some examples of samples.

Sample	Variable	Observational unit
150 babies born in a certain hospital	Birthweight (kg)	A baby
73 *Cecropia* moths caught in a trap	Sex	A moth
81 plants that are a progeny of a single parental cross	Flower color	A plant
Bacterial colonies in each of six petri dishes	Number of colonies	A petri dish

Notation for Variables and Observations

We will adopt a notational convention to distinguish between a variable and an observed value of that variable. We will denote variables by uppercase letters such as Y. We will denote the observations themselves (that is, the data) by lowercase letters such as y. Thus, we distinguish, for example, between Y = birthweight (the variable) and $y = 7.9$ lb (the observation). This distinction will be helpful in explaining some fundamental ideas concerning variability.

Exercises 2.1.1–2.1.4

For each of the following settings in Exercises 2.1.1–2.1.4, (i) identify the variable(s) in the study, (ii) for each variable tell the type of variable (e.g., categorical and ordinal, discrete, etc.), (iii) identify the observational unit (the thing sampled), and (iv) determine the sample size.

2.1.1

(a) A paleontologist measured the width (in mm) of the last upper molar in 36 specimens of the extinct mammal *Acropithecus rigidus*.

(b) The birthweight, date of birth, and the mother's race were recorded for each of 65 babies.

2.1.2

(a) A physician measured the height and weight of each of 37 children.

(b) During a blood drive, a blood bank offered to check the cholesterol of anyone who donated blood. A total of 129 persons donated blood. For each of them, the blood type and cholesterol levels were recorded.

2.1.3

(a) A biologist measured the number of leaves on each of 25 plants.

(b) A physician recorded the number of seizures that each of 20 patients with severe epilepsy had during an eight-week period.

2.1.4

(a) A conservationist recorded the weather (clear, partly cloudy, cloudy, rainy) and number of cars parked at noon at a trailhead on each of 18 days.

(b) An enologist measured the pH and residual sugar content (g/l) of seven barrels of wine.

2.2 Frequency Distributions

A first step toward understanding a set of data on a given variable is to explore the data and describe the data in summary form. In this chapter we discuss three mutually complementary aspects of summary data description: frequency distributions, measures of center, and measures of dispersion. These tell us about the shape, center, and spread of the data.

A **frequency distribution** is simply a display of the **frequency**, or number of occurrences, of each value in the data set. The information can be presented in tabular form or, more vividly, with a graph. A **bar chart** is a simple graphic showing the categories that a categorical variable takes on and the number of observations in each category for the data in the sample. Here are two examples of frequency distributions for categorical data.

Example 2.2.1

Color of Poinsettias Poinsettias can be red, pink, or white. In one investigation of the hereditary mechanism controlling the color, 182 progeny of a certain parental cross were categorized by color.[1] The bar graph in Figure 2.2.1 is a visual display of the results given in Table 2.2.1. ■

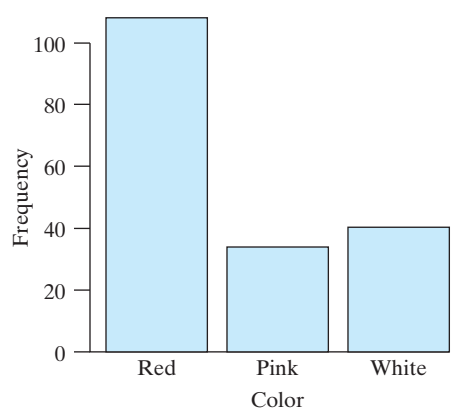


Figure 2.2.1 Bar chart of color of 182 poinsettias

Table 2.2.1 Color of one hundred eighty-two poinsettias

Color	Frequency (number of plants)
Red	108
Pink	34
White	40
Total	182

Example 2.2.2

School Bags and Neck Pain Physiologists in Australia were concerned that carrying a school bag loaded with heavy books was a cause of neck pain in adolescents, so they asked a sample of 585 teenage girls how often they get neck pain when carrying their school bag (e.g., never, almost never, sometimes, often, always). A summary of the results reported to them is given in Table 2.2.2 and displayed as a bar graph in Figure 2.2.2(a).[2] As the variable incidence is an ordinal categorical variable, our tables and graphs should respect the natural ordering. Figure 2.2.2(b) shows the same data but with the categories in alphabetical order (a default setting for much software), which obscures the information in the data. ■

Table 2.2.2 Neck pain associated with carrying a school bag

Incidence	Frequency (number of girls)
Never	179
Almost never	159
Sometimes	173
Often	64
Always	10
Total	585

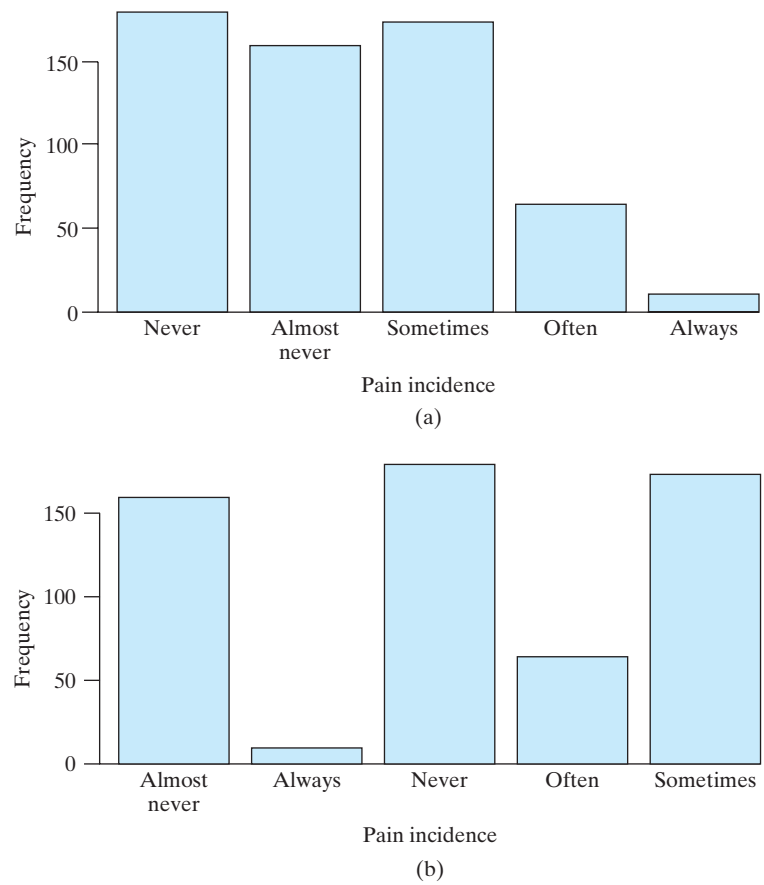


Figure 2.2.2 (a) Bar chart of incidence of neck pain reported by 585 adolescents; (b) the same data but with the categories in alphabetical order

A **dotplot** is a simple graph that can be used to show the distribution of a numeric variable when the sample size is small. To make a dotplot, we draw a number line covering the range of the data and then put a dot above the number line for each observation, as the following example shows.

Example 2.2.3 Infant Mortality Table 2.2.3 shows the infant mortality rate (infant deaths per 1,000 live births) in each of 12 countries in South America, as of 2009.[3] The distribution is shown in Figure 2.2.3. ■

Table 2.2.3 Infant mortality in 12 South American countries

Country	Infant mortality rate
Argentina	11.4
Bolivia	44.7
Brazil	22.6
Chile	7.7
Colombia	18.9
Ecuador	20.9
Guyana	30.0
Paraguay	24.7
Peru	28.6
Suriname	18.8
Uruguay	11.3
Venezuela	26.5

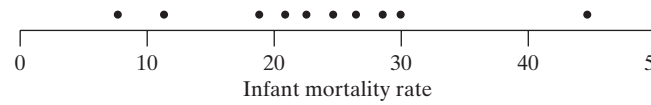


Figure 2.2.3 Dotplot of infant mortality in 12 South American countries

When two or more observations take on the same value, we stack the dots in a dotplot on top of each other. This gives an effect similar to the effect of the bars in a bar chart. If we create bars, in place of the stacks of dots, we then have a **histogram**. A histogram is like a bar chart, except that a histogram displays a numeric variable, which means that there is a natural order and scale for the variable. In a bar chart the amount of space between the bars (if any) is arbitrary, since the data being displayed are categorical. In a histogram the scale of the variable determines the placement of the bars. The following example shows a dotplot and a histogram for a frequency distribution.

Example 2.2.4 Litter Size of Sows A group of thirty-six 2-year-old sows of the same breed ($\frac{3}{4}$ Duroc, $\frac{1}{4}$ Yorkshire) were bred to Yorkshire boars. The number of piglets surviving to 21 days of age was recorded for each sow.[4] The results are given in Table 2.2.4 and displayed as a dotplot in Figure 2.2.4 and as a histogram in Figure 2.2.5. ■

Table 2.2.4 Number of surviving piglets of 36 sows

Number of piglets	Frequency (number of sows)
5	1
6	0
7	2
8	3
9	3
10	9
11	8
12	5
13	3
14	2
Total	36

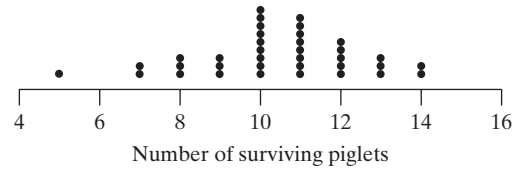


Figure 2.2.4 Dotplot of number of surviving piglets of 36 sows

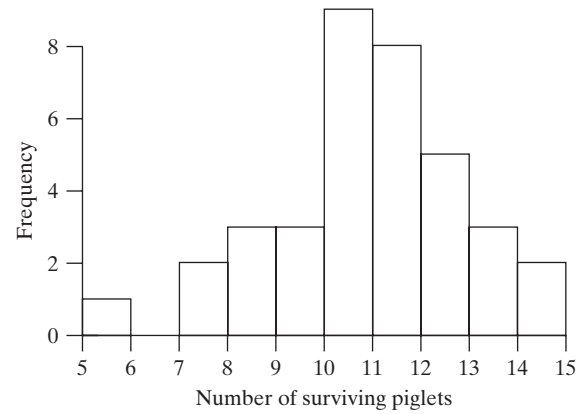


Figure 2.2.5 Histogram of number of surviving piglets of 36 sows

Relative Frequency

The frequency scale is often replaced by a **relative frequency** scale:

$$\text{Relative frequency} = \frac{\text{Frequency}}{n}$$

The relative frequency scale is useful if several data sets of different sizes (n's) are to be displayed together for comparison. As another option, a relative frequency can be expressed as a percentage frequency. The shape of the display is not affected by the choice of frequency scale, as the following example shows.

Example 2.2.5

Color of Poinsettias The poinsettia color distribution of Example 2.2.1 is expressed as frequency, relative frequency, and percent frequency in Table 2.2.5 and Figure 2.2.6. ■

Table 2.2.5 Color of one hundred eighty-two poinsettias

Color	Frequency	Relative frequency	Percent frequency
Red	108	.59	59
Pink	34	.19	19
White	40	.22	22
Total	182	1.00	100

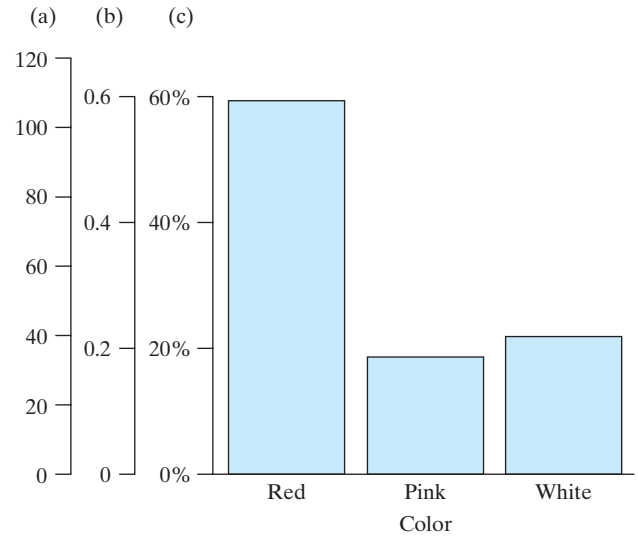


Figure 2.2.6 Bar chart of poinsettia colors on three scales:
(a) Frequency
(b) Relative frequency
(c) Percent frequency

Grouped Frequency Distributions

In the preceding examples, simple ungrouped frequency distributions provided concise summaries of the data. For many data sets, it is necessary to group the data in order to condense the information adequately. (This is usually the case with continuous variables.) The following example shows a grouped frequency distribution.

Example 2.2.6

Serum CK Creatine phosphokinase (CK) is an enzyme related to muscle and brain function. As part of a study to determine the natural variation in CK concentration, blood was drawn from 36 male volunteers. Their serum concentrations of CK (measured in U/l) are given in Table 2.2.6.[5] Table 2.2.7 shows these data grouped into **classes**. For instance, the frequency of the class [20,40) (all values in the interval $20 \leq y < 40$) is 1, which means that one CK value fell in this range. The grouped frequency distribution is displayed as a histogram in Figure 2.2.7. ■

Table 2.2.6 Serum CK values for 36 men

121	82	100	151	68	58
95	145	64	201	101	163
84	57	139	60	78	94
119	104	110	113	118	203
62	83	67	93	92	110
25	123	70	48	95	42

Table 2.2.7 Frequency distribution of serum CK values for 36 men

Serum CK (U/l)	Frequency (number of men)
[20,40)	1
[40,60)	4
[60,80)	7
[80,100)	8
[100,120)	8
[120,140)	3
[140,160)	2
[160,180)	1
[180,200)	0
[200,220)	2
Total	36

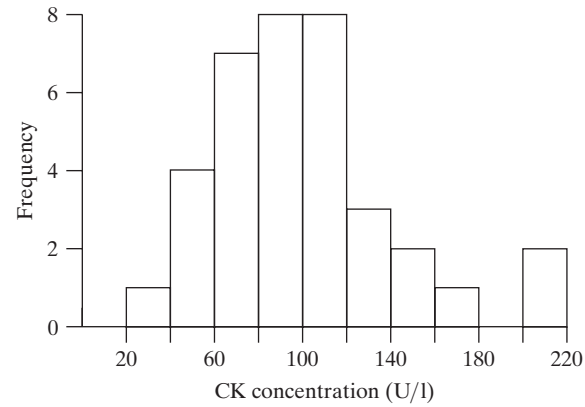


Figure 2.2.7 Histogram of serum CK concentrations for 36 men

A grouped frequency distribution should display the essential features of the data. For instance, the histogram of Figure 2.2.7 shows that the average CK value is about 100 U/l, with the majority of the values falling between 60 and 140 U/l. In addition, the histogram shows the *shape* of the distribution. Note that the CK values are piled up around a central peak, or **mode**. On either side of this mode, the frequencies decline and ultimately form the **tails** of the distribution. These shape features are labeled in Figure 2.2.8. The CK distribution is not symmetric but is a bit **skewed to the right**, which means that the right tail is more stretched out than the left.*

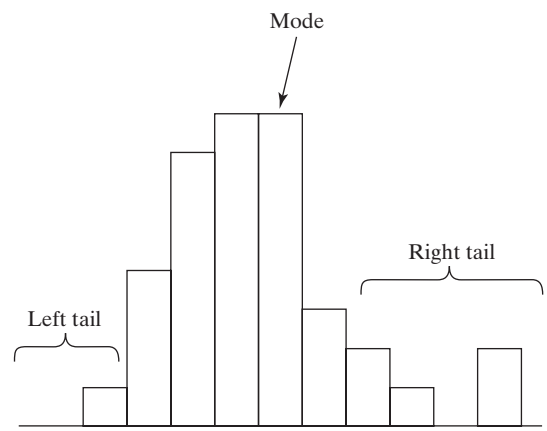


Figure 2.2.8 Shape features of the CK distribution

When making a histogram, we need to decide how many classes to have and how wide the classes should be. If we use computer software to generate a histogram, the program will choose the number of classes and the class width for us, but most software allows the user to change the number of classes and to specify the class width. If a data set is large and is quite spread out, it is a good idea to look at more than one histogram of the data, as is done in Example 2.2.7.

Example 2.2.7

Heights of Students A sample of 510 college students were asked how tall they were. Note that they were not measured; rather, they just reported their heights.[6] Figure 2.2.9 shows the distribution of the self-reported values, using 7 classes and a

*To help remember which tail of a skewed distribution is the longer tail, think of skew as stretch. Which side of the distribution is more stretched away from the center? A distribution that is skewed to the right is one in which the right tail stretches out more than the left.

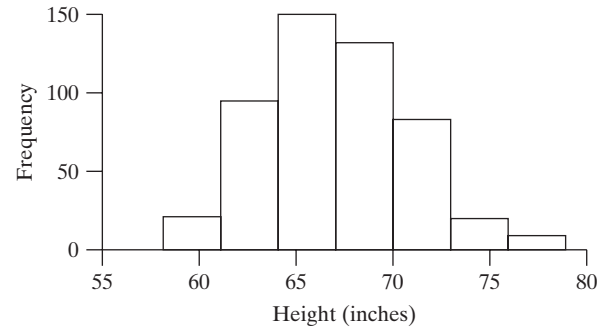


Figure 2.2.9 Heights of students, using 7 classes (class width = 3)

class width of 3 (inches). By using only 7 classes, the distribution appears to be reasonably symmetric, with a single peak around 66 inches.

Figure 2.2.10 shows the height data, but in a histogram that uses 18 classes and a class width of 1.1. This view of the data shows two modes—one for women and one for men.

Figure 2.2.11 shows the height data again, this time using 37 classes, each of width 0.5. Using such a large number of classes makes the distribution look jagged. In this case, we see an alternating pattern between classes with lots of observations and classes with few observations. In the middle of the distribution we see that there were many students who reported a height of 63 inches, few who reported a height of 63.5 inches, many who reported a height of 64 inches, and so on. It seems that most students round off to the nearest inch! ■

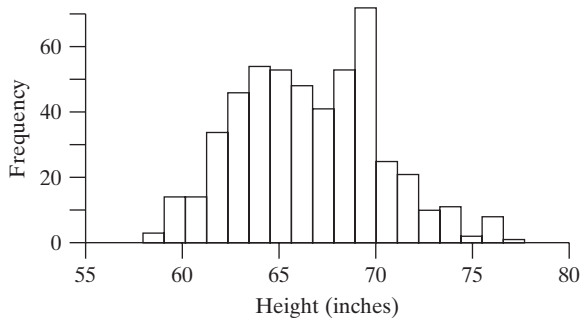


Figure 2.2.10 Heights of students, using 18 classes (class width = 1.1)

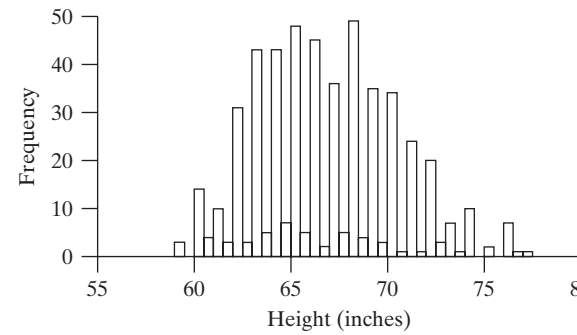


Figure 2.2.11 Heights of students, using 37 classes (class width = 0.5)

Interpreting Areas in a Histogram

A histogram can be looked at in two ways. The tops of the bars sketch out the shape of the distribution. But the *areas* within the bars also have a meaning. The area of each bar is proportional to the corresponding frequency. Consequently, the area of one or several bars can be interpreted as expressing the number of observations in the classes represented by the bars. For example, Figure 2.2.12 shows a histogram of the CK distribution of Example 2.2.6. The shaded area is 42% of the total area in all the bars. Accordingly, 42% of the CK values are in the corresponding classes; that is, 15 of 36 or 42% of the values are between 60 U/I and 100 U/l.*

*Strictly speaking, between 60 U/l and 99 U/l, inclusive.

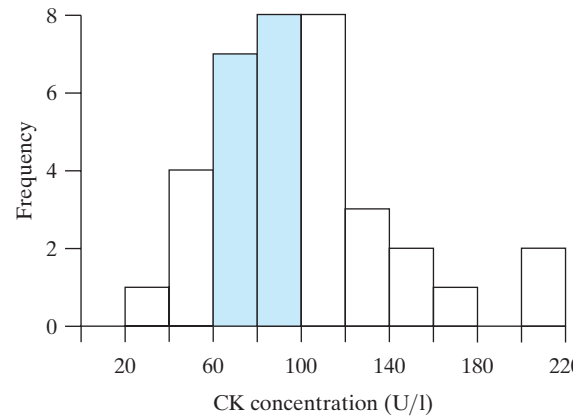


Figure 2.2.12 Histogram of CK distribution. The shaded area is 42% of the total area and represents 42% of the observations.

The area interpretation of histograms is a simple but important idea. In our later work with distributions we will find the idea to be indispensable.

Shapes of Distributions

When discussing a set of data, we want to describe the shape, center, and spread of the distribution. In this section we concentrate on the shapes of frequency distributions and illustrate some of the diversity of distributions encountered in the life sciences. The shape of a distribution can be indicated by a smooth curve that approximates the histogram, as shown in Figure 2.2.13.

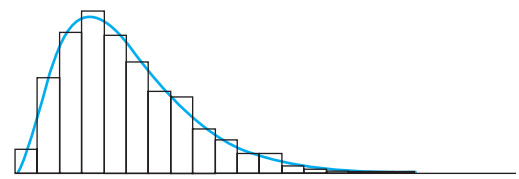


Figure 2.2.13 Approximation of a histogram by a smooth curve

Some distributional shapes are shown in Figure 2.2.14. A common shape for biological data is **unimodal** (has one mode) and is somewhat skewed to the right, as in (c). Approximately bell-shaped distributions, as in (a), also occur. Sometimes a distribution is symmetric but differs from a bell in having long tails; an exaggerated version is shown in (b). Left-skewed (d) and exponential (e) shapes are less common. **Bimodality** (two modes), as in (f), can indicate the existence of two distinct subgroups of observational units.

Notice that the shape characteristics we are emphasizing, such as number of modes and degree of symmetry, are *scale free;* that is, they are not affected by the arbitrary choices of vertical and horizontal scale in plotting the distribution. By contrast, a characteristic such as whether the distribution appears short and fat, or tall and skinny, is affected by how the distribution is plotted and so is not an inherent feature of the biological variable.

The following three examples illustrate biological frequency distributions with various shapes. In the first example, the shape provides evidence that the distribution is in fact biological rather than nonbiological.

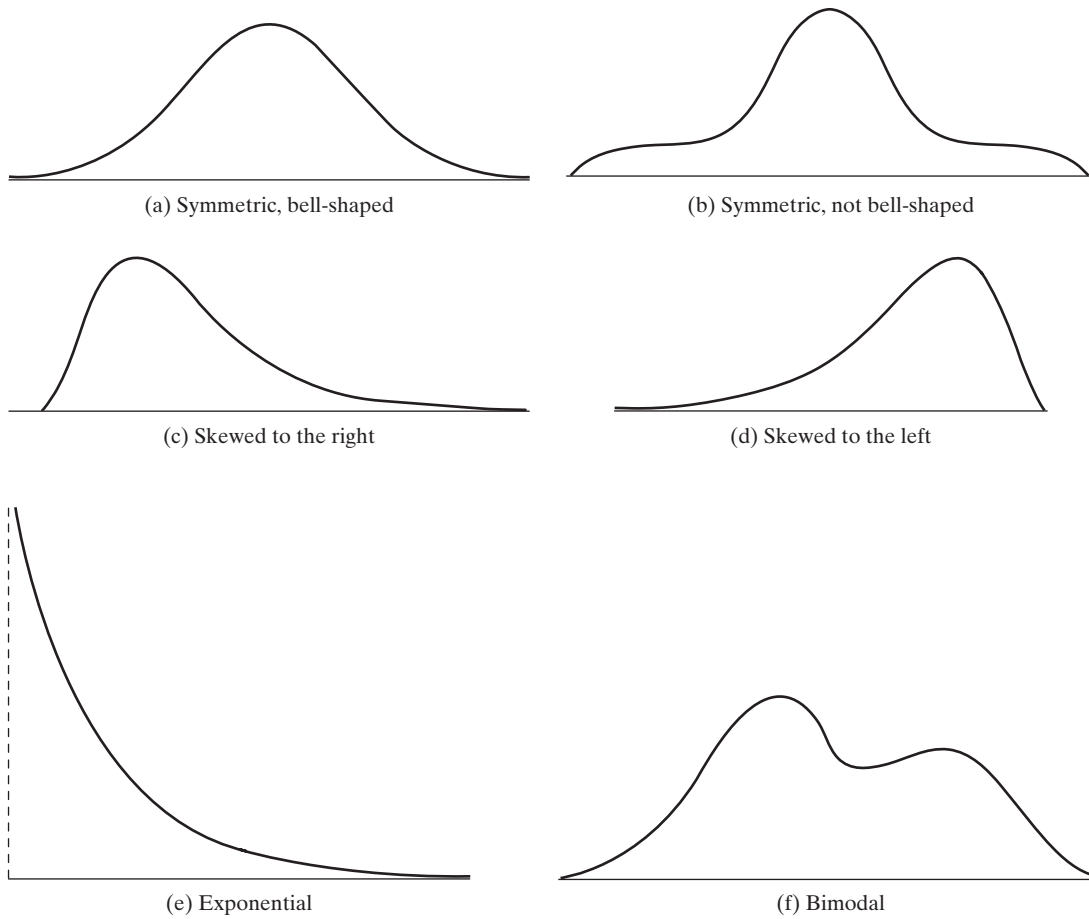
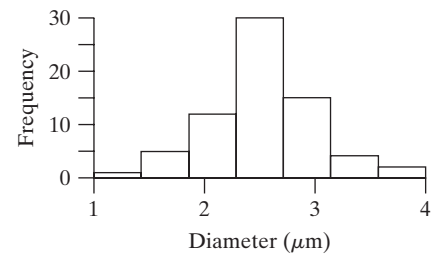


Figure 2.2.14 Shapes of distributions

Example 2.2.8

Microfossils In 1977, paleontologists discovered microscopic fossil structures, resembling algae, in rocks 3.5 billion years old. A central question was whether these structures were biological in origin. One line of argument focused on their size distribution, which is shown in Figure 2.2.15. This distribution, with its unimodal and rather symmetric shape, resembles that of known microbial populations, but not that of known nonbiological structures.[7] ■

Figure 2.2.15 Sizes of microfossils



Example 2.2.9

Cell Firing Times A neurobiologist observed discharges from rat muscle cells grown in culture together with nerve cells. The time intervals between 308 successive discharges were distributed as shown in Figure 2.2.16. Note the exponential shape of the distribution.[8]

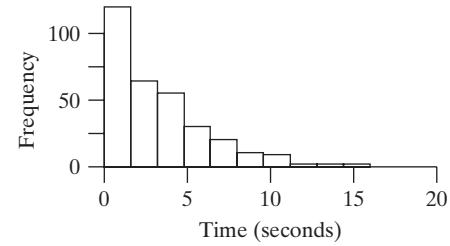


Figure 2.2.16 Time intervals between electrical discharges in rat muscle cells

Example 2.2.10

Brain Weight In 1888, P. Topinard published data on the brain weights of hundreds of French men and women. The data for males and females are shown in Figure 2.2.17(a) and (b). The male distribution is fairly symmetric and bell shaped; the female distribution is somewhat skewed to the right. Part (c) of the figure shows the brain weight distribution for males and females combined. This combined distribution is slightly bimodal.[9]

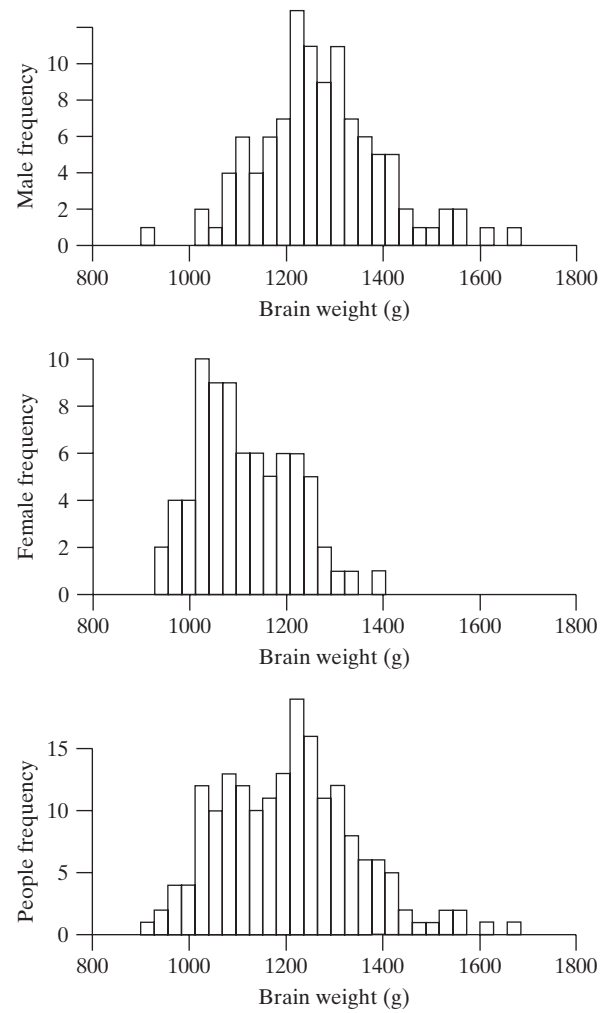


Figure 2.2.17 Brain weights

Sources of Variation

In interpreting biological data, it is helpful to be aware of sources of variability. The variation among observations in a data set often reflects the combined effects of several underlying factors. The following two examples illustrate such situations.

Example 2.2.11

Weights of Seeds In a classic experiment to distinguish environmental from genetic influence, a geneticist weighed seeds of the princess bean *Phaseolus vulgaris*. Figure 2.2.18 shows the weight distributions of (a) 5,494 seeds from a commercial seed lot, and (b) 712 seeds from a highly inbred line that was derived from a single seed from the original lot. The variability in (a) is due to both environmental and genetic factors; in (b), because the plants are nearly genetically identical, the variation in weights is due largely to environmental influence.[10] Thus, there is less variability in the inbred line. ■

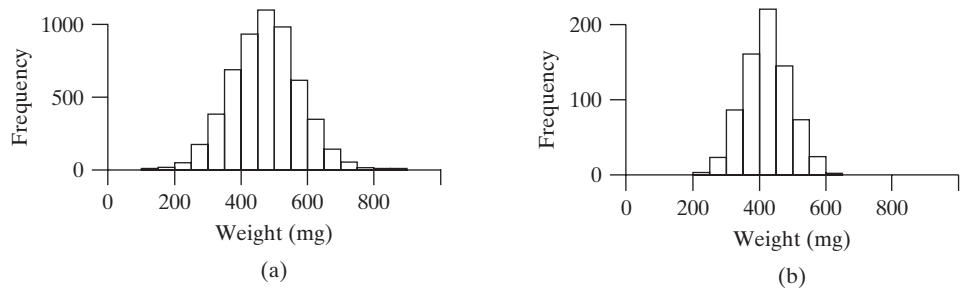


Figure 2.2.18 Weights of princess bean seeds: (a) from an open-bred population; (b) from an inbred line

Example 2.2.12

Serum ALT Alanine aminotransferase (ALT) is an enzyme found in most human tissues. Part (a) of Figure 2.2.19 shows the serum ALT concentrations for 129 adult volunteers. The following are potential sources of variability among the measurements:

1. Interindividual
   (a) Genetic
   (b) Environmental
2. Intraindividual
   (a) Biological: changes over time
   (b) Analytical: imprecision in assay

The effect of the last source—analytical variation—can be seen in part (b) of Figure 2.2.19, which shows the frequency distribution of 109 assays of the *same* specimen of serum; the figure shows that the ALT assay is fairly imprecise.[11] ■

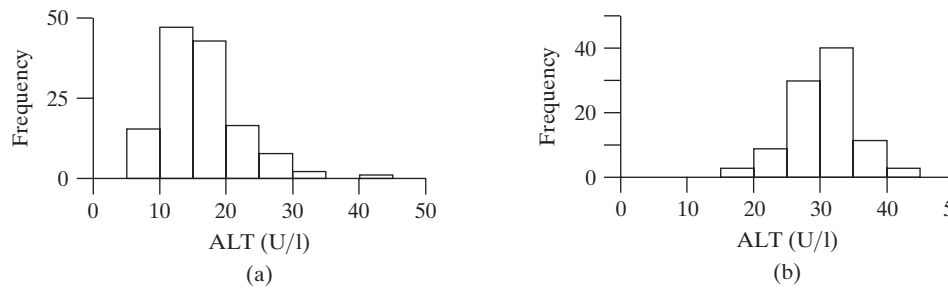


Figure 2.2.19 Distribution of serum ALT measurements (a) for 129 volunteers; (b) for 109 assays of the same specimen

Exercises 2.2.1–2.2.9

2.2.1 A paleontologist measured the width (in mm) of the last upper molar in 36 specimens of the extinct mammal *Acropithecus rigidus*. The results were as follows:[12]

6.1	5.7	6.0	6.5	6.0	5.7
6.1	5.8	5.9	6.1	6.2	6.0
6.3	6.2	6.1	6.2	6.0	5.7
6.2	5.8	5.7	6.3	6.2	5.7
6.2	6.1	5.9	6.5	5.4	6.7
5.9	6.1	5.9	5.9	6.1	6.1

(a) Construct a frequency distribution and display it as a table and as a histogram.

(b) Describe the shape of the distribution.

2.2.2 In a study of schizophrenia, researchers measured the activity of the enzyme monoamine oxidase (MAO) in the blood platelets of 18 patients. The results (expressed as nmoles benzylaldehyde product per 108 platelets) were as follows:[13]

6.8	8.4	8.7	11.9	14.2	18.8
9.9	4.1	9.7	12.7	5.2	7.8
7.8	7.4	7.3	10.6	14.5	10.7

Construct a dotplot of the data.

2.2.3 Consider the data presented in Exercise 2.2.2. Construct a frequency distribution and display it as a table and as a histogram.

2.2.4 A dendritic tree is a branched structure that emanates from the body of a nerve cell. As part of a study of brain development, 36 nerve cells were taken from the brains of newborn guinea pigs. The investigators counted the number of dendritic branch segments emanating from each nerve cell. The numbers were as follows:[14]

23	30	54	28	31	29	34	35	30
27	21	43	51	35	51	49	35	24
26	29	21	29	37	27	28	33	33
23	37	27	40	48	41	20	30	57

Construct a dotplot of the data.

2.2.5 Consider the data presented in Exercise 2.2.4. Construct a frequency distribution and display it as a table and as a histogram.

2.2.6 The total amount of protein produced by a dairy cow can be estimated from periodic testing of her milk. The following are the total annual protein production values (lb) for twenty-eight 2-year-old Holstein cows. Diet, milking procedures, and other conditions were the same for all the animals.[15]

425	481	477	434	410	397	438
545	528	496	502	529	500	465
539	408	513	496	477	445	546
471	495	445	565	499	508	426

Construct a frequency distribution and display it as a table and as a histogram.

2.2.7 For each of 31 healthy dogs, a veterinarian measured the glucose concentration in the anterior chamber of the right eye and also in the blood serum. The following data are the anterior chamber glucose measurements, expressed as a percentage of the blood glucose.[16]

81	85	93	93	99	76	75	84
78	84	81	82	89	81	96	82
74	70	84	86	80	70	131	75
88	102	115	89	82	79	106	

Construct a frequency distribution and display it as a table and as a histogram.

2.2.8 Agronomists measured the yield of a variety of hybrid corn in 16 locations in Illinois. The data, in bushels per acre, were[17]

241	230	207	219	266	167
204	144	178	158	153	
187	181	196	149	183	

(a) Construct a dotplot of the data.

(b) Describe the shape of the distribution.

2.2.9 (Computer problem) Trypanosomes are parasites that cause disease in humans and animals. In an early study of trypanosome morphology, researchers measured the lengths of 500 individual trypanosomes taken from the blood of a rat. The results are summarized in the accompanying frequency distribution.[18]

LENGTH (μm)	FREQUENCY (NUMBER OF INDIVIDUALS)	LENGTH (μm)	FREQUENCY (NUMBER OF INDIVIDUALS)
15	1	27	36
16	3	28	41
17	21	29	48
18	27	30	28
19	23	31	43
20	15	32	27
21	10	33	23
22	15	34	10
23	19	35	4
24	21	36	5
25	34	37	1
26	44	38	1

(a) Construct a histogram of the data using 24 classes (i.e., one class for each integer length, from 15 to 38).

(b) What feature of the histogram suggests the interpretation that the 500 individuals are a mixture of two distinct types?

(c) Construct a histogram of the data using only 6 classes. Discuss how this histogram gives a qualitatively different impression than the histogram from part (a).

2.3 Descriptive Statistics: Measures of Center

For categorical data, the frequency distribution provides a concise and complete summary of a sample. For numeric variables, the frequency distribution can usefully be supplemented by a few numerical measures. A numerical measure calculated from sample data is called a **statistic**.* **Descriptive statistics** are statistics that describe a set of data. Usually the descriptive statistics for a sample are calculated in order to provide information about a population of interest (see Section 2.8). In this section we discuss measures of the center of the data. There are several different ways to define the "center" or "typical value" of the observations in a sample. We will consider the two most widely used measures of center: the median and the mean.

The Median

Perhaps the simplest measure of the center of a data set is the sample **median**. The sample median is the value that most nearly lies in the middle of the sample—it is the data value that splits the ordered data into two equal halves. To find the median, first arrange the observations in increasing order. In the array of ordered observations, the median is the middle value (if n is odd) or midway between the two middle values (if n is even). We denote the median of the sample by the symbol $\tilde{y}$ (read "y-tilde"). Example 2.3.1 illustrates these definitions.

Example 2.3.1

Weight Gain of Lambs The following are the two-week weight gains (lb) of six young lambs of the same breed that had been raised on the same diet:[19]

$$11 \quad 13 \quad 19 \quad 2 \quad 10 \quad 1$$

The ordered observations are

$$1 \quad 2 \quad 10 \quad 11 \quad 13 \quad 19$$

The median weight gain is

$$\tilde{y} = \frac{10 + 11}{2} = 10.5 \text{ lb}$$

The median divides the sorted data into two equal pieces (the same number of observations fall above and below the median). Figure 2.3.1 shows a dotplot of the lamb weight-gain data, along with the location of $\tilde{y}$. ■

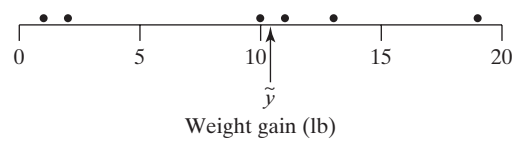


Figure 2.3.1 Plot of the lamb weight-gain data

*Numerical measures based on the entire population are called **parameters**, which are discussed in greater detail in Section 2.8.

Example 2.3.2 Weight Gain of Lambs Suppose the sample contained one more lamb, with the seven ranked observations as follows:

$$1 \;\; 2 \;\; 10 \;\; 10 \;\; 11 \;\; 13 \;\; 19$$

For this sample, the median weight gain is

$$\tilde{y} = 10 \text{ lb}$$

(Notice that in this example there are two lambs whose weight gain is equal to the median. The fourth observation—the second 10—is the median.) ■

A more formal way to define the median is in terms of rank position in the ordered array (counting the smallest observation as rank 1, the next as 2, and so on). The rank position of the median is equal to

$$(0.5)(n + 1)$$

Thus, if $n = 7$, we calculate $(0.5)(n + 1) = 4$, so that the median is the fourth largest observation; if $n = 6$, we have $(0.5)(n + 1) = 3.5$, so that the median is midway between the third and fourth largest observations. Note that the formula $(0.5)(n + 1)$ does not give the median, it gives the location of the median within the ordered list of the data.

The Mean

The most familiar measure of center is the ordinary average or **mean** (sometimes called the arithmetic mean). The mean of a sample (or "the sample mean") is the sum of the observations divided by the number of observations. If we denote a variable by Y, then we denote the observations in a sample by $y_1, y_2, \ldots, y_n$ and we denote the mean of the sample by the symbol $\bar{y}$ (read "y-bar"). Example 2.3.3 illustrates this notation.

Example 2.3.3 Weight Gain of Lambs The following are the data from Example 2.3.1:

$$11 \;\; 13 \;\; 19 \;\; 2 \;\; 10 \;\; 1$$

Here $y_1 = 11$, $y_2 = 13$, and so on, and $y_6 = 1$. The sum of the observations is $11 + 13 + \ldots + 1 = 56$. We can write this using "summation notation" as $\sum_{i=1}^{n} y_i = 56$. The symbol $\sum_{i=1}^{n} y_i$ means to "add up the y_i's." Thus, when $n = 6$, $\sum_{i=1}^{n} y_i = y_1 + y_2 + y_3 + y_4 + y_5 + y_6$. In this case we get $\sum_{i=1}^{n} y_i = 11 + 13 + 19 + 2 + 10 + 1 = 56$.

The mean weight gain of the six lambs in this sample is

$$\begin{aligned} \bar{y} &= \frac{11 + 13 + 19 + 2 + 10 + 1}{6} \\ &= \frac{56}{6} \\ &= 9.33 \text{ lb} \end{aligned}$$

The Sample Mean The general definition of the sample mean is

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

where the y_i's are the observations in the sample and n is the sample size (that is, the number of y_i's).

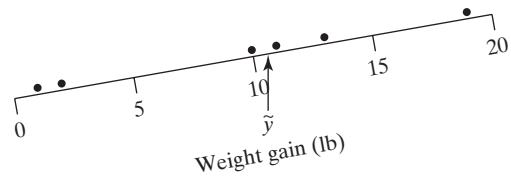


Figure 2.3.2 Plot of the lamb weight-gain data with the sample median as the fulcrum of a balance

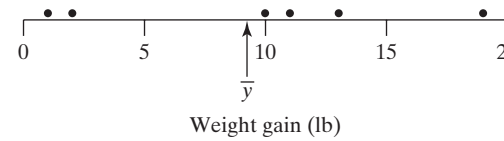


Figure 2.3.3 Plot of the lamb weight-gain data with the sample mean as the fulcrum of a balance

While the median divides the data into two equal pieces (i.e., the same number of observations above and below), the mean is the "point of balance" of the data. Figure 2.3.2 shows a dotplot of the lamb weight-gain data, along with the location of $\tilde{y}$. If the data points were children on a weightless seesaw, then the seesaw would tip if the fulcrum were placed at $\tilde{y}$ despite there being the same number of children on either side. The children on the left side (below $\tilde{y}$) tend to sit further from $\tilde{y}$ than the children on the right (above $\tilde{y}$) causing the seesaw to tip. However, if the fulcrum were placed at $\bar{y}$, the seesaw would exactly balance as in Figure 2.3.3. ■

The difference between a data point and the mean is called a **deviation**: $\text{deviation}_i = y_i - \bar{y}$. The mean has the property that the sum of the deviations from the mean is zero—that is, $\sum_{i=1}^{n}(y_i - \bar{y}) = 0$. In this sense, the mean is a center of the distribution—the positive deviations balance the negative deviations.

Example 2.3.4

Weight Gain of Lambs For the lamb weight-gain data, the deviations are as follows:

$$\begin{aligned}
\text{deviation}_1 &= y_1 - \bar{y} = 11 - 9.33 = 1.67\\
\text{deviation}_2 &= y_2 - \bar{y} = 13 - 9.33 = 3.67\\
\text{deviation}_3 &= y_3 - \bar{y} = 19 - 9.33 = 9.67\\
\text{deviation}_4 &= y_4 - \bar{y} = 2 - 9.33 = -7.33\\
\text{deviation}_5 &= y_5 - \bar{y} = 10 - 9.33 = 0.67\\
\text{deviation}_6 &= y_6 - \bar{y} = 1 - 9.33 = -8.33
\end{aligned}$$

The sum of the deviations is $\sum_{i=1}^{n}(y_i - \bar{y}) = 1.67 + 3.67 + 9.67 - 7.33 + 0.67 - 8.33 = 0$. ■

Robustance A statistic is said to be **robust** or **resistant** if the value of the statistic is relatively unaffected by changes in a small portion of the data, even if the changes are dramatic ones. The median is a robust statistic, but the mean is not robust because it can be greatly shifted by changes in even one observation. Example 2.3.5 illustrates this behavior.

Example 2.3.5

Weight Gain of Lambs Recall that for the lamb weight-gain data

$$1\ \ 2\ \ 10\ \ 11\ \ 13\ \ 19$$

we found

$$\bar{y} = 9.3 \text{ and } \tilde{y} = 10.5$$

Suppose now that the observation 19 is changed, or even omitted. How would the mean and median be affected? You can visualize the effect by imagining moving or removing the right-hand dot in Figure 2.3.3. Clearly the mean could change a great deal; the median would generally be less affected. For instance,

If the 19 is changed to 12, the mean becomes 8.2 and the median does not change.

If the 19 is omitted, the mean becomes 7.4 and the median becomes 10.

These changes are not wild ones; that is, the changed samples might well have arisen from the same feeding experiment. Of course, a huge change, such as changing the 19 to 100, would shift the mean very drastically. Note that it would not shift the median at all. ■

Visualizing the Mean and Median

We can visualize the mean and the median in relation to the histogram of a distribution. The median divides the area under the histogram roughly in half because it divides the observations roughly in half ["roughly" because some observations may be tied at the median, as in Example 2.3.3(b), and because the observations within each class are not uniformly distributed across the class]. The mean can be visualized as the point of balance of the histogram: If the histogram were made out of plywood, it would balance if supported at the mean.

If the frequency distribution is symmetric, the mean and the median are equal and fall in the center of the distribution. If the frequency distribution is skewed, both measures are pulled toward the longer tail, but the mean is usually pulled farther than the median. The effect of skewness is illustrated by the following example.

Example 2.3.6

Cricket Singing Times Male Mormon crickets *(Anabrus simplex)* sing to attract mates. A field researcher measured the duration of 51 unsuccessful songs—that is, the time until the singing male gave up and left his perch.[20] Figure 2.3.4 shows the histogram of the 51 singing times. Table 2.3.1 gives the raw data. The median is 3.7 min and the mean is 4.3 min. The discrepancy between these measures is due largely to the long straggly tail of the distribution; the few unusually long singing times influence the mean, but not the median. ■

Table 2.3.1 Fifty-one cricket singing times (min)

4.3	3.9	17.4	2.3	0.8	1.5	0.7	3.7
24.1	9.4	5.6	3.7	5.2	3.9	4.2	3.5
6.6	6.2	2.0	0.8	2.0	3.7	4.7	
7.3	1.6	3.8	0.5	0.7	4.5	2.2	
4.0	6.5	1.2	4.5	1.7	1.8	1.4	
2.6	0.2	0.7	11.5	5.0	1.2	14.1	
4.0	2.7	1.6	3.5	2.8	0.7	8.6	

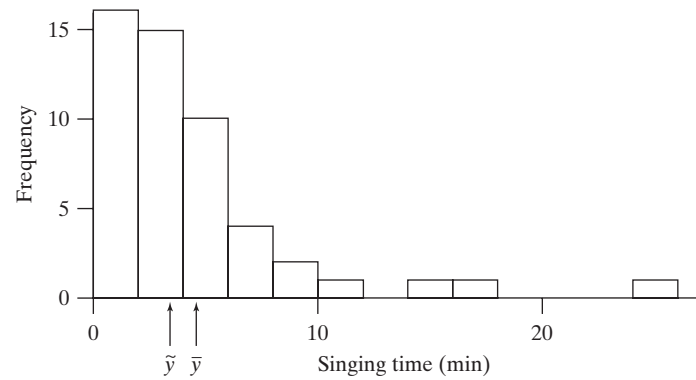


Figure 2.3.4 Histogram of cricket singing times

Mean versus Median

Both the mean and the median are usually reasonable measures of the center of a data set. The mean is related to the sum; for example, if the mean weight gain of 100 lambs is 9 lb, then the total weight gain is 900 lb, and this total may be of primary interest since it translates more or less directly into profit for the farmer. In some

situations the mean makes very little sense. Suppose, for example, that the observations are survival times of cancer patients on a certain treatment protocol, and that most patients survive less than 1 year, while a few respond well and survive for 5 or even 10 years. In this case, the mean survival time might be greater than the survival time of most patients; the median would more nearly represent the experience of a "typical" patient. Note also that the mean survival time cannot be computed until the last patient has died; the median does not share this disadvantage. Situations in which the median can readily be computed, but the mean cannot, are not uncommon in bioassay, survival, and toxicity studies.

We have noted that the median is more resistant than the mean. If a data set contains a few observations rather distant from the main body of the data—that is, a long "straggly" tail—then the mean may be unduly influenced by these few unusual observations. Thus, the "tail" may "wag the dog"—an undesirable situation. In such cases, the resistance of the median may be advantageous.

An advantage of the mean is that in some circumstances it is more efficient than the median. Efficiency is a technical notion in statistical theory; roughly speaking, a method is efficient if it takes full advantage of all the information in the data. Partly because of its efficiency, the mean has played a major role in classical methods in statistics.

Exercises 2.3.1–2.3.16

2.3.1 Invent a sample of size 5 for which the sample mean is 20 and not all the observations are equal.

2.3.2 Invent a sample of size 5 for which the sample mean is 20 and the sample median is 15.

2.3.3 A researcher applied the carcinogenic (cancer-causing) compound benzo(a)pyrene to the skin of five mice, and measured the concentration in the liver tissue after 48 hours. The results (nmol/gm) were as follows:[21]

6.3 5.9 7.0 6.9 5.9

Determine the mean and the median.

2.3.4 Consider the data from Exercise 2.3.3. Do the calculated mean and median support the claim that, in general, liver tissue concentration after 48 hours differs from 6.3 nmol/gm?

2.3.5 Six men with high serum cholesterol participated in a study to evaluate the effects of diet on cholesterol level. At the beginning of the study their serum cholesterol levels (mg/dl) were as follows:[22]

366 327 274 292 274 230

Determine the mean and the median.

2.3.6 Consider the data from Exercise 2.3.5. Suppose an additional observation equal to 400 were added to the sample. What would be the mean and the median of the seven observations?

2.3.7 The weight gains of beef steers were measured over a 140-day test period. The average daily gains (lb/day) of 9 steers on the same diet were as follows:[23]

3.89 3.51 3.97 3.31 3.21
3.36 3.67 3.24 3.27

Determine the mean and median.

2.3.8 Consider the data from Exercise 2.3.7. Are the calculated mean and median consistent with the claim that, in general, steers gain 3.5 lb/day? Are they consistent with a claim of 4.0 lb/day?

2.3.9 Consider the data from Exercise 2.3.7. Suppose an additional observation equal to 2.46 were added to the sample. What would be the mean and the median of the 10 observations?

2.3.10 As part of a classic experiment on mutations, 10 aliquots of identical size were taken from the same culture of the bacterium *E. coli*. For each aliquot, the number of bacteria resistant to a certain virus was determined. The results were as follows:[24]

14 15 13 21 15
14 26 16 20 13

(a) Construct a frequency distribution of these data and display it as a histogram.

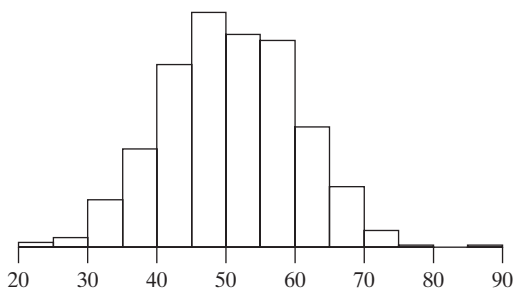
(b) Determine the mean and the median of the data and mark their locations on the histogram.

2.3.11 The accompanying table gives the litter size (number of piglets surviving to 21 days) for each of 36 sows (as in Example 2.2.4). Determine the median litter size. (*Hint*: Note that there is one 5, but there are two 7's, three 8's, etc.)

NUMBER OF PIGLETS	FREQUENCY (NUMBER OF SOWS)
5	1
6	0
7	2
8	3
9	3
10	9
11	8
12	5
13	3
14	2
Total	36

2.3.12 Consider the data from Exercise 2.3.11. Determine the mean of the 36 observations. (*Hint*: Note that there is one 5 but there are two 7's, three 8's, etc. Thus, $\Sigma y_i = 5 + 7 + 7 + 8 + 8 + 8 + \ldots = 5 + 2(7) + 3(8) + \ldots$)

2.3.13 Here is a histogram.

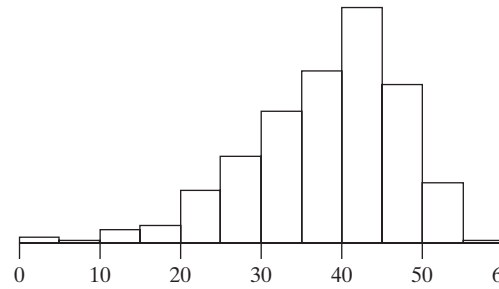


(a) Estimate the median of the distribution.

(b) Estimate the mean of the distribution.

2.3.14 Consider the histogram from Exercise 2.3.13. By "reading" the histogram, estimate the percentage of observations that are less than 40. Is this percentage closest to 15%, 25%, 35%, or 45%? (*Note*: The frequency scale is not given for this histogram, because there is no need to calculate the number of observations in each class. Rather, the percentage of observations that are less than 40 can be estimated by looking at area.)

2.3.15 Here is a histogram.



(a) Estimate the median of the distribution.

(b) Estimate the mean of the distribution.

2.3.16 Consider the histogram from Exercise 2.3.15. By "reading" the histogram, estimate the percentage of observations that are greater than 45. Is this percentage closest to 15%, 25%, 35%, or 45%? (*Note*: The frequency scale is not given for this histogram, because there is no need to calculate the number of observations in each class. Rather, the percentage of observations that are greater than 45 can be estimated by looking at area.)

2.4 Boxplots

One of the most efficient graphics, both for examining a single distribution and for making comparisons between distributions, is known as a boxplot, which is the topic of this section. Before discussing boxplots, however, we need to discuss quartiles.

Quartiles and the Interquartile Range

The median of a distribution splits the distribution into two parts, a lower part and an upper part. The **quartiles** of a distribution divide each of these parts in half, thereby dividing the distribution into four quarters. The **first quartile**, denoted by Q_1, is

the median of the data values in the lower half of the data set. The **third quartile**, denoted by Q_3, is the median of the data values in the upper half of the data set.* The following example illustrates these definitions.

Example 2.4.1

Blood Pressure The systolic blood pressures (mm Hg) of seven middle-aged men were as follows:[25]

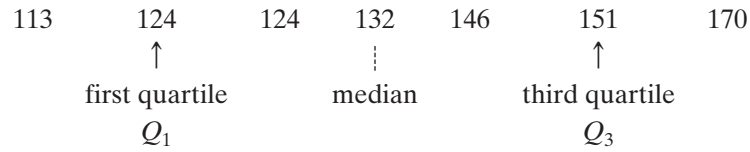
$$151 \quad 124 \quad 132 \quad 170 \quad 146 \quad 124 \quad 113$$

Putting these values in rank order, the sample is

$$113 \quad 124 \quad 124 \quad 132 \quad 146 \quad 151 \quad 170$$

The median is the fourth largest observation, which is 132. There are three data points in the lower part of the distribution: 113, 124, and 124. The median of these three values is 124. Thus, the first quartile, Q_1, is 124.

Likewise, there are three data points in the upper part of the distribution: 146, 151 and 170. The median of these three values is 151. Thus, the third quartile, Q_3, is 151.



Note that the median is not included in either the lower part or the upper part of the distribution. If the sample size, n, is even, then exactly one-half of the observations are in the lower part of the distribution and one-half are in the upper part.

The **interquartile range** is the difference between the first and third quartiles and is abbreviated as **IQR**: $\text{IQR} = Q_3 - Q_1$. For the blood pressure data in Example 2.4.1, the IQR is $151 - 124 = 27$.

Example 2.4.2

Pulse The pulses of 12 college students were measured.[26] Here are the data, arranged in order, with the position of the median indicated by a dashed line:

$$62 \quad 64 \quad 68 \quad 70 \quad 70 \quad 74 \;\vdots\; 74 \quad 76 \quad 76 \quad 78 \quad 78 \quad 80$$

The median is $\dfrac{74 + 74}{2} = 74$. There are six observations in the lower part of the distribution: 62, 64, 68, 70, 70, 74. Thus, the first quartile is the average of the third and fourth largest data values:

$$Q_1 = \frac{68 + 70}{2} = 69$$

There are six observations in the upper part of the distribution: 74, 76, 76, 78, 78, 80. Thus, the third quartile is the average of the ninth and tenth largest data values (the third and fourth values in the upper part of the distribution):

$$Q_3 = \frac{76 + 78}{2} = 77$$

*Some authors use other definitions of quartiles, as does some computer software. A common alternative definition is to say that the first quartile has rank position $(.25)(n + 1)$ and that the third quartile has rank position $(.75)(n + 1)$. Thus, if $n = 10$, the first quartile would have rank position $(.25)(11) = 2.75$—that is, to find the first quartile we would have to interpolate between the second and third largest observations. If n is large, then there is little practical difference between the definitions that various authors use.

Thus, the interquartile range is

$$\text{IQR} = 77 - 69 = 8$$

We have

62 64 68 70 70 74 | 74 76 76 78 78 80

↑ median ↑

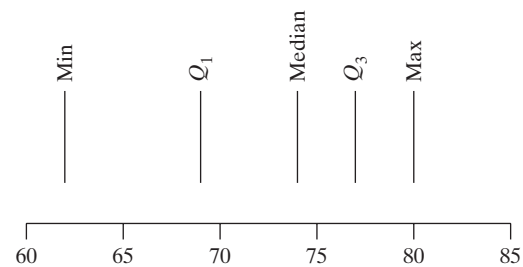
first quartile Q_1 third quartile Q_3

The minimum pulse value is 62 and the maximum is 80. ■

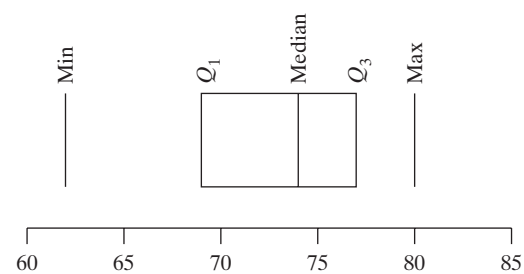
The minimum, the maximum, the median, and the quartiles, taken together, are referred to as the **five-number summary** of the data.

Boxplots

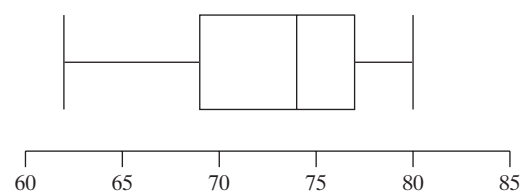
A **boxplot** is a visual representation of the five-number summary. To make a boxplot, we first make a number line; then we mark the positions minimum, Q_1, the median, Q_3, and the maximum:



Next, we make a box connecting the quartiles:



Note that the interquartile range is equal to the length of the box. Finally, we extend "whiskers" from Q_1 down to the minimum and from Q_3 up to the maximum:



A boxplot gives a quick visual summary of the distribution. We can immediately see where the center of the data is from the line within the box that locates the median. We see the spread of the total distribution, from the minimum up to the maximum, as well as the spread of the middle half of the distribution—the interquartile range—from the length of the box. The boxplot also gives an indication of the shape of the distribution; the preceding boxplot has a long lower whisker, indicating that the distribution is skewed to the left. Example 2.4.3 shows a boxplot for data from a radish growth experiment.*

Example 2.4.3

Radish Growth A common biology experiment involves growing radish seedlings under various conditions. In one version of this experiment, a moist paper towel is put into a plastic bag. Staples are put in the bag about one-third of the way from the bottom of the bag and then radish seeds are placed along the staple seam. One group of students kept their radish seed bags in total darkness for three days and then measured the length, in mm, of each radish shoot at the end of the three days. They collected 14 observations; the data are shown in Table 2.4.1.[27]

Table 2.4.1 Radish growth, in mm, after three days in total darkness

15	20	11	30	33
20	29	35	8	10
22	37	15	25	

Here are the data in order from smallest to largest:

8 10 11 **15** 15 20 **20** ¦ **22** 25 29 **30** 33 35 37

↑ median ↑

first quartile Q_1 third quartile Q_3

The quartiles are $Q_1 = 15$ and $Q_3 = 30$. The median, $\tilde{y} = 21$, is the average of the two middle values of 20 and 22. Figure 2.4.1 shows a boxplot of the same data. ■

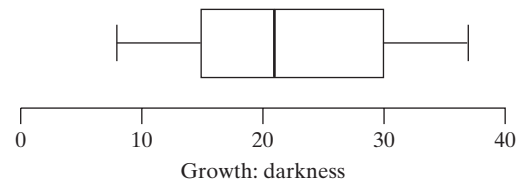


Figure 2.4.1 Boxplot of data on radish growth in darkness

Outliers

Sometimes a data point differs so much from the rest of the data that it doesn't seem to belong with the other data. Such a point is called an **outlier**. An outlier might occur because of a recording error or typographical error when the data are recorded, because of an equipment failure during an experiment, or for many other rea-

*This and subsequent boxplots in our text are slightly stylized. Different computer packages present the plot somewhat differently, but all boxplots have the same basic five-number summary.

sons. Outliers are the most interesting points in a data set. Sometimes outliers tell us about a problem with the experimental protocol (e.g., an equipment failure or a failure of a patient to take his or her medication consistently during a medical trial). At other times an outlier might alert us to the fact that a special circumstance has happened (e.g., an abnormally high or low value on a medical test could indicate the presence of a disease in a patient).

People often use the term "outlier" informally. There is, however, a common definition of "outlier" in statistical practice. To give a definition of outlier, we first discuss what are known as fences. The **lower fence** of a distribution is

$$\text{lower fence} = Q_1 - 1.5 \times \text{IQR}$$

The **upper fence** of a distribution is

$$\text{upper fence} = Q_3 + 1.5 \times \text{IQR}$$

This means that the fences are located 1.5 IQRs (i.e., 1.5 × the length of the box) beyond the end of the box in a boxplot.

Note that the fences need not be data values; indeed, there might be no data near the fences. The fences just locate limits within the sample distribution. These limits give us a way to define outliers. *An outlier is a data point that falls outside of the fences*. That is, if

$$\text{data point} < Q_1 - 1.5 \times \text{IQR}$$

or

$$\text{data point} > Q_3 + 1.5 \times \text{IQR}$$

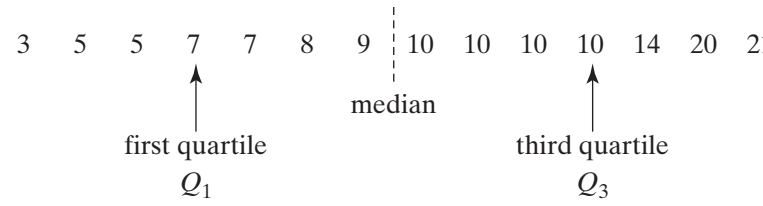
then we call the point an outlier.

Example 2.4.4

Pulse In Example 2.4.2 we saw that $Q_1 = 69, Q_3 = 77$, and IQR = 8. Thus, the lower fence is $69 - 1.5 \times 8 = 69 - 12 = 57$. Any point less than 57 would be an outlier. The upper fence is $77 + 1.5 \times 8 = 77 + 12 = 89$. Any point greater than 89 would be an outlier. Since there are no points less than 57 or greater than 89, there are no outliers in this data set. ■

Example 2.4.5

Radish Growth in Light The data in Example 2.4.3 were for radish seedlings grown in total darkness. In another part of the experiment students grew 14 radish seedlings in constant light. The observations, in order, are



Thus, the median is $\frac{9 + 10}{2} = 9.5$, Q_1 is 7, and Q_3 is 10. The interquartile range is IQR $= 10 - 7 = 3$. The lower fence is $7 - 1.5 \times 3 = 7 - 4.5 = 2.5$, so any point less than 2.5 would be an outlier. The upper fence is $10 + 1.5 \times 3 = 10 + 4.5 = 14.5$, so any point greater than 14.5 is an outlier. Thus, the two largest observations in this data set are outliers: 20 and 21. ■

The method we have defined for identifying outliers allows the bulk of the data to determine how extreme an observation must be before we consider it to be an outlier, since the quartiles and the IQR are determined from the data themselves. Thus, a point that is an outlier in one data set might not be an outlier in another data set. We label a point as an outlier if it is unusual relative to the inherent variability in the entire data set.

After an outlier has been identified, people are often tempted to remove the outlier from the data set. In general this is not a good idea. If we can identify that an outlier occurred due to an equipment error, for example, then we have good reason to remove the outlier before analyzing the rest of the data. However, quite often outliers appear in data sets without any identifiable, external reason for them. In such cases, we simply proceed with our analysis, aware that there is an outlier present. In some cases, we might want to calculate the mean, for example, with and without the outlier and then report both calculations, to show the effect of the outlier in the overall analysis. This is preferable to removing the outlier, which obscures the fact that there was an unusual data point present. In presenting data graphically, we can draw attention to outliers by using modified boxplots, which we now introduce.

Modified Boxplot

A standard variation on the idea of a boxplot is what is known as a modified boxplot. A **modified boxplot** is a boxplot in which the outliers, if any, are graphed as separate points. The advantage of a modified boxplot is that it lets us quickly see where the outliers are, if there are any.

To make a modified boxplot, we proceed as we did when first making a boxplot, except for the last step. After drawing the box for the boxplot, we check to see if there are outliers. If there are no outliers, then we extend whiskers from the box out to the extremes (the minimum and the maximum). However, if there are outliers in the upper part of the distribution, then we identify them with a dot or other plotting symbol. We then extend a whisker from Q_3 up to the largest data point that is not an outlier. Likewise, if there are outliers in the lower part of the distribution, we identify them with asterisks and extend a whisker from Q_1 down to the smallest observation that is not an outlier. Figure 2.4.2 shows the distribution of radish seedlings grown under constant light. The area between the lower and upper fences is white, while the outlying region is blue.

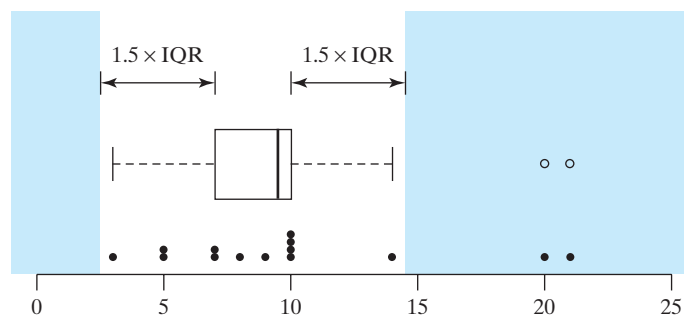


Figure 2.4.2 Dotplot and boxplot of data on radish growth in constant light. The points in the blue region are outliers.

Figure 2.4.3 shows a boxplot and a modified boxplot of the data on radish seedlings grown in constant light.

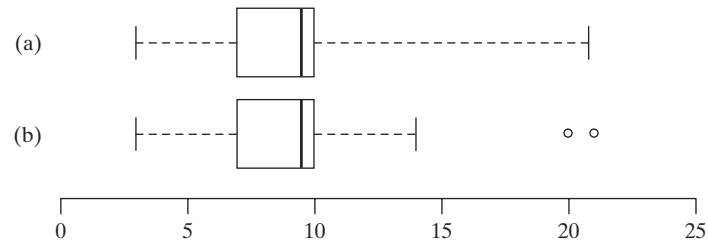


Figure 2.4.3 (a) Boxplot of data on radish growth in constant light; (b) modified boxplot of radish growth data

Most often, when people make boxplots, they make modified boxplots. Computer software is typically programmed to produce a modified boxplot when the user asks for a boxplot. Thus, we will use the term "boxplot" to mean "modified boxplot."

Exercises 2.4.1–2.4.8

2.4.1 Here are the data from Exercise 2.3.10 on the number of virus-resistant bacteria in each of 10 aliquots:

14	15	13	21	15
14	26	16	20	13

(a) Determine the median and the quartiles.
(b) Determine the interquartile range.
(c) How large would an observation in this data set have to be in order to be an outlier?

2.4.2 Here are the 18 measurements of MAO activity reported in Exercise 2.2.2:

6.8	8.4	8.7	11.9	14.2	18.8
9.9	4.1	9.7	12.7	5.2	7.8
7.8	7.4	7.3	10.6	14.5	10.7

(a) Determine the median and the quartiles.
(b) Determine the interquartile range.
(c) How large would an observation in this data set have to be in order to be an outlier?
(d) Construct a (modified) boxplot of the data.

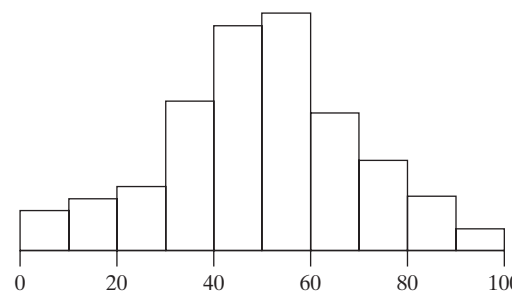
2.4.3 In a study of milk production in sheep (for use in making cheese), a researcher measured the three-month milk yield for each of 11 ewes. The yields (liters) were as follows:[28]

56.5	89.8	110.1	65.6	63.7	82.6
75.1	91.5	102.9	44.4	108.1	

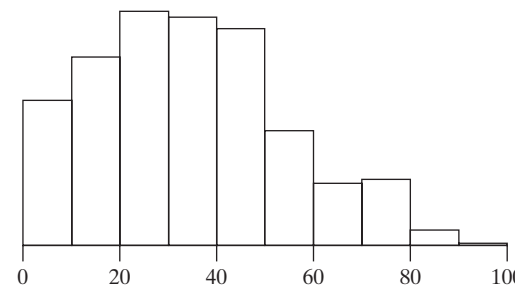
(a) Determine the median and the quartiles.
(b) Determine the interquartile range.
(c) Construct a (modified) boxplot of the data.

2.4.4 For each of the following histograms, use the histogram to estimate the median and the quartiles; then construct a boxplot for the distribution.

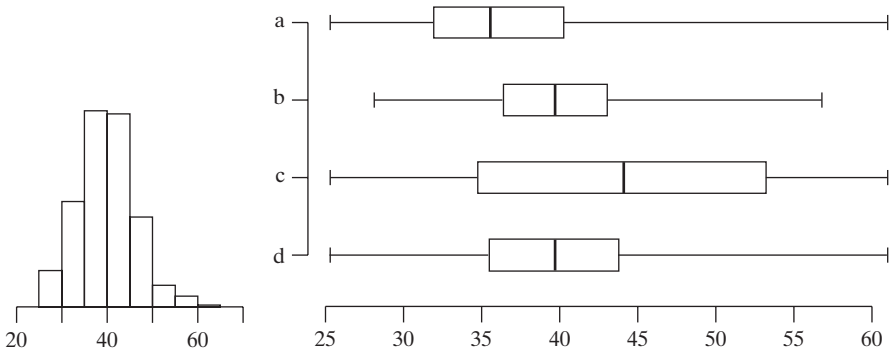
(a)



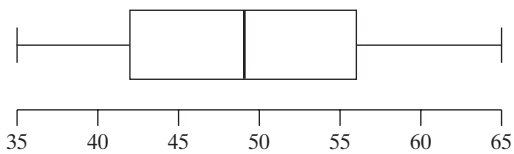
(b)



2.4.5 The following histogram shows the same data that are shown in one of the four boxplots. Which boxplot goes with the histogram? Explain your answer.



2.4.6 The following boxplot shows the five-number summary for a data set. For these data the minimum is 35, Q_1 is 42, the median is 49, Q_3 is 56, and the maximum is 65. Is it possible that no observation in the data set equals 42? Explain your answer.



2.4.7 Statistics software can be used to find the five-number summary of a data set. Here is an example of MINITAB's descriptive statistics summary for a variable stored in column 1 (Cl) of MINITAB's worksheet.

```
Variable   N   Mean   Median  TrMean  StDev  SEMean
   Cl     75 119.94  118.40  119.98   9.98    1.15

Variable    Min     Max      Q1      Q3
   Cl     95.16  145.11  113.59  127.42
```

(a) Use the MINITAB output to calculate the interquartile range.

(b) Are there any outliers in this set of data?

2.4.8 Consider the data from Exercise 2.4.7. Use the five-number summary that is given to create a boxplot of the data.

2.5 Relationships between Variables

In the previous sections we have studied **univariate** summaries of both numeric and categorical variables. A univariate summary is a graphical or numeric summary of a single variable.

The histogram, boxplot, sample mean, and median are all examples of univariate summaries for numeric data. The bar chart, frequency, and relative frequency tables are examples of univariate summaries for categorical data. In this section we present some common **bivariate** graphical summaries used to examine the *relationship* between pairs of variables.

Categorical–Categorical Relationships

To understand the relationship between two categorical variables, we first summarize the data in a **bivariate frequency table**. Unlike the frequency table presented in Section 2.2 (a univariate table), the bivariate frequency table has both rows and columns—one dimension for each variable. The choice of which variable to list with the rows and which to list with the columns is arbitrary. The following example considers the relationship between two categorical variables: *E. Coli* Source and Sampling Location.

Example 2.5.1

E. Coli Watershed Contamination In an effort to determine if there are differences in the primary sources of fecal contamination at different locations in the Morro Bay watershed, $n = 623$ water specimens were collected at three primary locations that feed into Morro Bay: Chorro Creek ($n_1 = 241$), Los Osos Creek ($n_2 = 256$), and Baywood Seeps ($n_3 = 126$).[29] DNA fingerprinting techniques were used to determine the intestinal origin of the dominant *E. coli* strain in each water specimen. *E. coli* origins were classified into the following five categories: bird, domestic pet (e.g., cat or dog), farm animal (e.g., horse, cow, pig), human, or other terrestrial mammal (e.g., fox, mouse, coyote . . .). Thus, each water specimen had *two* categorical variables measured: location (Chorro, Los Osos, or Baywood) and *E. coli* source (bird, . . . , terrestrial mammal). Table 2.5.1 presents a frequency table of the data. ■

Table 2.5.1 Frequency table of *E. coli* source by location

	E. Coli Source					
Location	Bird	Domestic pet	Farm animal	Human	Terrestrial mammal	Total
Chorro Creek	46	29	106	38	22	**241**
Los Osos Creek	79	56	32	63	26	**256**
Baywood	35	23	0	60	8	**126**
Total	**160**	**108**	**138**	**161**	**56**	**623**

While Table 2.5.1 provides a concise summary of the data, it is difficult to discover any patterns in the data. Examining relative frequencies (row or column proportions) often helps us make meaningful comparisons as seen in the following example.

Example 2.5.2

E. Coli Watershed Contamination Are domestic pets more of an *E. coli* problem (i.e., source) at Chorro Creek or Baywood? Table 2.5.1 shows that the domestic pet *E. coli* source count at Chorro (29) is higher than Baywood (23), so at first glance it seems that pets are more problematic at Chorro. However, as more water specimens were collected at Chorro ($n_1 = 241$) than Baywood ($n_2 = 126$), the relative frequency of domestic pet source *E. coli* is actually lower at Chorro ($29/241 = 0.120$) than Baywood ($23/126 = 0.183$). Table 2.5.2 displays row percentages and thus facilitates comparisons of *E. coli* sources among the locations. (Note that column percentages would not be meaningful in this context since the water was sampled by location and not by *E. coli* source.). ■

Table 2.5.2 Bivariate relative frequency table (row percentages) of *E. coli* source by location

	E. Coli Source					
Location	Bird	Domestic pet	Farm animal	Human	Terrestrial mammal	Total
Chorro Creek	19.1	12.0	44.0	15.8	9.1	**100**
Los Osos Creek	30.9	21.9	12.5	24.6	10.2	**100**
Baywood	27.8	18.3	0.0	47.6	6.3	**100**
Total	**25.7**	**17.3**	**22.2**	**25.8**	**9.0**	**100**

To visualize the data in Tables 2.5.1 and 2.5.2 we can examine **stacked bar charts**. With a stacked frequency bar chart, the overall height of each bar reflects the sample size for a level of the X categorical variable (e.g., location) while the height or thickness of a slice that makes up a bar represents the count of the Y categorical variable (e.g., *E. coli* source) for that level of X. Figure 2.5.1 displays a stacked bar chart for the *E. coli* watershed count data in Table 2.5.1.

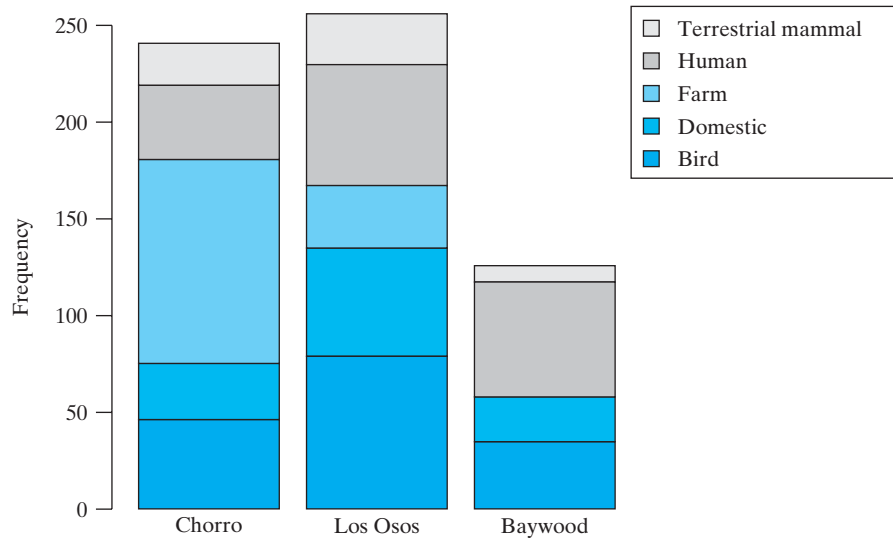


Figure 2.5.1 Stacked frequency chart of *E. coli* source by location

Like the frequency table, the stacked frequency bar chart is not conducive to making comparisons across the three locations as the sample sizes differ for these locations. (This graph does help highlight the difference in sample sizes; for example, it is very clear that many fewer water specimens were collected at Baywood.) A chart that better displays the distribution of one categorical variable across levels of another is a **stacked relative frequency** (or percentage) bar chart, which graphs the summaries from a bivariate relative frequency table such as Table 2.5.2. Figure 2.5.2 provides an example using the *E. coli* watershed contamination data. This plot normalizes the bars of Figure 2.5.1 to have the same height (100%) to facilitate comparisons across the three locations.

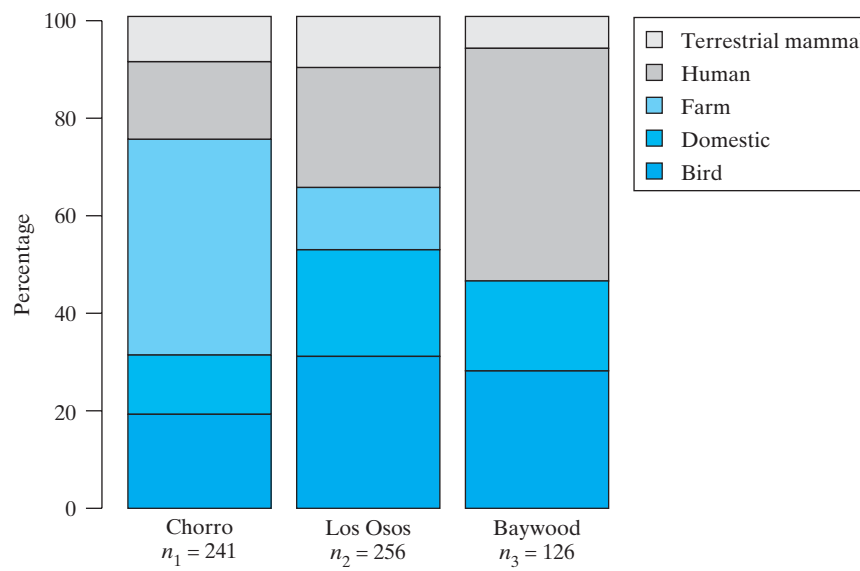


Figure 2.5.2 Stacked relative frequency (percentage) chart of *E. coli* source by location

Figure 2.5.2 makes it very easy to see that farm animals are the largest contributors of *E. coli* to Chorro Creek while humans are primarily responsible for the pollution at Baywood. The distribution of the slices in the three bars appears quite different, suggesting that the distribution of *E. coli* sources is not the same at the three locations. In Chapter 10 we will learn how to determine if these apparent differences are large enough to be compelling evidence for real differences in the distribution of *E. coli* source by location, or whether they are likely due to chance variation.

Numeric–Categorical Relationships

In Section 2.4 we learned that boxplots are graphs based on only five numbers: the minimum, first quartile, median, third quartile, and maximum. They are appealing plots because they are very simple and uncluttered, yet contain easy to read information about center, spread, skewness, and even outliers of a data set. By displaying **side-by-side boxplots** on the same graph, we are able to compare numeric data among several groups. We now consider an extension of the radish shoot growth problem in Example 2.4.3.

Example 2.5.3

Radish Growth Does light exposure alter initial radish shoot growth? The complete radish growth experiment of Example 2.4.3 actually involved a total of 42 radish seeds randomly divided to receive one of three lighting conditions for germination (14 seeds in each lighting condition): 24-hour light, diurnal light (12 hours of light and 12 hours of darkness each day), and 24 hours of darkness. At the end of three days, shoot length was measured (mm). Thus, each shoot has two variables that are measured in this study: the categorical variable lighting condition (light, diurnal, dark) and the numeric variable sprout length (mm). Figure 2.5.3 displays side-by-side boxplots of the data. The boxplots make it very easy to compare the growth under the three conditions: It appears that light inhibits shoot growth. Are the observed differences in growth among the lighting conditions just due to chance variation, or is light really altering growth? We will learn how to numerically measure the strength of this evidence and answer this question in Chapters 7 and 11. ■

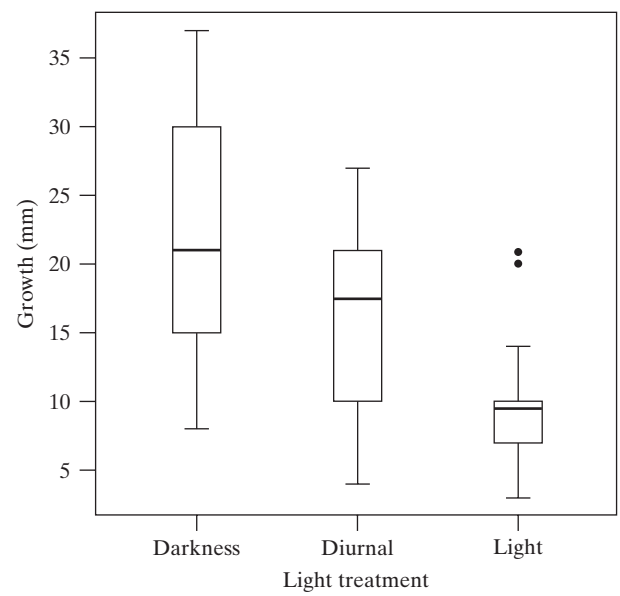


Figure 2.5.3 Side-by-side boxplots of radish growth under three conditions: constant darkness, half light–half darkness, and constant light

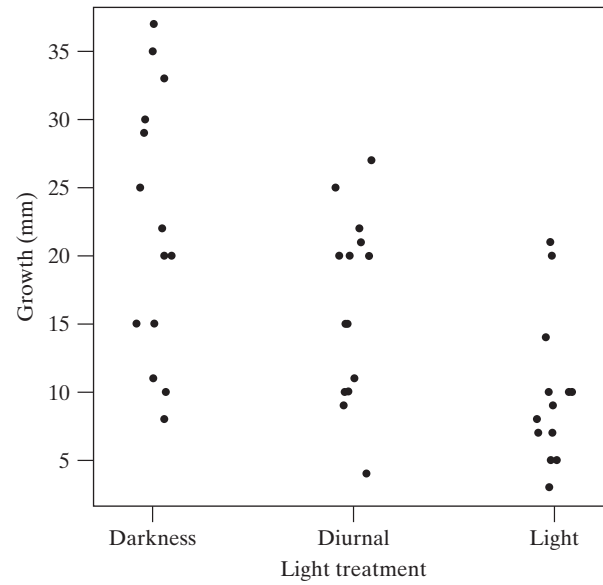


Figure 2.5.4 Side-by-side jittered dotplots of radish growth under three conditions: constant darkness, half light–half darkness, and constant light

For smaller data sets, we also may consider side-by-side dotplots of the data. Figure 2.5.4 displays a jittered side-by-side dotplot of the radish growth data of Example 2.5.3. The "jitter" is a common software option that adds horizontal scatter to the plot, helping to reduce the overlap of the dots. Choosing between side-by-side boxplots and dotplots is matter of personal preference. A good rule of thumb is to choose the plot that accurately reflects patterns in the data in the cleanest (least ink on the paper) way possible. For the radish growth example, the boxplot enables a very clean comparison of the growth under the three light treatments without hiding any information revealed by the dotplot.

Numeric–Numeric Relationships

Each of the previous examples considered comparing the distribution of one variable (either categorical or numeric) among several groups (i.e., across levels of a categorical variable). In the next example we illustrate the **scatterplot** as a tool to examine the relationship between two numeric variables, X and Y. A scatterplot plots each observed (x,y) pair as a dot on the x–y plane.

Example 2.5.4

Whale Selenium Can metal concentration in marine mammal teeth be used as a bioindicator for body burden? Selenium (Se) is an essential element that has been shown to play an important role in protecting marine mammals against the toxic effects of mercury (Hg) and other metals. Twenty beluga whales (*Delphinapterus leucas*) were harvested from the Mackenzie Delta, Northwest Territories, as part of an annual traditional Inuit hunt.[30] Each whale yielded two numeric measurements: Tooth Se (μg/g) and Liver Se (ng/g). Selenium concentrations for the whales are listed in Table 2.5.3. Tooth Se concentration (Y) is graphed against Liver Se concentration (X) in the scatterplot of Figure 2.5.5. ■

Table 2.5.3 Liver and tooth selenium concentrations of twenty belugas

Whale	Liver Se (μg/g)	Tooth Se (ng/g)	Whale	Liver Se (μg/g)	Tooth Se (ng/g)
1	6.23	140.16	11	15.28	112.63
2	6.79	133.32	12	18.68	245.07
3	7.92	135.34	13	22.08	140.48
4	8.02	127.82	14	27.55	177.93
5	9.34	108.67	15	32.83	160.73
6	10.00	146.22	16	36.04	227.60
7	10.57	131.18	17	37.74	177.69
8	11.04	145.51	18	40.00	174.23
9	12.36	163.24	19	41.23	206.30
10	14.53	136.55	20	45.47	141.31

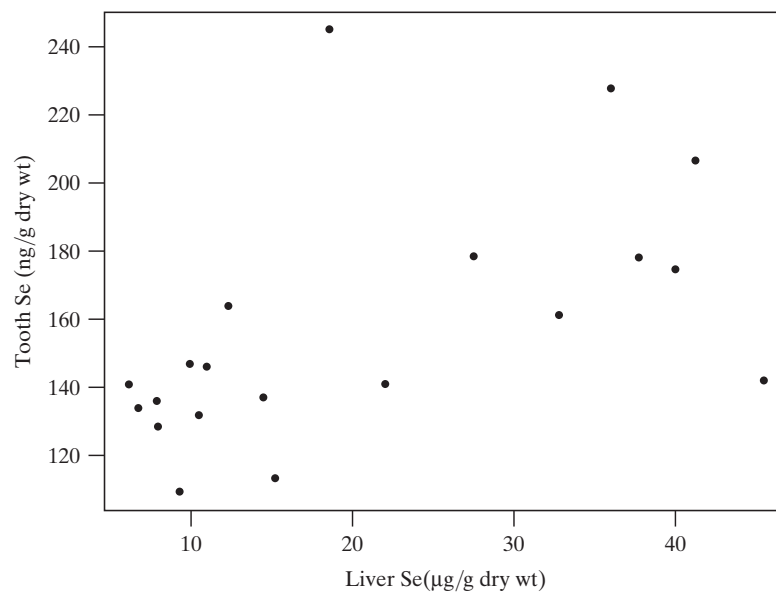


Figure 2.5.5 Scatterplot of tooth selenium concentration against liver selenium concentration for 20 belugas

Scatterplots are helpful in revealing relationships between numeric variables. In Figure 2.5.6 two lines have been added to the whale selenium scatterplot of Figure 2.5.5 to highlight the increasing trend in the data: Tooth Se concentration tends to increase with liver Se concentration. The dashed line is called a **lowess smooth** whereas the straight solid line is called a **regression line**. Many software packages allow one to easily add these lines to a scatterplot. The lowess smooth is particularly helpful in visualizing curved or nonlinear relationships in data, while the regression line is used to highlight linear trend. Generally speaking, we would choose only one of these to display on our graph. In this case, since the pattern is fairly linear (the lowess smooth is fairly straight), we would choose the solid regression line. In Chapter 12 we will learn how to identify the equation of the regression line that best summarizes the data and determine if the apparent trend in the data is likely to be just due to chance or if there is evidence for a real relationship between X and Y.

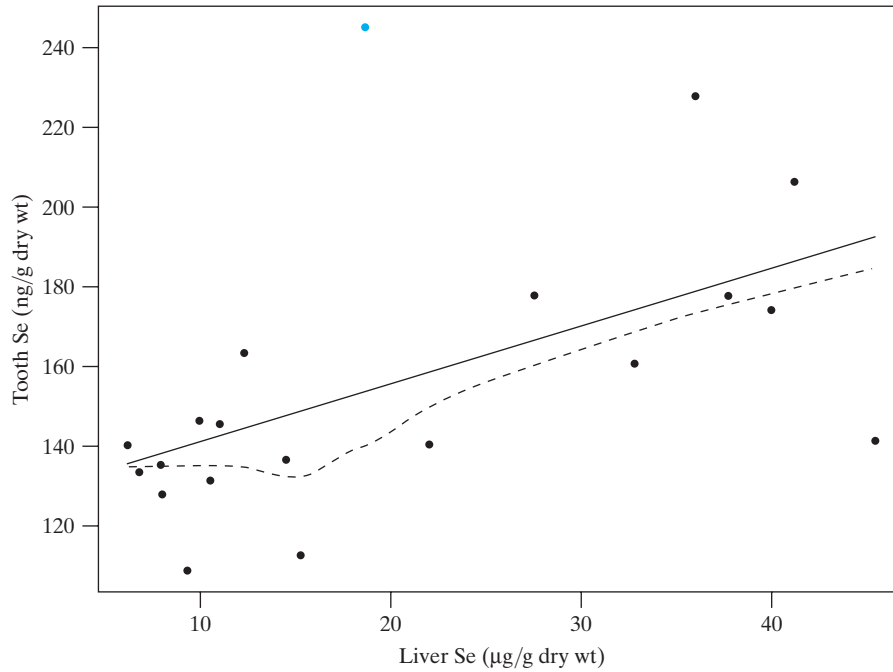


Figure 2.5.6 Scatterplot of tooth selenium concentration against liver selenium concentration for 20 belugas with regression (solid) and lowess (dashed) summary lines and outlier marked in blue

In addition to revealing relationships between two numeric variables, scatterplots also help reveal outliers that might otherwise be unnoticed in univariate plots (e.g., histograms, single boxplots, etc.). The colored point on Figure 2.5.6 falls far from the scatter of the other points. The X value of this point is not unusual in any way, and even the Y value, though large, doesn't appear extreme. The scatterplot, however, shows that the particular (x,y) pair for this whale is unusual.

Exercises 2.5.1–2.5.3

2.5.1 The two claws of the lobster (*Homarus americanus)* are identical in the juvenile stages. By adulthood, however, the two claws normally have differentiated into a stout claw called a "crusher" and a slender claw called a "cutter." In a study of the differentiation process, 26 juvenile animals were reared in smooth plastic trays and 18 were reared in trays containing oyster chips (which they could use to exercise their claws). Another 23 animals were reared in trays containing only one oyster chip. The claw configurations of all the animals as adults are summarized in the table.[31]

	CLAW CONFIGURATION		
TREATMENT	RIGHT CRUSHER, LEFT CUTTER	RIGHT CUTTER, LEFT CRUSHER	RIGHT AND LEFT CUTTER (NO CRUSHER)
Oyster chips	8	9	1
Smooth plastic	2	4	20
One oyster chip	7	9	7

(a) Create a stacked frequency bar chart to display these data.

(b) Create a stacked relative frequency bar chart to display these data.

(c) Of the two charts you created in parts (a) and (b), which is more useful for comparing the claw configurations across the three treatments? Why?

2.5.2 Does the length (mm) of the golden mantled ground squirrel (*Spermophilus lateralis*) differ by latitude in California? A graduate student captured squirrels at four locations across California. Listed from south to north the locations are Hemet, Big Bear, Susanville, and Loop Hill.[32]

HEMET	BIG BEAR	SUSANVILLE	LOOP HILL
263	274	245	273
256	256	272	291
251	249	263	278
242	264	260	281
248		271	
281			

(a) Create side-by-side dotplots of the data. Consider the geography of these four locations when making your plot. Is alphabetic order of the locations the most appropriate, or is there a better way to order the location categories?

(b) Create side-by-side boxplots of the data. Again, consider the geography of these four locations when making your plot.

(c) Of the two plots created in parts (a) and (b), which do you prefer and why?

2.5.3 The rowan (*Sorbus aucuparia*) is a tree that grows in a wide range of altitudes. To study how the tree adapts to its varying habitats, researchers collected twigs with attached buds from 12 trees growing at various altitudes in North Angus, Scotland. The buds were brought back to the laboratory and measurements were made of the dark respiration rate. The accompanying table shows the altitude of origin (in meters) of each batch of buds and the dark respiration rate (expressed as μl of oxygen per hour per mg dry weight of tissue).[33]

TREE	ALTITUDE OF ORIGIN X (M)	RESPIRATION RATE Y (μl/hr · mg)
1	90	0.11
2	230	0.20
3	240	0.13
4	260	0.15
5	330	0.18
6	400	0.16
7	410	0.23
8	550	0.18
9	590	0.23
10	610	0.26
11	700	0.32
12	790	0.37

(a) Create a scatterplot of the data.

(b) If your software allows, add a regression line to summarize the trend.

(c) If your software allows, create a scatterplot with a lowess smooth to summarize the trend.

2.6 Measures of Dispersion

We have considered the shapes and centers of distributions, but a good description of a distribution should also characterize how spread out the distribution is—are the observations in the sample all nearly equal, or do they differ substantially? In Section 2.4 we defined the interquartile range, which is one measure of dispersion. We will now consider other measures of dispersion: the range, the standard deviation, and the coefficient of variation.

The Range

The sample **range** is the difference between the largest and smallest observations in a sample. Here is an example.

Example 2.6.1

Blood Pressure The systolic blood pressures (mm Hg) of seven middle-aged men were given in Example 2.4.1 as follows:

$$113 \quad 124 \quad 124 \quad 132 \quad 146 \quad 151 \quad 170$$

For these data, the sample range is

$$170 - 113 = 57 \text{ mm Hg}$$ ■

The range is easy to calculate, but it is very sensitive to extreme values; that is, it is not robust. If the maximum in the blood pressure sample had been 190 rather than 170, the range would have been changed from 57 to 77.

We defined the interquartile range (IQR) in Section 2.4 as the difference between the quartiles. Unlike the range, the IQR is robust. The IQR of the blood

pressure data is 151 − 124 = 17. If the maximum in the blood pressure sample had been 190 rather than 170, the IQR would not have changed; it would still be 17.

The Standard Deviation

The standard deviation is the classical and most widely used measure of dispersion. Recall that a *deviation* is the difference between an observation and the sample mean:

$$\text{deviation} = \text{observation} - \bar{y}$$

The standard deviation of the sample, or sample **standard deviation**, is determined by combining the deviations in a special way, as described in the following box.

The Sample Standard Deviation The sample standard deviation is denoted by s and is defined by the following formula:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

In this formula, the expression $\sum_{i=1}^{n}(y_i - \bar{y})^2$ denotes the sum of the squared deviations.

So, to find the standard deviation of a sample, first find the deviations. Then

1. square
2. add
3. divide by $n - 1$
4. take the square root

To illustrate the use of the formula, we have chosen a data set that is especially simple to handle because the mean happens to be an integer.

Example 2.6.2

Growth of Chrysanthemums In an experiment on chrysanthemums, a botanist measured the stem elongation (mm in 7 days) of five plants grown on the same greenhouse bench. The results were as follows:[34]

$$76 \quad 72 \quad 65 \quad 70 \quad 82$$

The data are tabulated in the first column of Table 2.6.1. The sample mean is

$$\bar{y} = \frac{365}{5} = 73 \text{ mm}$$

The deviations $(y_i - \bar{y})$ are tabulated in the second column of Table 2.6.1; the first observation is 3 mm above the mean, the second is 1 mm below the mean, and so on.

The third column of Table 2.6.1 shows that the sum of the squared deviations is

$$= \sum_{i=1}^{n}(y_i - \bar{y})^2 = 164$$

Table 2.6.1 Illustration of the formula for the sample standard deviation

Observation (y_i)	Deviation ($y_i - \bar{y}$)	Squared deviation $(y_i - \bar{y})^2$
76	3	9
72	−1	1
65	−8	64
70	−3	9
82	9	81
Sum $365 = \sum_{i=1}^{n} y_i$	0	$164 = \sum_{i=1}^{n} (y_i - \bar{y})^2$

Since $n = 5$, the standard deviation is

$$\begin{aligned} s &= \sqrt{\frac{164}{4}} \\ &= \sqrt{41} \\ &= 6.4 \text{ mm} \end{aligned}$$

Note that the units of s (mm) are the same as the units of Y. This is because we have squared the deviations and then later taken the square root. ■

The sample **variance**, denoted by s^2, is simply the standard deviation squared: variance $= s^2$. Thus, $s = \sqrt{\text{variance}}$.

Example 2.6.3 Chrysanthemum Growth The variance of the chrysanthemum growth data is

$$s^2 = 41 \text{ mm}^2$$

Note that the units of the variance (mm^2) are not the same as the units of Y. ■

An abbreviation We will frequently abbreviate "standard deviation" as "SD"; the symbol "s" will be used in formulas.

Interpretation of the Definition of s

The magnitude (disregarding sign) of each deviation $(y_i - \bar{y})$ can be interpreted as the *distance* of the corresponding observation from the sample mean $\bar{y}$. Figure 2.6.1 shows a plot of the chrysanthemum growth data (Example 2.6.2) with each distance marked.

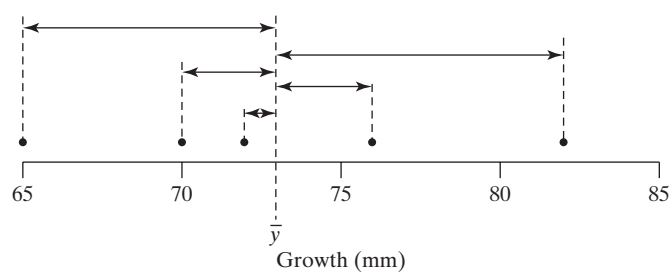


Figure 2.6.1 Plot of chrysanthemum growth data with deviations indicated as distances

From the formula for s, you can see that each deviation contributes to the SD. Thus, a sample of the same size but with less dispersion will have a smaller SD, as illustrated in the following example.

Example 2.6.4

Chrysanthemum Growth If the chrysanthemum growth data of Example 2.6.2 are changed to

$$75 \quad 72 \quad 73 \quad 75 \quad 70$$

then the mean is the same ($\bar{y} = 73$ mm), but the SD is smaller ($s = 2.1$ mm), because the observations lie closer to the mean. The relative dispersion of the two samples can easily be seen from Figure 2.6.2. ■

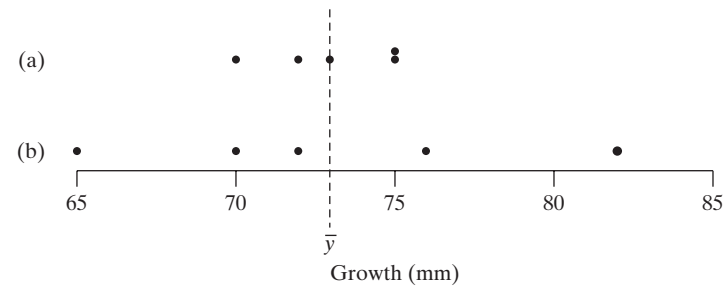


Figure 2.6.2 Two samples of chrysanthemum growth data with the same mean but different standard deviations: (a) $s = 2.1$ mm; (b) $s = 6.3$ mm

Let us look more closely at the way in which the deviations are combined to form the SD. The formula calls for dividing by $(n - 1)$. If the divisor were n instead of $(n - 1)$, then the quantity inside the square root sign would be the average (the mean) of the squared deviations. Unless n is very small, the inflation due to dividing by $(n - 1)$ instead of n is not very great, so that the SD can be interpreted approximately as

$$s \approx \sqrt{\text{sample average value of } (y_i - \bar{y})^2}$$

Thus, it is roughly appropriate to think of the SD as a "typical" distance of the observations from their mean.

Why $n - 1$? Since dividing by n seems more natural, you may wonder why the formula for the SD specifies dividing by $(n - 1)$. Note that the sum of the deviations $y_i - \bar{y}$ is always zero. Thus, once the first $n - 1$ deviations have been calculated, the last deviation is constrained. This means that in a sample with n observations there are only $n - 1$ units of information concerning deviation from the average. The quantity $n - 1$ is called the **degrees of freedom** of the standard deviation or variance. We can also give an intuitive justification of why $n - 1$ is used by considering the extreme case when $n = 1$, as in the following example.

Example 2.6.5

Chrysanthemum Growth Suppose the chrysanthemum growth experiment of Example 2.6.2 had included only one plant, so that the sample consisted of the single observation

$$73$$

For this sample, $n = 1$ and $\bar{y} = 73$. However, the SD formula breaks down (giving $\frac{0}{0}$), so the SD cannot be computed. This is reasonable, because the sample gives no information about variability in chrysanthemum growth under the experimental conditions. If the formula for the SD said to divide by n, we would obtain an SD of zero,

suggesting that there is little or no variability; such a conclusion hardly seems justified by observation of only one plant. ■

The Coefficient of Variation

The **coefficient of variation** is the standard deviation expressed as a percentage of the mean: coefficient of variation $= \frac{s}{\bar{y}} \times 100\%$. Here is an example.

Example 2.6.6

Chrysanthemum Growth For the chrysanthemum growth data of Example 2.6.2, we have $\bar{y} = 73.0$ mm and $s = 6.4$ mm. Thus,

$$\frac{s}{\bar{y}} \times 100\% = \frac{6.4}{73.0} \times 100\% = 0.088 \times 100\% = 8.8\%$$

The sample coefficient of variation is 8.8%. Thus, the standard deviation is 8.8% as large as the mean. ■

Note that the coefficient of variation is not affected by multiplicative changes of scale. For example, if the chrysanthemum data were expressed in inches instead of mm, then both $\bar{y}$ and s would be in inches, and the coefficient of variation would be unchanged. Because of its imperviousness to scale change, the coefficient of variation is a useful measure for comparing the dispersions of two or more variables that are measured on different scales.

Example 2.6.7

Girls' Height and Weight As part of the Berkeley Guidance Study,[35] the heights (in cm) and weights (in kg) of 13 girls were measured at age two. At age two, the average height was 86.6 cm and the SD was 2.9 cm. Thus, the coefficient of variation of height at age two is

$$\frac{s}{\bar{y}} \times 100\% = \frac{2.9}{86.6} \times 100\% = .033 \times 100\% = 3.3\%$$

For weight at age two the average was 12.6 kg and the SD was 1.4 kg. Thus, the coefficient of variation of weight at age two is

$$\frac{s}{\bar{y}} \times 100\% = \frac{1.4}{12.6} \times 100\% = .111 \times 100\% = 11.1\%$$

There is considerably more variability in weight than there is in height, when we express each measure of variability as a percentage of the mean. The SD of weight is a fairly large percentage of the average weight, but the SD of height is a rather small percentage of the average height. ■

Visualizing Measures of Dispersion

The range and the interquartile range are easy to interpret. The range is the spread of all the observations and the interquartile range is the spread of (roughly) the middle 50% of the observations. In terms of the histogram of a data set, the range can be visualized as (roughly) the width of the histogram. The quartiles are (roughly) the values that divide the area into four equal parts and the interquartile range is the distance between the first and third quartiles. The following example illustrates these ideas.

Example 2.6.8

Daily Gain of Cattle The performance of beef cattle was evaluated by measuring their weight gain during a 140-day testing period on a standard diet. Table 2.6.2 gives the average daily gains (kg/day) for 39 bulls of the same breed (Charolais); the observations are listed in increasing order.[36] The values range from 1.18 kg/day to 1.92 kg/day. The quartiles are 1.29, 1.41, and 1.58 kg/day. Figure 2.6.3 shows a histogram of the data, the range, the quartiles, and the interquartile range (IQR). The shaded area represents the middle 50% (approximately) of the observations. ■

Table 2.6.2 Average daily gain (kg/day) of thirty-nine Charolais bulls

1.18	1.24	1.29	1.37	1.41	1.51	1.58	1.72
1.20	1.26	1.33	1.37	1.41	1.53	1.59	1.76
1.23	1.27	1.34	1.38	1.44	1.55	1.64	1.83
1.23	1.29	1.36	1.40	1.48	1.57	1.64	1.92
1.23	1.29	1.36	1.41	1.50	1.58	1.65	

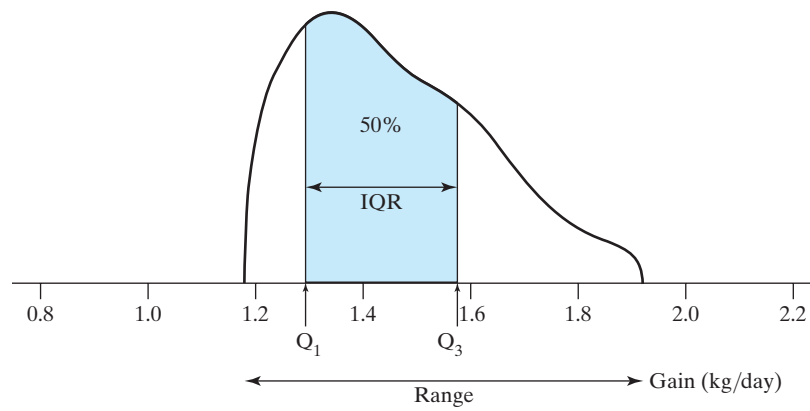


Figure 2.6.3 Smoothed histogram of 39 daily gain measurements, showing the range, the quartiles, and the interquartile range (IQR). The shaded area represents about 50% of the observations.

Visualizing the Standard Deviation

We have seen that the SD is a combined measure of the distances of the observations from their mean. It is natural to ask how many of the observations are within ±1 SD of the mean, within ±2 SDs of the mean, and so on. The following example explores this question.

Example 2.6.9

Daily Gain of Cattle For the daily-gain data of Example 2.6.8, the mean is $\bar{y} = 1.445$ kg/day and the SD is $s = 0.183$ kg/day. In Figure 2.6.4 the intervals $\bar{y} \pm s$, $\bar{y} \pm 2s$, and $\bar{y} \pm 3s$ have been marked on a histogram of the data. The interval $\bar{y} \pm s$ is

$$1.445 \pm 0.183 \text{ or } 1.262 \text{ to } 1.628$$

You can verify from Table 2.6.2 that this interval contains 25 of the 39 observations. Thus, $\frac{25}{39}$ or 64% of the observations are within ±1 SD of the mean; the corresponding area is shaded in Figure 2.6.4. The intervals $\bar{y} \pm 2s$ is

$$1.445 \pm 0.366 \text{ or } 1.079 \text{ to } 1.811$$

This interval contains $\frac{37}{39}$ or 95% of the observations. You may verify that the interval $y \pm 3s$ contains all the observations. ■

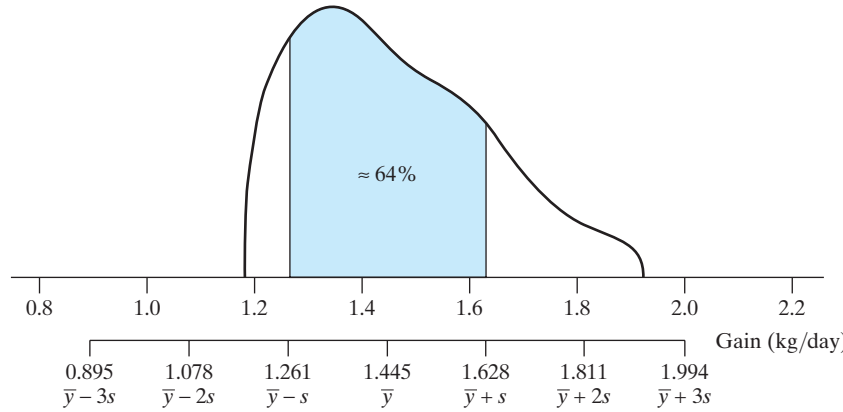


Figure 2.6.4 Histogram of daily-gain data showing intervals 1, 2, and 3 standard deviations from the mean. The shaded area represents about 64% of the observations.

It turns out that the percentages found in Example 2.6.9 are fairly typical of distributions that are observed in the life sciences.

Typical Percentages: The Empirical Rule

For "nicely shaped" distributions—that is, unimodal distributions that are not too skewed and whose tails are not overly long or short—we usually expect to find

about 68% of the observations within ±1 SD of the mean.

about 95% of the observations within ±2 SDs of the mean.

>99% of the observations within ±3 SDs of the mean.

The typical percentages enable us to construct a rough mental image of a frequency distribution if we know just the mean and SD. (The value 68% may seem to come from nowhere. Its origin will become clear in Chapter 4.)

Estimating the SD from a Histogram

The empirical rule gives us a way to construct a rough mental image of a frequency distribution if we know just the mean and SD: We can envision a histogram centered at the mean and extending out a bit more than 2 SDs in either direction. Of course, the actual distribution might not be symmetric, but our rough mental image will often be fairly accurate.

Thinking about this the other way around, we can look at a histogram and estimate the SD. To do this, we need to estimate the endpoints of an interval that is centered at the mean and that contains about 95% of the data. The empirical rule implies that this interval is roughly the same as $(\bar{y} - 2s, \bar{y} + 2s)$, so the length of the interval should be about 4 times the SD:

$$(\bar{y} - 2s, \bar{y} + 2s) \text{ has length of } 2s + 2s = 4s$$

This means

$$\text{length of interval} = 4s$$

so

$$\text{estimate of } s = \frac{\text{length of interval}}{4}$$

Of course, our visual estimate of the interval that covers the middle 95% of the data could be off. Moreover, the empirical rule works best for distributions that are symmetric. Thus, this method of estimating the SD will give only a general estimate. The method works best when the distribution is fairly symmetric, but it works reasonably well even if the distribution is somewhat skewed.

Example 2.6.10

Pulse after Exercise A group of 28 adults did some moderate exercise for five minutes and then measured their pulses. Figure 2.6.5 shows the distribution of the data.[37] We can see that about 95% of the observations are between about 75 and 125.* Thus, an interval of length 50 (125 −75) covers the middle 95% of the data. From this, we can estimate the SD to be $\frac{50}{4} = 12.5$. The actual SD is 13.4, which is not far off from our estimate. ■

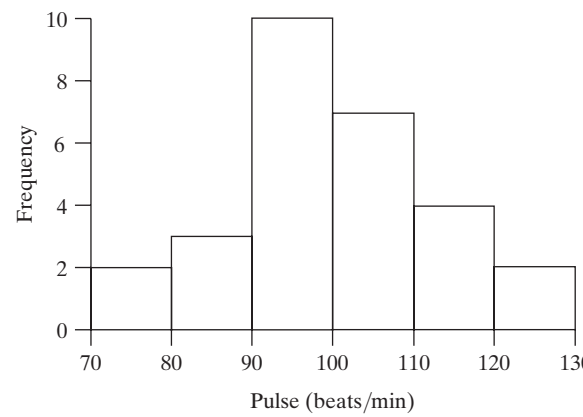


Figure 2.6.5 Pulse after moderate exercise for a group of adults

The typical percentages given by the empirical rule may be grossly wrong if the sample is small or if the shape of the frequency distribution is not "nice." For instance, the cricket singing time data (Table 2.3.1 and Figure 2.3.4) has $s = 4.4$ mm, and the interval $\bar{y} \pm s$ contains 90% of the observations. This is much higher than the "typical" 68% because the SD has been inflated by the long straggly tail of the distribution.

Comparison of Measures of Dispersion

The dispersion, or spread, of the data in a sample can be described by the standard deviation, the range, or the interquartile range. The range is simple to understand, but it can be a poor descriptive measure because it depends only on the extreme tails of the distribution. The interquartile range, by contrast, describes the spread in the central "body" of the distribution. The standard deviation takes account of all the observations and can roughly be interpreted in terms of the spread of the observations around their mean. However, the SD can be inflated by observations in the extreme tails. The interquartile range is a resistant measure, while the SD is nonresistant. Of course, the range is very highly nonresistant.

The descriptive interpretation of the SD is less straightforward than that of the range and the interquartile range. Nevertheless, the SD is the basis for most

*It is difficult to visually assess exactly where the middle 95% of the data lay using a histogram, but as this is only a visual estimate, we need not concern ourselves with producing an exact value. Our visual estimates of the SD might differ from one another, but they should all be relatively close.

standard classical statistical methods. The SD enjoys this classic status for various technical reasons, including efficiency in certain situations.

The developments in later chapters will emphasize classical statistical methods, in which the mean and SD play a central role. Consequently, in this book we will rely primarily on the mean and SD rather than other descriptive measures.

Exercises 2.6.1–2.6.16

2.6.1 Calculate the standard deviation of each of the following fictitious samples:

(a) 16, 13,18, 13

(b) 38, 30, 34, 38, 35

(c) 1, −1, 5, −1

(d) 4, 6, −1, 4, 2

2.6.2 Calculate the standard deviation of each of the following fictitious samples:

(a) 8, 6, 9, 4, 8

(b) 4, 7, 5, 4

(c) 9, 2, 6, 7, 6

2.6.3

(a) Invent a sample of size 5 for which the deviations $(y_i - \bar{y})$ are −3, −1, 0, 2, 2.

(b) Compute the standard deviation of your sample.

(c) Should everyone get the same answer for part (b)? Why?

2.6.4 Four plots of land, each 346 square feet, were planted with the same variety ("Beau") of wheat. The plot yields (lb) were as follows:[38]

35.1 30.6 36.9 29.8

(a) Calculate the mean and the standard deviation.

(b) Calculate the coefficient of variation.

2.6.5 A plant physiologist grew birch seedlings in the greenhouse and measured the ATP content of their roots. (See Example 1.1.3.) The results (nmol ATP/mg tissue) were as follows for four seedlings that had been handled identically.[39]

1.45 1.19 1.05 1.07

(a) Calculate the mean and the standard deviation.

(b) Calculate the coefficient of variation.

2.6.6 Ten patients with high blood pressure participated in a study to evaluate the effectiveness of the drug Timolol in reducing their blood pressure. The accompanying table shows systolic blood pressure measurements taken before and after two weeks of treatment with Timolol.[40] Calculate the mean and standard deviation of the *change* in blood pressure (note that some values are negative).

	BLOOD PRESSURE (mm HG)		
PATIENT	BEFORE	AFTER	CHANGE
1	172	159	−13
2	186	157	−29
3	170	163	−7
4	205	207	2
5	174	164	−10
6	184	141	−43
7	178	182	4
8	156	171	15
9	190	177	−13
10	168	138	−30

2.6.7 Dopamine is a chemical that plays a role in the transmission of signals in the brain. A pharmacologist measured the amount of dopamine in the brain of each of seven rats. The dopamine levels (nmoles/g) were as follows:[41]

6.8 5.3 6.0 5.9 6.8 7.4 6.2

(a) Calculate the mean and standard deviation.

(b) Determine the median and the interquartile range.

(c) Calculate the coefficient of variation.

(d) Replace the observation 7.4 by 10.4 and repeat parts (a) and (b). Which of the descriptive measures display resistance and which do not?

2.6.8 In a study of the lizard *Sceloporus occidentalis*, biologists measured the distance (m) run in two minutes for each of 15 animals. The results (listed in increasing order) were as follows:[42]

18.4 22.2 24.5 26.4 27.5 28.7 30.6 32.9
32.9 34.0 34.8 37.5 42.1 45.5 45.5

(a) Determine the quartiles and the interquartile range.

(b) Determine the range.

2.6.9 Refer to the running-distance data of Exercise 2.6.8. The sample mean is 32.23 m and the SD is 8.07 m. What percentage of the observations are within

(a) 1 SD of the mean?

(b) 2 SDs of the mean?

2.6.10 Compare the results of Exercise 2.6.9 with the predictions of the empirical rule.

2.6.11 Listed in increasing order are the serum creatine phosphokinase (CK) levels (U/l) of 36 healthy men (these are the data of Example 2.2.6):

25	62	82	95	110	139
42	64	83	95	113	145
48	67	84	100	118	151
57	68	92	101	119	163
58	70	93	104	121	201
60	78	94	110	123	203

The sample mean CK level is 98.3 U/l and the SD is 40.4 U/l. What percentage of the observations are within

(a) 1 SD of the mean?

(b) 2 SDs of the mean?

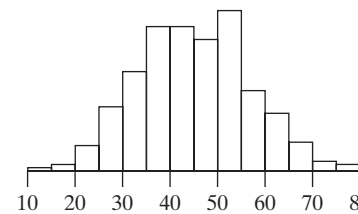
(c) 3 SDs of the mean?

2.6.12 Compare the results of Exercise 2.6.11 with the predictions of the empirical rule.

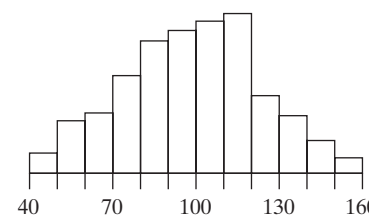
2.6.13 The girls in the Berkeley Guidance Study (Example 2.6.7) who were measured at age two were measured again at age nine. Of course, the average height and weight were much greater at age nine than at age two. Likewise, the SDs of height and of weight were much greater at age nine, than they were at age two. But what about the coefficient of variation of height and the coefficient of variation of weight? It turns out that one of these went up a moderate amount from age two to age nine, but for the other variable the increase in the coefficient of variation was fairly large. For which variable, height or weight, would you expect the coefficient of variation to change more between age two and age nine? Why? (*Hint*: Think about how genetic factors influence height and weight and how environmental factors influence height and weight.)

2.6.14 Consider the 13 girls mentioned in Example 2.6.7. At age 18 their average height was 166.3 cm and the SD of their heights was 6.8 cm. Calculate the coefficient of variation.

2.6.15 Here is a histogram. Estimate the mean and the SD of the distribution.



2.6.16 Here is a histogram. Estimate the mean and the SD of the distribution.



2.7 Effect of Transformation of Variables (Optional)

Sometimes when we are working with a data set, we find it convenient to transform a variable. For example, we might convert from inches to centimeters or from °F to °C. Transformation, or reexpression, of a variable Y means replacing Y by a new variable, say Y'. To be more comfortable working with data, it is helpful to know how the features of a distribution are affected if the observed variable is transformed.

The simplest transformations are **linear** transformations, so called because a graph of Y against Y' would be a straight line. A familiar reason for linear transformation is a change in the scale of measurement, as illustrated in the following two examples.

Example 2.7.1

Weight Suppose Y represents the weight of an animal in kg, and we decide to reexpress the weight in lb. Then

$$Y = \text{Weight in kg}$$
$$Y' = \text{Weight in lb}$$

so

$$Y' = 2.2Y$$

This is a **multiplicative** transformation, because Y' is calculated from Y by multiplying by the constant value 2.2. ■

Example 2.7.2

Body Temperature Measurements of basal body temperature (temperature on waking) were made on 47 women.[43]

Typical observations Y, in °C, were

$$Y: \quad 36.23, \quad 36.41, \quad 36.77, \quad 36.15, \quad \ldots$$

Suppose we convert these data from °C to °F, and call the new variable Y':

$$Y': \quad 97.21, \quad 97.54, \quad 98.19, \quad 97.07, \quad \ldots$$

The relation between Y and Y' is

$$Y' = 1.8Y + 32$$

The combination of **additive** (+32) and multiplicative (×1.8) changes indicates a linear relationship. ■

Another reason for linear transformation is **coding**, which means transforming the data for convenience in handling the numbers. The following is an example.

Example 2.7.3

Body Temperature Consider the temperature data of Example 2.7.2. If we subtract 36 from each observation, the data become

$$0.23, \quad 0.41, \quad 0.77, \quad 0.15, \quad \ldots$$

This is additive coding, since we added a constant value (−36) to each observation. Now suppose we further transform the data to the form

$$23, \quad 41, \quad 77, \quad 15, \quad \ldots$$

This step of the coding is multiplicative, since each observation is multiplied by a constant value (100). ■

As the foregoing examples illustrate, a linear transformation consists of (1) multiplying all the observations by a constant, or (2) adding a constant to all the observations, or (3) both.

How Linear Transformations Affect the Frequency Distribution

A linear transformation of the data does not change the essential shape of its frequency distribution; by suitably scaling the horizontal axis, you can make the transformed histogram identical to the original histogram. Example 2.7.4 illustrates this idea.

Example 2.7.4

Body Temperature Figure 2.7.1 shows the distribution of 47 temperature measurements that have been transformed by first subtracting 36 from each observation and then multiplying by 100 (as in Examples 2.7.2 and 2.7.3). That is, $Y' = (Y - 36) \times 100$. The figure shows that the two distributions can be represented by the same histogram with different horizontal scales. ■

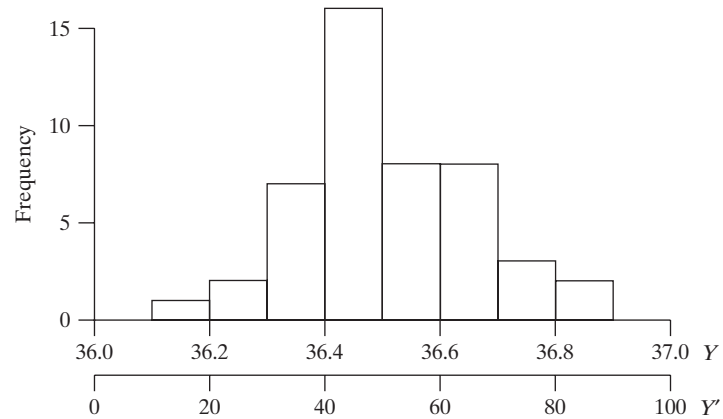


Figure 2.7.1 Distribution of 47 temperature measurements showing original and linearly transformed scales

How Linear Transformations Affect $\bar{y}$ and s

The effect of a linear transformation on $\bar{y}$ is "natural"; that is, **under a linear transformation,** $\bar{y}$ changes like Y. For instance, if temperatures are converted from °C to °F, then the mean is similarly converted:

$$Y' = 1.8Y + 32 \qquad \text{so} \qquad \bar{y}' = 1.8\bar{y}' + 32$$

The effect of multiplying Y by a positive constant on s is "natural"; if $Y' = c \times Y$, with $c > 0$, then $s' = c \times s$. For instance, if weights are converted from kg to lb, the SD is similarly converted: $s' = 2.2s$. If $Y' = c \times Y$ and $c < 0$, then $s' = -c \times s$. In general, if $Y' = c \times Y$ then $s' = |c| \times s$.

However, an additive transformation does not affect s. If we add or subtract a constant, we do not change how spread out the distribution is, so s does not change. Thus, for example, we would *not* convert the SD of temperature data from °C to °F in the same way as we convert each observation; we would multiply the SD by 1.8, but we would *not* add 32. The fact that the SD is unchanged by additive transformation will appear less surprising if you recall (from the definition) that s depends only on the deviations $(y_i - \bar{y})$, and these are not changed by an additive transformation. The following example illustrates this idea.

Example 2.7.5

Additive Transformation Consider a simple set of fictitious data, coded by subtracting 20 from each observation. The original and transformed observations are shown in Table 2.7.1.

The SD for the original observations is

$$s = \sqrt{\frac{(-1)^2 + (0)^2 + (2)^2 + (-1)^2}{3}}$$
$$= 1.4$$

Table 2.7.1 Effect of additive transformation

	Original observations (y)	Deviations ($y_i - \bar{y}$)	Transformed observations (y')	Deviations ($y'_i - \bar{y}$)
	25	−1	5	−1
	26	0	6	0
	28	2	8	2
	25	−1	5	−1
Mean	26		6	

Because the deviations are unaffected by the transformation, the SD for the transformed observations is the same:

$$s' = 1.4$$ ■

An additive transformation effectively picks up the histogram of a distribution and moves it to the left or to the right on the number line. The shape of the histogram does not change and the deviations do not change, so the SD does not change. A multiplicative transformation, on the other hand, stretches or shrinks the distribution, so the SD gets larger or smaller accordingly.

Other Statistics Under linear transformations, other measures of center (for instance, the median) change like $\bar{y}$, and other measures of dispersion (for instance, the interquartile range) change like s. The quartiles themselves change like $\bar{y}$.

Nonlinear Transformations

Data are sometimes reexpressed in a nonlinear way. Examples of nonlinear transformations are

$$Y' = \sqrt{Y}$$
$$Y' = \log(Y)$$
$$Y' = \frac{1}{Y}$$
$$Y' = Y^2$$

These transformations are termed "nonlinear" because a graph of Y' against Y would be a curve rather than a straight line. Computers make it easy to use nonlinear transformations. The logarithmic transformation is especially common in biology because many important relationships can be simply expressed in terms of logs. For instance, there is a phase in the growth of a bacterial colony when log(colony size) increases at a constant rate with time. [Note that logarithms are used in some familiar scales of measurement, such as pH measurement or earthquake magnitude (Richter scale).]

Nonlinear transformations can affect data in complex ways. For example, the mean does not change "naturally" under a log transformation; the log of the mean is *not* the same as the mean of the logs. Furthermore, nonlinear transformations (unlike linear ones) *do* change the essential shape of a frequency distribution.

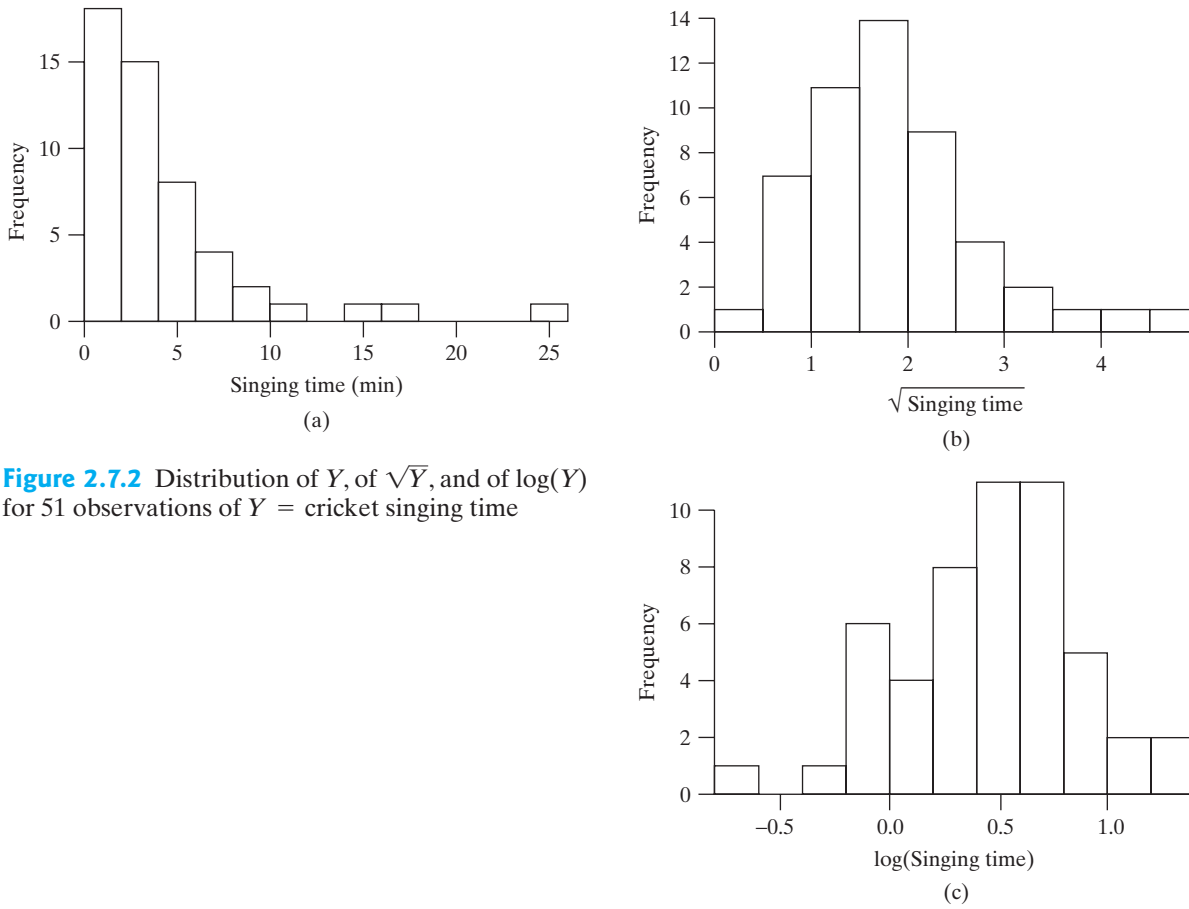


Figure 2.7.2 Distribution of Y, of $\sqrt{Y}$, and of $\log(Y)$ for 51 observations of Y = cricket singing time

In future chapters we will see that if a distribution is skewed to the right, such as the cricket singing-time distribution shown in Figure 2.7.2, then we may wish to apply a transformation that makes the distribution more symmetric, by pulling in the right-hand tail. Using $Y' = \sqrt{Y}$ will pull in the right-hand tail of a distribution and push out the left-hand tail. The transformation $Y' = \log(Y)$ is more severe than $\sqrt{Y}$ in this regard. The following example shows the effect of these transformations.

Example 2.7.6

Cricket Singing Times Figure 2.7.2(a) shows the distribution of the cricket singing-time data of Table 2.3.1. If we transform these data by taking square roots, the transformed data have the distribution shown in Figure 2.7.2(b). Taking logs (base 10) yields the distribution shown in Figure 2.7.2(c). Notice that the transformations have the effect of "pulling in" the straggly upper tail and "stretching out" the clumped values on the lower end of the original distribution. ■

Exercises 2.7.1–2.7.6

2.7.1 A biologist made a certain pH measurement in each of 24 frogs; typical values were[44]

$$7.43, \quad 7.16, \quad 7.51, \ldots$$

She calculated a mean of 7.373 and a standard deviation of 0.129 for these original pH measurements. Next, she transformed the data by subtracting 7 from each observation and then multiplying by 100. For example, 7.43 was transformed to 43. The transformed data are

$$43, \quad 16, \quad 51, \quad \ldots$$

What are the mean and standard deviation of the transformed data?

2.7.2 The mean and SD of a set of 47 body temperature measurements were as follows:[45]

$$\bar{y} = 36.497\ ^\circ\text{C} \quad s = 0.172\ ^\circ\text{C}$$

If the 47 measurements were converted to °F,

(a) What would be the new mean and SD?

(b) What would be the new coefficient of variation?

2.7.3 A researcher measured the average daily gains (in kg/day) of 20 beef cattle; typical values were[46]

$$1.39, \quad 1.57, \quad 1.44, \quad \ldots$$

The mean of the data was 1.461 and the standard deviation was 0.178.

(a) Express the mean and standard deviation in lb/day. (*Hint*: 1 kg = 2.20 lb.)

(b) Calculate the coefficient of variation when the data are expressed (i) in kg/day; (ii) in lb/day.

2.7.4 Consider the data from Exercise 2.7.3. The mean and SD were 1.461 and 0.178. Suppose we transformed the data from

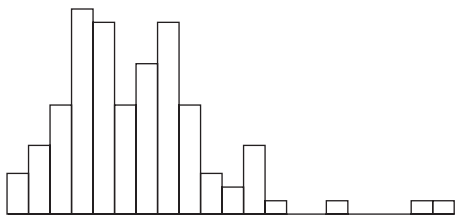
$$1.39, \quad 1.57, \quad 1.44, \quad \ldots$$

to

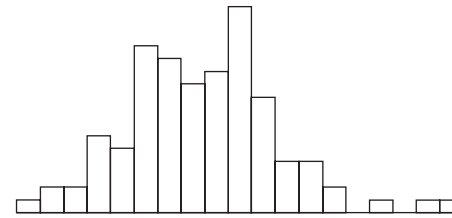
$$39, \quad 57, \quad 44, \quad \ldots$$

What would be the mean and standard deviation of the transformed data?

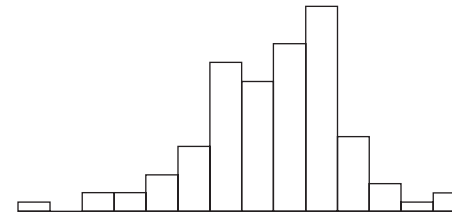
2.7.5 The following histogram shows the distribution for a sample of data:



One of the following histograms is the result of applying a square root transformation and the other is the result of applying a log transformation. Which is which? How do you know?



(a)



(b)

2.7.6 (Computer problem) The file 'Exer2.7.6.csv' is included on the data disk packaged with this text. This file contains 36 observations on the number of dendritic branch segments emanating from nerve cells taken from the brains of newborn guinea pigs. (These data were used in Exercise 2.2.4.) Open the file and enter the data into a statistics package. Make a histogram of the data, which are skewed to the right. Now consider the following possible transformations: sqrt(Y), log(Y), and 1/sqrt(Y). Which of these transformations does the best job of meeting the goal of making the resulting distribution reasonably symmetric?

2.8 Statistical Inference

The description of a data set is sometimes of interest for its own sake. Usually, however, the researcher hopes to generalize, to extend the findings beyond the limited scope of the particular group of animals, plants, or other units that were actually observed. Statistical theory provides a rational basis for this process of generalization, building on the random sampling model from Section 1.3 and taking into account the variability of the data. The key idea of the statistical approach is to view the particular data in a study as a sample from a larger population; the population is the real focus of scientific and/or practical interest. The following example illustrates this idea.

Example 2.8.1

Table 2.8.1 Blood types of 3,696 persons

Blood type	Frequency
A	1,634
B	327
AB	119
O	1,616
Total	3,696

Blood Types In an early study of the ABO blood-typing system, researchers determined blood types of 3,696 persons in England. The results are given in Table 2.8.1.[47]

These data were not collected for the purpose of learning about the blood types of those particular 3,696 people. Rather, they were collected for their scientific value as a source of information about the distribution of blood types in a larger population. For instance, one might presume that the blood type distribution of all English people should resemble the distribution for these 3,696 people. In particular, the observed relative frequency of type A blood was

$$\frac{1634}{3696} \text{ or } 44\% \text{ type A}$$

One might conclude from this that approximately 44% of the people in England have type A blood. ■

The process of drawing conclusions about a population, based on observations in a sample from that population, is called **statistical inference**. For instance, in Example 2.8.1 the conclusion that approximately 44% of the people in England have type A blood would be a statistical inference. The inference is shown schematically in Figure 2.8.1. Of course, such an inference might be entirely wrong—perhaps the 3,696 people are not at all representative of English people in general. We might be worried about two possible sources of difficulty: (1) the 3,696 people might have been selected in a way that was systematically biased for (or against) type A people, and (2) the number of people examined might have been too small to permit generalization to a population of many millions. In general, it turns out that the population size being in the millions is *not* a problem, but bias in the way people are selected is a big concern.

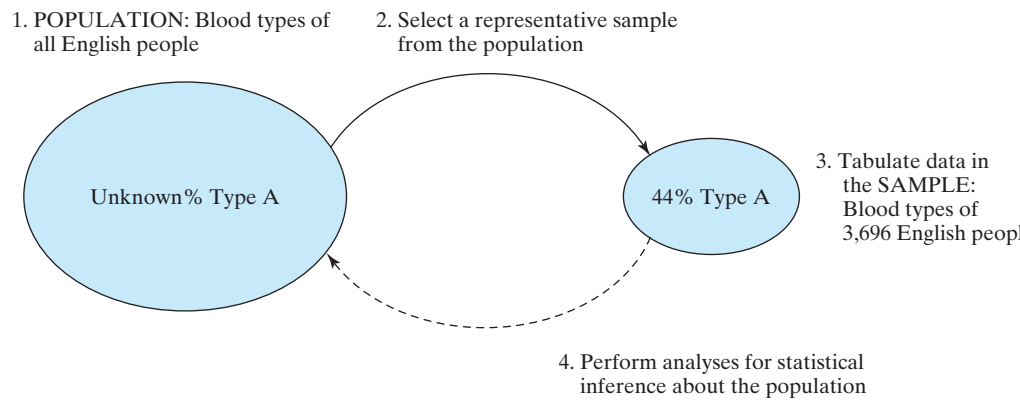


Figure 2.8.1 Schematic representation of inference from sample to population regarding prevalence of blood type A

In making a statistical inference, we hope that the sample resembles the population closely—that the sample is *representative* of the population. In Section 1.3 we saw how sampling errors and nonsampling errors can lead to nonrepresentative samples. However, even in the absence of bias we must ask how likely it is that a particular sample will provide a good representation of the population. The important question is: *How representative (of the population) is a sample likely to be?* We will see in Chapter 5 how statistical theory can help to answer this question.

Specifying the Population

In Section 1.3 we emphasized that the collection of individuals that comprise a sample should be representative of the population. In fact, this requirement is a bit stronger than what is actually necessary. Ultimately, what matters is that the measurements that we obtain on the variable of interest are representative of the values present in the population. The following provides an example of a case where the sample members might not be representative of the population, but one could argue that the measurements taken from this sample could be viewed as representative of the larger population.

Example 2.8.2

Blood Types How were the 3,696 English people of Example 2.8.1 actually chosen? It appears from the original paper that this was a "sample of convenience," that is, friends of the investigators, employees, and sundry unspecified sources. There is little basis for believing that the *people* themselves would be representative of the entire English population. Nevertheless, one might argue that their *blood types* might be (more or less) representative of the population. The argument would be that the biases that entered into the selection of those particular people were probably not related to blood type. [Nonetheless, an objection to this argument might be made on the basis of race. For example, the racial distribution of the sample could differ substantially from the racial distribution of England (the population) and there are known differences in blood type distributions among races.] The argument for representativeness would be much less plausible if the observed variable were blood pressure rather than blood type; we know that blood pressure tends to increase with age, and the selection procedure was undoubtedly biased against certain age groups (for example, elderly people). ■

As Example 2.8.2 shows, whether the measurements obtained from a sample are likely to be representative of the measurements from a population depends not only on how the observational units (in this case people) were chosen, but also on the variable that was observed. Ideally we would always work with random samples, but we have noted that in some instances random samples are not possible or convenient. However, by turning our attention to the measurements themselves rather than the individuals from which they came, we can often make an argument for the generalizabiltity (or lack of generalizability) of our results to a larger population. We do this by thinking of the population as consisting of observations or a collection of values from a measurement process, rather than of people or other observational units. The following is another example.

Example 2.8.3

Alcohol and MOPEG The biochemical MOPEG plays a role in brain function. Seven healthy male volunteers participated in a study to determine whether drinking alcohol might elevate the concentration of MOPEG in the cerebrospinal fluid. The MOPEG concentration was measured twice for each man—once at the start of the experiment, and again after he drank 80 gm of ethanol. The results (in pmol/ml) are given in Table 2.8.2.[48]

Let us focus on the rightmost column, which shows the change in MOPEG concentration (that is, the difference between the "after" and the "before" measurements). In thinking of these values as a sample from a population, we need to specify all the details of the experimental conditions—how the cerebrospinal specimens were obtained, the exact timing of the measurements and the alcohol

Table 2.8.2 Effect of alcohol on MOPEG

	MOPEG concentration		
Volunteer	Before	After	Change
1	46	56	10
2	47	52	5
3	41	47	6
4	45	48	3
5	37	37	0
6	48	51	3
7	58	62	4

consumption, and so on—as well as relevant characteristics of the volunteers themselves. Thus, the definition of the population might be something like this:

> Population Change in cerebrospinal MOPEG concentration in healthy young men when measured before and after drinking 80 gm of ethanol, both measurements being made at 8:00 A.M., . . . (other relevant experimental conditions are specified here).

There is no single "correct" definition of a population for an experiment like this. A scientist reading a report of the experiment might find this definition too narrow (for instance, perhaps it does not matter that the volunteers were measured at 8:00 A.M.) or too broad. She might use her knowledge of alcohol and brain chemistry to formulate her own definition, and she would then use that definition as a basis for interpreting these seven observations. ■

Describing a Population

Because observations are made only on a sample, characteristics of biological populations are almost never known exactly. Typically, our knowledge of a population characteristic comes from a sample. In statistical language, we say that the sample characteristic is an estimate of the corresponding population characteristic. Thus, estimation is a type of statistical inference.

Just as each sample has a distribution, a mean, and an SD, so also we can envision a population distribution, a population mean, and a population SD. In order to discuss inference from a sample to a population, we will need a language for describing the population. This language parallels the language that describes the sample. A sample characteristic is called a **statistic**; a population characteristic is called a **parameter**.

Proportions

For a categorical variable, we can describe a population by simply stating the proportion, or relative frequency, of the population in each category. The following is a simple example.

Example 2.8.4

Oat Plants In a certain population of oat plants, resistance to crown rust disease is distributed as shown in Table 2.8.3.[49] ■

Table 2.8.3 Disease resistance in oats

Resistance	Proportion of plants
Resistant	0.47
Intermediate	0.43
Susceptible	0.10
Total	1.00

Remark The population described in Example 2.8.4 is realistic, but it is not a specific real population; the exact proportions for any real population are not known. For similar reasons, we will use fictitious but realistic populations in several other examples, here and in Chapters 3, 4, and 5.

For categorical data, the sample proportion of a category is an estimate of the corresponding population proportion. Because these two proportions are not necessarily the same, it is essential to have a notation that distinguishes between them. We denote the population proportion of a category by p and the sample proportion by $\hat{p}$ (read "p-hat"):

$$p = \text{Population proportion}$$
$$\hat{p} = \text{Sample proportion}$$

The symbol "^" can be interpreted as "estimate of." Thus,

$$\hat{p} \text{ is an estimate of } p.$$

We illustrate this notation with an example.

Example 2.8.5

Lung Cancer Eleven patients suffering from adenocarcinoma (a type of lung cancer) were treated with the chemotherapeutic agent Mitomycin. Three of the patients showed a positive response (defined as shrinkage of the tumor by at least 50%).[50] Suppose we define the population for this study as "responses of all adenocarcinoma patients." Then we can represent the sample and population proportions of the category "positive response" as follows:

p = Proportion of positive responders among all adenocarcinoma patients

$\hat{p}$ = Proportion of positive responders among the 11 patients in the study

$$\hat{p} = \frac{3}{11} = 0.27$$

Note that p is unknown, and $\hat{p}$, which is known, is an estimate of p. ■

We should emphasize that an "estimate," as we are using the term, may or may not be a *good* estimate. For instance, the estimate $\hat{p}$ in Example 2.8.5 is based on very few patients; estimates based on a small number of observations are subject to considerable uncertainty. Of course, the question of whether an estimation procedure is good or poor is an important one, and we will show in later chapters how this question can be answered.

Other Descriptive Measures

If the observed variable is quantitative, one can consider descriptive measures other than proportions—the mean, the quartiles, the SD, and so on. Each of these quantities can be computed for a sample of data, and each is an estimate of its corresponding

population analog. For instance, the sample median is an estimate of the population median. In later chapters, we will focus especially on the mean and the SD, and so we will need a special notation for the population mean and SD. **The population mean is denoted by μ (mu), and the population SD is denoted by σ (sigma).** We may define these as follows for a quantitative variable Y:

$$\mu = \text{Population average value of } Y$$

$$\sigma = \sqrt{\text{Population average value of } (Y - \mu)^2}$$

The following example illustrates this notation.

Example 2.8.6

Tobacco Leaves An agronomist counted the number of leaves on each of 150 tobacco plants of the same strain (Havana). The results are shown in Table 2.8.4.[51]

The sample mean is

$$\bar{y} = 19.78 = \text{Mean number of leaves on the 150 plants}$$

Table 2.8.4 Number of leaves on tobacco plants

Number of leaves	Frequency (number of plants)
17	3
18	22
19	44
20	42
21	22
22	10
23	6
24	1
Total	150

The population mean is

$$\mu = \text{Mean number of leaves on Havana tobacco plants grown under these conditions}$$

We do not know μ, but we can regard $\bar{y} = 19.78$ as an estimate of μ. The sample SD is

$$s = 1.38 = \text{SD of number of leaves on the 150 plants}$$

The population SD is

$$\sigma = \text{SD of number of leaves on Havana tobacco plants grown under these conditions}$$

We do not know σ, but we can regard $s = 1.38$ as an estimate of σ.* ■

*You may wonder why we use $\bar{y}$ and s instead of $\hat{\mu}$ and $\hat{\sigma}$. One answer is tradition. Another answer is that since "^" means estimate, you might have other estimates in mind.

2.9 Perspective

In this chapter we have considered various ways of describing a set of data. We have also introduced the notion of regarding features of a sample as estimates of corresponding features of a suitably defined population.

Parameters and Statistics

Some features of a distribution—for instance, the mean—can be represented by a single number, while some—for instance, the shape—cannot. We have noted that a numerical measure that describes a sample is called a statistic. Correspondingly, a numerical measure that describes a population is called a parameter. For the most important numerical measures, we have defined notations to distinguish between the statistic and the parameter. These notations are summarized in Table 2.9.1 for convenient reference.

Table 2.9.1 Notation for some important statistics and parameters

Measure	Sample value (statistic)	Population value (parameter)
Proportion	$\hat{p}$	p
Mean	$\bar{y}$	μ
Standard deviation	s	σ

A Look Ahead

It is natural to view a sample characteristic (for instance, $\bar{y}$) as an estimate of the corresponding population characteristic (for instance, μ). But in taking such a view, one must guard against unjustified optimism. Of course, if the sample were perfectly representative of the population, then the estimate would be perfectly accurate. But this raises the central question: How representative (of the population) is a sample likely to be? Intuition suggests that, if the observational units are appropriately selected, then the sample should be more or less representative of the population. Intuition also suggests that larger samples should tend to be more representative than smaller samples. These intuitions are basically correct, but they are too vague to provide practical guidance for research in the life sciences. Practical questions that need to be answered are

1. How can an investigator judge whether a sample can be viewed as "more or less" representative of a population?
2. How can an investigator quantify "more or less" in a specific case?

In Section 1.3 we described a theoretical probability model based on random sampling that provides a framework for the judgment in question (1), and in Chapter 6 we will see how this model can provide a concrete answer to question (2). Specifically, in Chapter 6 we will see how to analyze a set of data so as to quantify how closely the sample mean ($\bar{y}$) estimates the population mean (μ). But before returning to data analysis in Chapter 6, we will need to lay some groundwork in Chapters 3, 4, and 5; the developments in these chapters are an essential prelude to understanding the techniques of statistical inference.

Supplementary Exercises 2.S.1–2.S.20

2.S.1 A sample of four students had the following heights (in cm): 180, 182, 179, 176. Suppose a fifth student were added to the group. How tall would that student have to be to make the mean height of the group equal 181?

2.S.2 A botanist grew 15 pepper plants on the same greenhouse bench. After 21 days, she measured the total stem length (cm) of each plant, and obtained the following values:[52]

12.4	12.2	13.4
10.9	12.2	12.1
11.8	13.5	12.0
14.1	12.7	13.2
12.6	11.9	13.1

(a) Construct a dotplot for these data, and mark the positions of the quartiles.

(b) Calculate the interquartile range.

2.S.3 In a behavioral study of the fruitfly *Drosophila melanogaster*, a biologist measured, for individual flies, the total time spent preening during a six-minute observation period. The following are the preening times (sec) for 20 flies:[53]

34	24	10	16	52
76	33	31	46	24
18	26	57	32	25
48	22	48	29	19

(a) Determine the median and the quartiles.

(b) Determine the interquartile range.

(c) Construct a (modified) boxplot of the data.

2.S.4 To calibrate a standard curve for assaying protein concentrations, a plant pathologist used a spectrophotometer to measure the absorbance of light (wavelength 500 nm) by a protein solution. The results of 27 replicate assays of a standard solution containing 60 μg protein per ml water were as follows:[54]

0.111	0.115	0.115	0.110	0.099
0.121	0.107	0.107	0.100	0.110
0.106	0.116	0.098	0.116	0.108
0.098	0.120	0.123	0.124	0.122
0.116	0.130	0.114	0.100	0.123
0.119	0.107			

Construct a frequency distribution and display it as a table and as a histogram.

2.S.5 Refer to the absorbance data of Exercise 2.S.4.

(a) Determine the median, the quartiles, and the interquartile range.

(b) How large must an observation be to be an outlier?

2.S.6 The midrange is defined as the average of the minimum and maximum of a distribution. Is the midrange a robust statistic? Why or why not?

2.S.7 Twenty patients with severe epilepsy were observed for eight weeks. The following are the numbers of major seizures suffered by each patient during the observation period:[55]

5 0 9 6 0 0 5 0 6 1
5 0 0 0 0 7 0 0 4 7

(a) Determine the median number of seizures.

(b) Determine the mean number of seizures.

(c) Construct a histogram of the data. Mark the positions of the mean and the median on the histogram.

(d) What feature of the frequency distribution suggests that neither the mean nor the median is a meaningful summary of the experience of these patients?

2.S.8 Calculate the standard deviation of each of the following fictitious samples:

(a) 11, 8, 4, 10, 7 (b) 23, 29, 24, 21, 23

(c) 6, 0, −3, 2, 5

2.S.9 To study the spatial distribution of Japanese beetle larvae in the soil, researchers divided a 12- × 12-foot section of a cornfield into 144 one-foot squares. They counted the number of larvae Y in each square, with the results shown in the following table.[56]

NUMBER OF LARVAE	FREQUENCY (NUMBER OF SQUARES)
0	13
1	34
2	50
3	18
4	16
5	10
6	2
7	1
Total	144

(a) The mean and standard deviation of Y are $\bar{y} = 2.23$ and $s = 1.47$. What percentage of the observations are within

(i) 1 standard deviation of the mean?
(ii) 2 standard deviations of the mean?

(b) Determine the total number of larvae in all 144 squares. How is this number related to $\bar{y}$?

(c) Determine the median value of the distribution.

2.S.10 One measure of physical fitness is maximal oxygen uptake, which is the maximum rate at which a person can consume oxygen. A treadmill test was used to determine the maximal oxygen uptake of nine college women before and after participation in a 10-week program of vigorous exercise. The accompanying table shows the before and after measurements and the change (after–before); all values are in ml O_2 per mm per kg body weight.[57]

	MAXIMAL OXYGEN UPTAKE		
PARTICIPANT	BEFORE	AFTER	CHANGE
1	48.6	38.8	−9.8
2	38.0	40.7	2.7
3	31.2	32.0	0.8
4	45.5	45.4	−0.1
5	41.7	43.2	1.5
6	41.8	45.3	3.5
7	37.9	38.9	1.0
8	39.2	43.5	4.3
9	47.2	45.0	−2.2

The following computations are to be done on the *change* in maximal oxygen uptake (the right-hand column).

(a) Calculate the mean and the standard deviation.

(b) Determine the median.

(c) Eliminate participant 1 from the data and repeat parts (a) and (b). Which of the descriptive measures display resistance and which do not?

2.S.11 A veterinary anatomist investigated the spatial arrangement of the nerve cells in the intestine of a pony. He removed a block of tissue from the intestinal wall, cut the block into many equal sections, and counted the number of nerve cells in each of 23 randomly selected sections. The counts were as follows.[58]

35 19 33 34 17 26 16 40
28 30 23 12 27 33 22 31
28 28 35 23 23 19 29

(a) Determine the median, the quartiles, and the interquartile range.

(b) Construct a boxplot of the data.

2.S.12 Exercise 2.S.11 asks for a boxplot of the nerve-cell data. Does this graphic support the claim that the data came from a reasonably symmetric distribution?

2.S.13 A geneticist counted the number of bristles on a certain region of the abdomen of the fruitfly *Drosophila melanogaster*. The results for 119 individuals were as shown in the table.[59]

NUMBER OF BRISTLES	NUMBER OF FLIES	NUMBER OF BRISTLES	NUMBER OF FLIES
29	1	38	18
30	0	39	13
31	1	40	10
32	2	41	15
33	2	42	10
34	6	43	2
35	9	44	2
36	11	45	3
37	12	46	2

(a) Find the median number of bristles.

(b) Find the first and third quartiles of the sample.

(c) Make a boxplot of the data.

(d) The sample mean is 38.45 and the standard deviation is 3.20. What percentage of the observations fall within 1 standard deviation of the mean?

2.S.14 The carbon monoxide in cigarettes is thought to be hazardous to the fetus of a pregnant woman who smokes. In a study of this hypothesis, blood was drawn from pregnant women before and after smoking a cigarette. Measurements were made of the percent of blood hemoglobin bound to carbon monoxide as carboxyhemoglobin (COHb). The results for 10 women are shown in the table.[60]

	BLOOD COHB (%)		
SUBJECT	BEFORE	AFTER	INCREASE
1	1.2	7.6	6.4
2	1.4	4.0	2.6
3	1.5	5.0	3.5
4	2.4	6.3	3.9
5	3.6	5.8	2.2
6	0.5	6.0	5.5
7	2.0	6.4	4.4
8	1.5	5.0	3.5
9	1.0	4.2	3.2
10	1.7	5.2	3.5

(a) Calculate the mean and standard deviation of the *increase* in COHb.

(b) Calculate the mean COHb before and the mean after. Is the mean increase equal to the increase in means?

(c) Determine the median increase in COHb.

(d) Repeat part (c) for the before measurements and for the after measurements. Is the median increase equal to the increase in medians?

2.S.15 (Computer problem) A medical researcher in India obtained blood specimens from 31 young children, all of whom were infected with malaria. The following data, listed in increasing order, are the numbers of malarial parasites found in 1 ml of blood from each child.[61]

100	140	140	271	400	435	455	770
826	1,400	1,540	1,640	1,920	2,280	2,340	3,672
4,914	6,160	6,560	6,741	7,609	8,547	9,560	10,516
14,960	16,855	18,600	22,995	29,800	83,200	134,232	

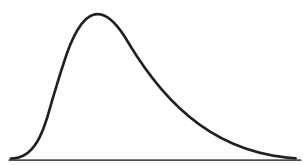
(a) Construct a frequency distribution of the data, using a class width of 10,000; display the distribution as a histogram.

(b) Transform the data by taking the logarithm (base 10) of each observation. Construct a frequency distribution of the transformed data and display it as a histogram. How does the log transformation affect the shape of the frequency distribution?

(c) Determine the mean of the original data and the mean of the log-transformed data. Is the mean of the logs equal to the log of the mean?

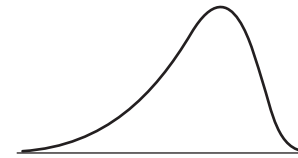
(d) Determine the median of the original data and the median of the log-transformed data. Is the median of the logs equal to the log of the median?

2.S.16 Rainfall, measured in inches, for the month of June in Cleveland, Ohio, was recorded for each of 41 years.[62] The values had a minimum of 1.2, an average of 3.6, and a standard deviation of 1.6. Which of the following is a rough histogram for the data? How do you know?



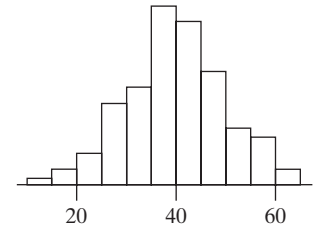
I

II

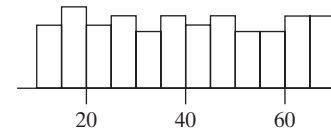


III

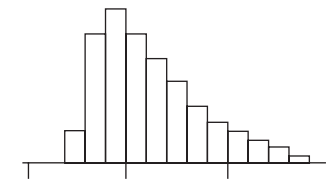
2.S.17 The following histograms (a), (b), and (c) show three distributions.



(a)



(b)



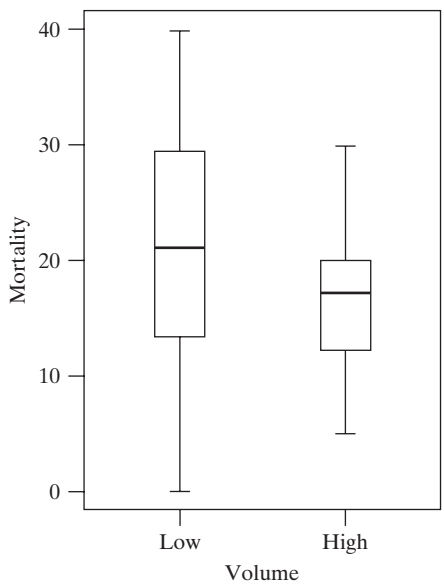
(c)

The accompanying computer output shows the mean, median, and standard deviation of the three distributions, plus the mean, median, and standard deviation for a fourth distribution. Match the histograms with the statistics. Explain your reasoning. (One set of statistics will not be used.)

```
1. Count     100          2. Count     100
   Mean      41.3522         Mean      39.6761
   Median    39.5585         Median    39.5377
   StdDev    13.0136         StdDev    10.0476

3. Count     100          4. Count     100
   Mean      37.7522         Mean      39.6493
   Median    39.5585         Median    39.5448
   StdDev    13.0136         StdDev    17.5126
```

2.S.18 The following boxplots show mortality rates (deaths within one year per 100 patients) for heart transplant patients at various hospitals. The low-volume hospitals are those that perform between 5 and 9 transplants per year. The high-volume hospitals perform 10 or more transplants per year.[63] Describe the distributions, paying special attention to how they compare to one another. Be sure to note the shape, center, and spread of each distribution.



2.S.19 (Computer problem) Physicians measured the concentration of calcium (nM) in blood samples from 38 healthy persons. The data are listed as follows:[64]

95	110	135	120	88	125
112	100	130	107	86	130
122	122	127	107	107	107
88	126	125	112	78	115
78	102	103	93	88	110
104	122	112	80	121	126
90	96				

Calculate appropriate measures of the center and spread of the distribution. Describe the shape of the distribution and any unusual features in the data.

2.S.20 The following boxplot shows the same data that are shown in one of the three histograms. Which histogram goes with the boxplot? Explain your answer.