CHAPTER TWO

DESCRIPTIVE STATISTICS

SECTION 2.1 Introduction

The first step in looking at data is to describe the data at hand in some concise way. In smaller studies this step can be accomplished by listing each data point. In general, however, this procedure is tedious or impossible and, even if it were possible, would not give an overall picture of what the data look like.

EXAMPLE 2.1

Cancer, Nutrition Some investigators have proposed that consumption of vitamin A prevents cancer. To test this theory, a dietary questionnaire to collect data on vitamin-A consumption among 200 hospitalized cancer cases and 200 controls might be used. The controls would be matched on age and sex to the cancer cases and would be in the hospital at the same time for an unrelated disease. What should be done with these data after they are collected? ■■■

Before any formal attempt to answer this question can be made, the vitamin-A consumption among cases and controls must be described. Consider Figure 2.1. The **bar graphs** show visually that the controls have a higher vitamin-A consumption than the cases do, particularly in doses higher than the recommended daily allowance (RDA).

EXAMPLE 2.2

Pulmonary Disease Medical researchers have often suspected that passive smokers—people who themselves do not smoke but who live or work in an environment where others smoke—might have impaired pulmonary function as a result. In 1980 a research group in San Diego published results indicating that passive smokers did indeed have significantly lower pulmonary function than comparable nonsmokers who did not work in smoky environments [1]. As supporting evidence, the authors measured the carbon-monoxide (CO) concentrations in the working environments of passive smokers and of nonsmokers (where no smoking was permitted in the workplace) to see if the relative CO concentration changed over the course of the day. These results are displayed in the form of a **scatter plot** in Figure 2.2. ■■■

Figure 2.2 clearly shows that the CO concentrations in the two working environments are about the same early in the day but diverge widely in the middle of the day and then converge again after the working day is over at 7 P.M.

Graphic displays illustrate the important role of descriptive statistics, which is to quickly display data to give the researcher a clue as to the principal trends in the data and suggest hints as to where a more detailed look at the data, using the methods of inferential statistics, might be worthwhile. Descriptive statistics are also crucially important in conveying the final results of studies in written publications. Unless it is one of their primary interests, most readers will not have time to critically evaluate the work of others but will be influenced mainly by the descriptive statistics presented.

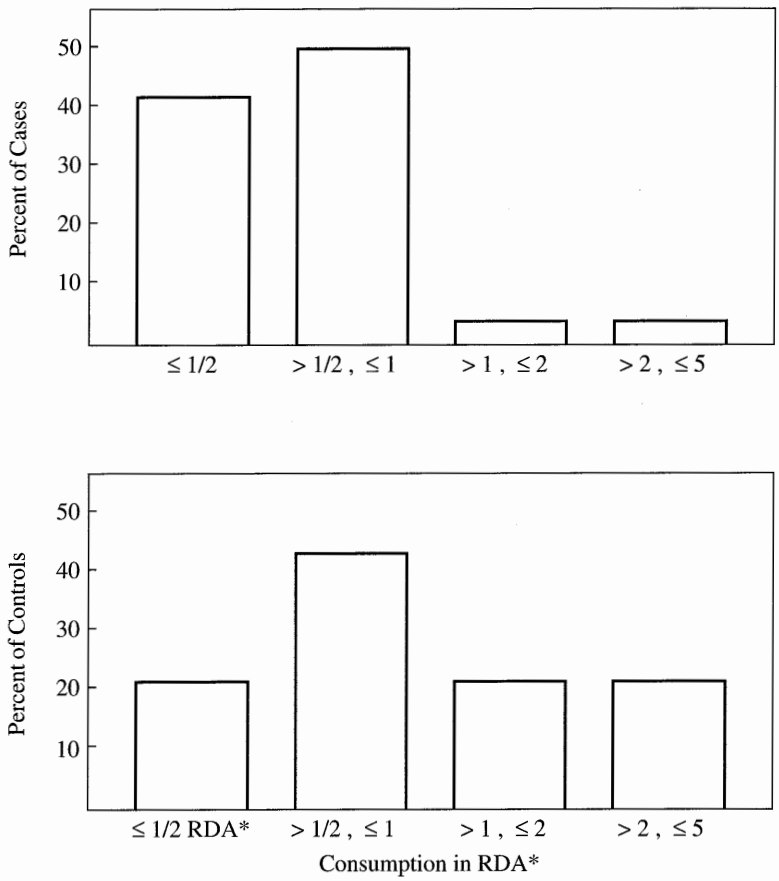

FIGURE 2.1
Daily vitamin-A consumption among cancer cases and controls

*RDA = Recommended Daily Allowance

What makes a good graphic or numeric display? The principal guideline is that the material should be as self-contained as possible and should be understandable without reading the text. These attributes require clear labeling. The captions, units, and axes on graphs should be clearly labeled, and the statistical terms used in tables and figures should be well defined. The quantity of material presented is equally important. If bar graphs are constructed, then care must be taken that neither too many nor too few groups be displayed. The same is true of tabular material.

Many methods are available for summarizing data in both numeric and graphic form. In this chapter the methods are summarized and their strengths and weaknesses given.

SECTION 2.2 Measures of Central Location

The basic problem of statistics can be stated as follows: Consider a sample of data $x_1, \ldots, x_n$, where x_1 corresponds to the first sample point and x_n corresponds to the nth sample point. Presuming that the sample is drawn from some population P, what inferences or conclusions can be made about P from the sample?

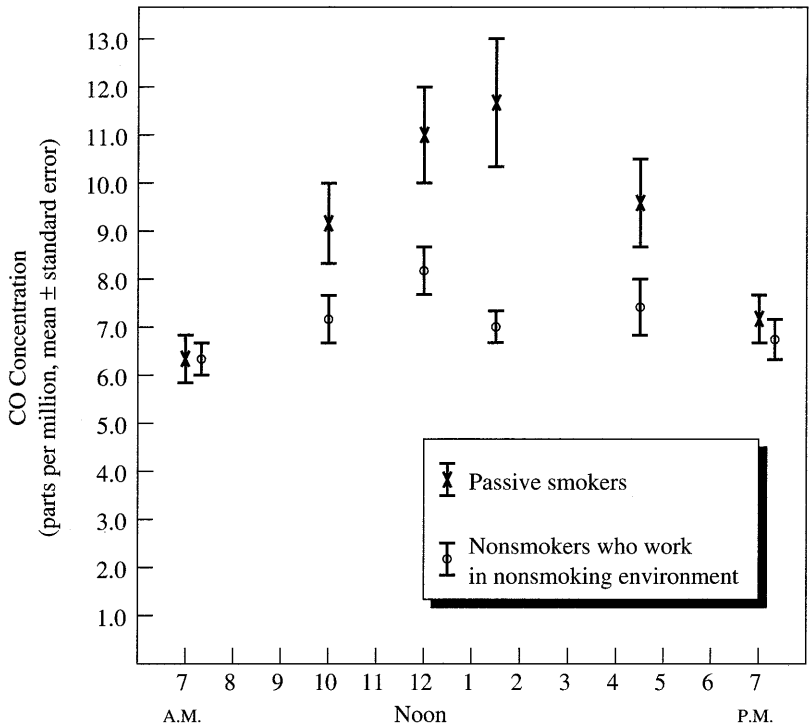

FIGURE 2.2
Mean carbon-monoxide concentration (± standard error) by time of day as measured in the working environment of passive smokers and nonsmokers who work in nonsmoking environments

Source: Reproduced with permission of *The New England Journal of Medicine*, *302*, 720-723, 1980.

Before this question can be answered, the data must be summarized as succinctly as possible, since the number of sample points is frequently large and it is easy to lose track of the overall picture by looking at all the data at once. One type of measure useful for summarizing data defines the center, or middle, of the sample. This type of measure is a **measure of central location**.

2.2.1 The Arithmetic Mean

How to define the middle of a sample may seem obvious, but the more you think about it, the less obvious it becomes. Suppose the sample consists of birthweights of all live-born infants born at a private hospital in San Diego, California, during a 1-week period. This sample is shown in Table 2.1.

TABLE 2.1
Sample of birthweights of live-born infants born at a private hospital in San Diego, California, during a 1-week period (g)

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

One measure of central location for this sample is the arithmetic mean (colloquially referred to as the average). The arithmetic mean (or mean or sample mean) is usually denoted by $\bar{x}$.

DEFINITION 2.1

The **arithmetic mean** is the sum of the all the observations divided by the number of observations. It is written in statistical terms as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

■

The sign Σ (sigma) in Definition 2.1 is referred to as a summation sign. The expression

$$\sum_{i=1}^{n} x_i$$

is simply a short way of writing the quantity $(x_1 + x_2 + \cdots + x_n)$.

If a and b are integers, then

$$\sum_{i=a}^{b} x_i$$

means that

(1) The value of a is substituted for i and x_a is used as the first term in the summation.

(2) The value of i is increased by 1 to $a + 1$ and x_{a+1} is added to x_a.

(3) Step 2 is repeated as many times as possible until $i = b$.

The term x_b is the last term in the summation. Thus,

$$\sum_{i=a}^{b} x_i = x_a + x_{a+1} + \cdots + x_b$$

If $a = b$, then $\sum_{i=a}^{b} x_i = x_a$. One fundamental property of summation signs is that if each term in the summation is a multiple of the same constant c, then c can be factored out from the summation; that is,

$$\sum_{i=1}^{n} cx_i = c\left(\sum_{i=1}^{n} x_i\right)$$

EXAMPLE 2.3 If

$$x_1 = 2 \qquad x_2 = 5 \qquad x_3 = -4$$

find

$$\sum_{i=1}^{3} x_i \quad \sum_{i=2}^{3} x_i \quad \sum_{i=1}^{3} x_i^2 \quad \sum_{i=1}^{3} 2x_i$$

SOLUTION

$$\sum_{i=1}^{3} x_i = 2 + 5 - 4 = 3 \qquad \sum_{i=2}^{3} x_i = 5 - 4 = 1$$

$$\sum_{i=1}^{3} x_i^2 = 4 + 25 + 16 = 45 \qquad \sum_{i=1}^{3} 2x_i = 2\sum_{i=1}^{3} x_i = 6$$

■■■

It is important to become familiar with summation signs, because they are used extensively in the remainder of this text.

EXAMPLE 2.4 What is the arithmetic mean for the sample of birthweights in Table 2.1?

$$\bar{x} = (3265 + 3260 + \cdots + 2834)/20 = 3166.9 \text{ g}$$

The arithmetic mean is, in general, a very natural measure of central location. One of its principal limitations, however, is that it is overly sensitive to extreme values. In this instance it may not be representative of the location of the great majority of the sample points. For example, if the first infant in Table 2.1 happened to be a premature infant weighing 500 g rather than 3265 g, then the arithmetic mean of the sample would be reduced to 3028.7 g. In this instance, 7 of the birthweights would be lower than the arithmetic mean, and 13 would be higher than the arithmetic mean. It is possible in extreme cases for all but one of the sample points to be on one side of the arithmetic mean. The arithmetic mean is a poor measure of central location in these types of samples, since it does not reflect the center of the sample. Nevertheless, the arithmetic mean is by far the most widely used measure of central location.

2.2.2 The Median

An alternative measure of central location, perhaps second in popularity to the arithmetic mean, is the **median** or, more precisely, the **sample median**.

Suppose there are n observations in a sample. If these observations are ordered from smallest to largest, then the median is defined as follows:

DEFINITION 2.2

The **sample median** is

(1) The $\left(\frac{n+1}{2}\right)$th largest observation if n is odd

(2) The average of the $\left(\frac{n}{2}\right)$th and $\left(\frac{n}{2}+1\right)$th largest observations if n is even

The rationale for these definitions is to ensure an equal number of sample points on both sides of the sample median. The median is defined differently when n is even and odd because it is impossible to achieve this goal with one uniform definition. For samples with an odd sample size, there is a unique central point; for example, for samples of size 7, the fourth largest point is the central point in the sense that 3 points are both smaller and larger than it. For samples with an even sample size, there is no unique central point and the middle 2 values must be averaged. Thus, for samples of size 8, the fourth and fifth largest points would be averaged to obtain the median, since neither is the central point.

EXAMPLE 2.5 Compute the sample median for the sample in Table 2.1

SOLUTION First, arrange the sample in ascending order:

2069, 2581, 2759, 2834, 2838, 2841, 3031, 3101, 3200, 3245, 3248, 3260, 3265, 3314, 3323, 3484, 3541, 3609, 3649, 4146

Since n is even,

$$\text{Sample median} = \text{average of the 10th and 11th largest observations}$$
$$= (3245 + 3248)/2 = 3246.5 \text{ g} \quad \blacksquare\blacksquare\blacksquare$$

EXAMPLE 2.6 **Infectious Disease** Consider the data set in Table 2.2, which consists of white-blood counts taken on admission of all patients entering a small hospital in Allentown, Pennsylvania, on a given day. Compute the median white-blood count.

TABLE 2.2 Sample of admission white-blood counts for all patients entering a hospital in Allentown, PA, on a given day (× 1000)

i	x_i	i	x_i
1	7	6	3
2	35	7	10
3	5	8	12
4	9	9	8
5	8		

SOLUTION First, order the sample as follows: 3, 5, 7, 8, 8, 9, 10, 12, 35. Since n is odd, the sample median is given by the fifth largest point, which equals 8. $\blacksquare\blacksquare\blacksquare$

The principal strength of the sample median is that it is insensitive to very large or very small values. In particular, if the second patient in Table 2.2 had a white count of 65,000 rather than 35,000, the sample median would remain unchanged, since the fifth largest value is still 8000. Conversely, the arithmetic mean would increase dramatically from 10,778 in the original sample to 14,111 in the new sample. The principal weakness of the sample median is that it is determined mainly by the middle points in a sample and is less sensitive to the actual numerical values of the remaining data points.

2.2.3 Comparison of the Arithmetic Mean and the Median

If a distribution is **symmetric**, then the relative position of the points on each side of the sample median will be the same. Examples of distributions expected to be roughly symmetric include the distribution of birthweights in Table 2.1 and the distribution of systolic blood-pressure measurements taken on all 30–39-year-old factory workers in a given workplace [Figure 2.3(a)].

If a distribution is **positively skewed** (or skewed to the right), then points above the median will tend to be farther from the median in absolute value than points below the median. An example of such a distribution would be the distribution of the number of years of oral contraceptive (OC) use by a group of women aged 20–29 years [Figure 2.3(b)]. Similarly, if a distribution is **negatively skewed** (or skewed to the left), then points below the median will tend to be farther from the median in absolute value than points above the median. An example of such a distribution would be the distribution of relative humidities observed in a humid climate at the same time of day over a number of days. In this case, most of the humidities will be at or close to 100%, with a few very low humidities on dry days [Figure 2.3(c)].

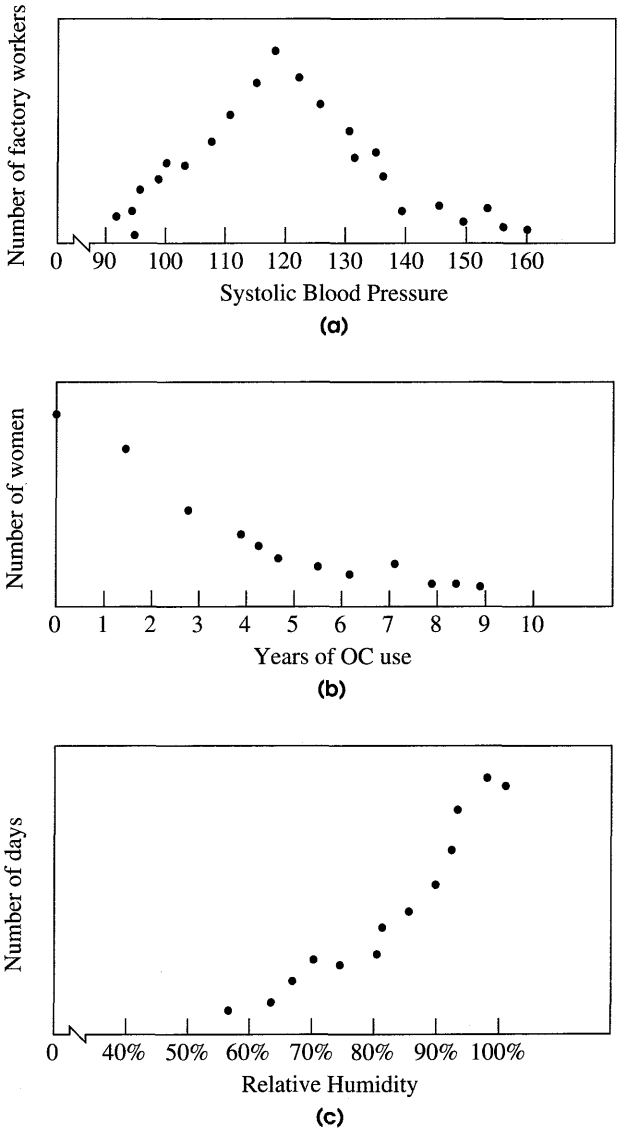

FIGURE 2.3
Graphic displays of (a) symmetric, (b) positively skewed, and (c) negatively skewed distributions

In many samples, the relationship between the arithmetic mean and the sample median can be used to assess the symmetry of a distribution. In particular, for symmetric distributions, the arithmetic mean will be approximately the same as the median. For positively skewed distributions, the arithmetic mean will tend to be larger than the median; for negatively skewed distributions, the arithmetic mean will tend to be smaller than the median.

2.2.4 The Mode

Another widely used measure of central location is the mode.

DEFINITION 2.3

The **mode** is the most frequently occurring value among all the observations in a sample. ■

EXAMPLE 2.7

Family Planning Consider the sample of time intervals between successive menstrual periods for a group of 500 college women aged 18–21, as shown in Table 2.3. The frequency column gives the number of women who reported each of the respective durations. Twenty-eight days is the mode, since it is the most frequently occurring value. ■■■

TABLE 2.3 Sample of time intervals between successive menstrual periods of college-aged women (days)

Value	Frequency	Value	Frequency	Value	Frequency
24	5	29	96	34	7
25	10	30	63	35	3
26	28	31	24	36	2
27	64	32	9	37	1
28	185	33	2	38	1

EXAMPLE 2.8

Compute the mode of the distribution in Table 2.2.

SOLUTION

The mode is 8000 because it occurs more frequently than any other white count. ■■■

Some distributions have more than one mode. In fact, one useful method of classifying distributions is by the number of modes present. A distribution with one mode is referred to as **unimodal**; two modes, **bimodal**; three modes, **trimodal**; and so forth.

EXAMPLE 2.9

Compute the mode of the distribution in Table 2.1.

SOLUTION

There is no mode, since all the values occur exactly once. ■■■

Example 2.9 illustrates a common problem with the mode: It is not a useful measure of location if there are a large number of possible values, each of which occurs infrequently. In such cases the mode will either be far from the center of the sample or, in extreme cases, will not exist, as in Example 2.9. The mode is not used much in this text because its mathematical properties are, in general, rather intractable, and in most common situations it is inferior to the arithmetic mean.

2.2.5 The Geometric Mean

Much laboratory data, specifically data in the form of concentrations of one substance in another, as assessed by serial dilution techniques, are either multiples of 2 or are a constant multiplied by a power of 2; that is, they can have outcomes only of the form $2^k c$, $k = 0, 1, \ldots$, for some constant c. For example, the data in Table 2.4 represent the minimal inhibitory concentration (MIC) of penicillin G in the urine for *N. gonorrhoeae* in 74 patients [2]. The arithmetic mean would not be appropriate as a measure of central location in this situation because the distribution is very skewed.

TABLE 2.4
Distribution of minimal inhibitory concentration (MIC) of penicillin G for *N. gonorrhoeae*

(μg/mL) Concentration	Frequency	(μg/mL) Concentration	Frequency
$0.03125 = 2^0(0.03125)$	21	$0.250 = 2^3(0.03125)$	19
$0.0625 = 2^1(0.03125)$	6	$0.50 = 2^4(0.03125)$	17
$0.125 = 2^2(0.03125)$	8	$1.0 = 2^5(0.03125)$	3

Source: Reproduced with permission from *JAMA, 220*, 205–208, 1972. Copyright 1972, American Medical Association.

However, the data do have a certain pattern since the only possible values are of the form $2^k(0.03125)$ for $k = 0, 1, 2 \ldots$. One solution is to work with the distribution of the logs of the concentrations. The log concentrations have the property that successive possible concentrations differ by a constant; that is, $\log(2^{k+1}c) - \log(2^k c) = \log(2^{k+1}) + \log c - \log(2^k) - \log c = (k + 1)\log 2 - k \log 2 = \log 2$. Thus, the log concentrations are equally spaced from each other, and the resulting distribution is now not as skewed as the concentrations themselves. The arithmetic mean could then be computed in the log scale, that is,

$$\overline{\log x} = \frac{1}{n}\sum_{i=1}^{n} \log x_i$$

and used as a measure of location. However, it is usually preferable to work in the original scale by taking the antilogarithm of $\overline{\log x}$ to form the geometric mean, which leads to the following definition:

DEFINITION 2.4

The **geometric mean** is the antilogarithm of $\overline{log\ x}$, where

$$\overline{log\ x} = \frac{1}{n}\sum_{i=1}^{n} log\ x_i$$

■

Any base can be used to compute logarithms for the geometric mean. The geometric mean will be the same regardless of which base is used. The only requirement is that the logs and antilogs in Definition 2.4 should be in the same base. Bases often used in practice are base 10 and base e; logs and antilogs using these bases can be computed using many pocket calculators.

EXAMPLE 2.10 **Infectious Disease** Compute the geometric mean for the sample in Table 2.4.

SOLUTION

1. For convenience, use base 10 to compute the logs and antilogs in this example.
2. Compute

$$\overline{\log x} = [21 \log(0.03125) + 6 \log(0.0625) + 8 \log(0.125) + 19 \log(0.250) + 17 \log(0.50) + 3 \log(1.0)]/74 = -0.846$$

3. The geometric mean = the antilogarithm of $-0.846 = 0.143$. ■■■

SECTION 2.3 Some Properties of the Arithmetic Mean

Consider a sample $x_1, \ldots, x_n$, which will be referred to as the original sample. To create a **translated sample** $x_1 + c, \ldots, x_n + c$, add a constant c to each data point. Let $y_i = x_i + c$, $i = 1, \ldots, n$. Suppose we want to compute the arithmetic mean of the translated sample. We can show that the following relationship holds:

2.1

If $y_i = x_i + c, \quad i = 1, \ldots, n$

then $\bar{y} = \bar{x} + c$

Therefore, to find the arithmetic mean of the y's, compute the arithmetic mean of the x's and add the constant c.

This principle is useful because it is sometimes convenient to change the "origin" of the sample data, that is, compute the arithmetic mean after the translation and transform back to the original origin.

EXAMPLE 2.11

In Table 2.3 it is more convenient to work with numbers that are near 0 than with numbers near 28 to compute the arithmetic mean of the time interval between menstrual periods. Thus, a translated sample might first be created by subtracting 28 days from each outcome in Table 2.3. The arithmetic mean of the translated sample could then be found and 28 added to get the actual arithmetic mean. The calculations are shown in Table 2.5.

TABLE 2.5 Translated sample for duration between successive menstrual periods in college-aged women

Value	Frequency	Value	Frequency	Value	Frequency
−4	5	1	96	6	7
−3	10	2	63	7	3
−2	28	3	24	8	2
−1	64	4	9	9	1
0	185	5	2	10	1

Note: $\bar{y} = [(-4)(5) + (-3)(10) + \cdots + (10)(1)]/500 = 0.54$

$\bar{x} = \bar{y} + 28 = 0.54 + 28 = 28.54$ days ■■■

Similarly, systolic blood-pressure scores are usually between 100 and 200. It is easy to subtract 100 from each blood-pressure score, find the mean of the translated sample, and add 100 to obtain the mean of the original sample.

What happens to the arithmetic mean if the units or scale being worked with are changed? A **rescaled sample** can be created:

$$y_i = cx_i, \quad i = 1, \ldots, n$$

The following result holds:

2.2

If $y_i = cx_i, \quad i = 1, \ldots, n$

then $\bar{y} = c\bar{x}$

Therefore, to find the arithmetic mean of the y's, compute the arithmetic mean of the x's and multiply it by the constant c.

EXAMPLE 2.12 Express the mean birthweight for the data in Table 2.1 in ounces rather than grams.

SOLUTION We know that 1 oz = 28.35 g and that $\bar{x} = 3166.9$ g. Thus, if the data were expressed in terms of ounces,

$$c = \frac{1}{28.35} \quad \text{and} \quad \bar{y} = \frac{1}{28.35}(3166.9) = 111.71 \text{ oz}$$ ■■■

Sometimes we want to change both the origin and the scale of the data at the same time. To do this, apply **(2.1)** and **(2.2)** as follows:

2.3

Let $x_1, \ldots, x_n$ be the original sample of data and let $y_i = c_1 x_i + c_2$, $i = 1, \ldots, n$, represent a transformed sample obtained by multiplying each original sample point by a factor c_1 and then shifting over by a constant c_2.

If $$y_i = c_1 x_i + c_2, \quad i = 1, \ldots, n$$

then $$\bar{y} = c_1 \bar{x} + c_2$$

EXAMPLE 2.13 If we have a sample of temperatures in °C with an arithmetic mean of 11.75°, then what is the arithmetic mean in °F?

SOLUTION Let y_i denote the °F temperature that corresponds to a °C temperature of x_i. Since the required transformation to convert the data to °F would be

$$y_i = \tfrac{9}{5}x_i + 32, \quad i = 1, \ldots, n$$

the arithmetic mean would be

$$\bar{y} = \tfrac{9}{5}(11.75) + 32 = 53.15°\text{F}$$ ■■■

SECTION 2.4 Measures of Spread

Consider the two samples shown in Figure 2.4. They represent two samples of cholesterol measurements, each on the same person, but using different measurement techniques. They appear to have about the same center, and whatever measure of central location is used will probably be about the same in the two samples. In fact, the arithmetic means are both 200 mg%/mL. However, the two samples visually appear to be radically different. This difference lies in the greater **variability**, or **spread**, of the Autoanalyzer method relative to the Microenzymatic method. In this section, the notion of variability will be quantified. Many samples can be well described by the combination of a measure of central location and a measure of spread.

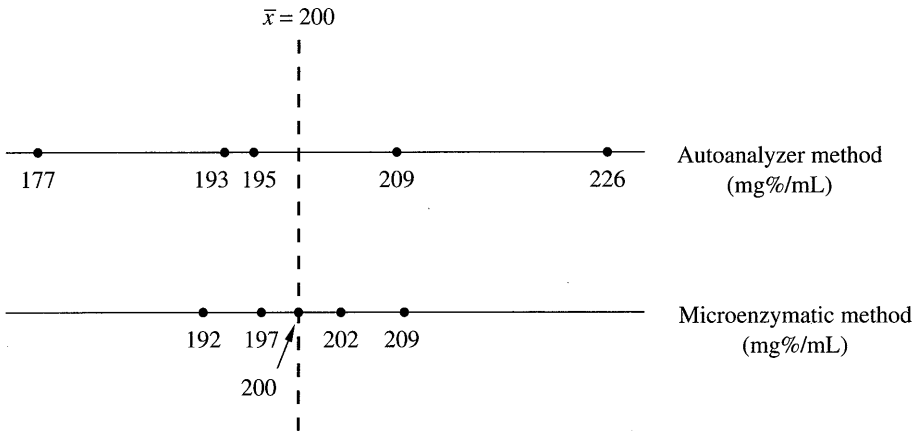

FIGURE 2.4 Two samples of cholesterol measurements on a given person using an Autoanalyzer and a Microenzymatic measurement technique

2.4.1 The Range

Several different measures can be used to describe the variability of a sample. Perhaps the simplest measure is the range.

DEFINITION 2.5

The **range** is the difference between the largest and smallest observations in the sample. ■

EXAMPLE 2.14 The range in the sample of birthweights in Table 2.1 is

$$4146 - 2069 = 2077 \text{ g}$$ ■■■

EXAMPLE 2.15 Compute the ranges for the Autoanalyzer- and Microenzymatic-method data in Figure 2.4 and compare the variability of the two methods.

SOLUTION The range for the Autoanalyzer method = 226 − 177 = 49 mg%/mL. The range for the Microenzymatic method = 209 − 192 = 17 mg%/mL. The Autoanalyzer method clearly seems more variable. ■■■

One advantage of the range is that it is very easy to compute once the sample points are ordered. One striking disadvantage is that it is very sensitive to extreme observations. Hence, if the lightest infant in Table 2.1 weighed 500 g rather than 2069 g, then the range would increase dramatically to 4146 − 500 = 3646 g. Another disadvantage of the range is that it depends on the sample size (n). That is, the larger n is, the larger the range tends to be. This complication makes it difficult to compare ranges from different sized data sets.

2.4.2 Quantiles

Another approach that addresses some of the shortcomings of the range in quantifying the spread in a data set is the use of **quantiles** or **percentiles**. Intuitively, the pth percentile is the value V_p such that p percent of the sample points are less than or equal to V_p. The median, being the 50th percentile, is a special case of a quantile. As was the case for the median, a different definition is needed for the pth percentile, depending on whether $np/100$ is an integer or not.

DEFINITION 2.6

The pth **percentile** is defined by

(1) The $(k + 1)$th largest sample point if $np/100$ is not an integer (where k is the largest integer less than $np/100$)

(2) The average of the $(np/100)$th and $(np/100 + 1)$th largest observations if $np/100$ is an integer. ■

The spread of a distribution can be characterized by specifying several percentiles. For example, the 10th and 90th percentiles are often used to characterize spread. Percentiles have the advantage over the range of being less sensitive to outliers and of not being much affected by the sample size (n).

EXAMPLE 2.16 Compute the 10th and 90th percentile for the birthweight data in Table 2.1.

SOLUTION Since $20 \times .1 = 2$ and $20 \times .9 = 18$ are integers, the 10th and 90th percentiles are defined by

10th percentile: average of the 2nd and 3rd largest values $= (2581 + 2759)/2 = 2670$ g

90th percentile: average of the 18th and 19th largest values $= (3609 + 3649)/2 = 3629$ g

We would estimate that 80 percent of birthweights will fall between 2670 g and 3629 g, which gives us an overall feel for the spread of the distribution. ■■■

EXAMPLE 2.17 Compute the 20th percentile for the white-count data in Table 2.2.

SOLUTION Since $np/100 = 9 \times .2 = 1.8$ is not an integer, the 20th percentile is defined by the $(1 + 1)$th largest value = 2nd largest value = 5000. ■■■

To compute percentiles, the sample points must be ordered. This can be difficult if n is even moderately large. An easy way to accomplish this is to use a stem-and-leaf plot (see Section 2.8.3).

There is no limit to the number of percentiles that can be computed. The most useful number is often determined by the sample size and by subject-matter considerations. Frequently used percentiles are quartiles (25th, 50th, and 75th percentiles), quintiles (20th, 40th, 60th, and 80th percentiles), and deciles (10th, 20th, . . . , 90th percentiles). It is almost always instructive to look at some of the quantiles to get an overall impression of the spread and the general shape of a distribution.

2.4.3 The Variance and Standard Deviation

The principal difference between the Autoanalyzer- and Microenzymatic-method data in Figure 2.4 is that the Microenzymatic-method values are in some sense closer to the center of the sample than the Autoanalyzer-method values are. If the center of the sample is defined as the arithmetic mean, then a measure that can summarize the difference (or deviations) between the individual sample points and the arithmetic mean, that is,

$$x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$$

is needed. One simple measure that would seem to accomplish this goal is

$$d = \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{n}$$

Unfortunately, this measure will not work because of the following principle:

2.4 The sum of the deviations of the individual observations of a sample about the sample mean is always 0.

This can be seen as follows

$$\text{Sum of deviations} = \sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x}$$

However, from the definition of the sample mean, $\sum_{i=1}^{n} x_i = n\bar{x}$. Furthermore, since $\bar{x}$ does not depend on i, $\sum_{i=1}^{n} \bar{x} = n\bar{x}$. Therefore,

$$\text{Sum of deviations} = n\bar{x} - n\bar{x} = 0$$

EXAMPLE 2.18 Compute the sum of the deviations about the mean for the Autoanalyzer- and Microenzymatic-method data in Figure 2.4.

SOLUTION For the Autoanalyzer-method data,

$$\begin{aligned} d &= (177 - 200) + (193 - 200) + (195 - 200) + (209 - 200) + (226 - 200) \\ &= -23 - 7 - 5 + 9 + 26 = 0 \end{aligned}$$

For the Microenzymatic-method data,

$$\begin{aligned} d &= (192 - 200) + (197 - 200) + (200 - 200) + (202 - 200) + (209 - 200) \\ &= -8 - 3 + 0 + 2 + 9 = 0 \end{aligned}$$

■■■

Thus, d does not help distinguish the difference in spreads between the two methods.

A second idea is to use the squares of the deviations from the sample mean rather than the deviations themselves. The resulting measure of spread, denoted by s^2 is

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

The more usual form for this measure is with $n - 1$ in the denominator rather than with n. The resulting measure is called the sample variance (or variance).

DEFINITION 2.7

The **sample variance**, or **variance**, is defined as follows:

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

■

A rationale for using $n - 1$ in the denominator rather than n is presented in the discussion of estimation in Chapter 6.

Another commonly used measure of spread is the sample standard deviation.

DEFINITION 2.8

The **sample standard deviation**, or **standard deviation**, is defined as follows:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}} = \sqrt{\text{sample variance}}$$

EXAMPLE 2.19

Compute the variance and standard deviation for the Autoanalyzer- and Microenzymatic-method data in Figure 2.4.

SOLUTION

Autoanalyzer method

$$s^2 = [(177 - 200)^2 + (193 - 200)^2 + (195 - 200)^2 + (209 - 200)^2 + (226 - 200)^2]/4$$
$$= (529 + 49 + 25 + 81 + 676)/4 = 1360/4 = 340$$
$$s = \sqrt{340} = 18.4$$

Microenzymatic method

$$s^2 = [(192 - 200)^2 + (197 - 200)^2 + (200 - 200)^2 + (202 - 200)^2 + (209 - 200)^2]/4$$
$$= (64 + 9 + 0 + 4 + 81)/4 = 158/4 = 39.5$$
$$s = \sqrt{39.5} = 6.3$$

Thus, the Autoanalyzer method has a standard deviation roughly three times as large as that of the Microenzymatic method.

One problem in using the variance is that it is difficult to compute in its original form, since the sample mean must first be computed, then the deviation of each sample point about the sample mean must be computed, and then the squares of these deviations about the sample mean must be summed. This procedure introduces two extra steps, which make the computation both more cumbersome and more error prone, especially because many pocket calculators can accumulate both the sum and sum of squares of a sample in one pass. This problem can be solved by using alternative expressions for the variance, as shown in equation 2.5.

2.5

Two alternative formulas for the sample variance,

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

are given by $\dfrac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n - 1}$ and $\dfrac{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n}{n - 1}$

To obtain these expressions, recall from algebra that $(a + b)^2 = a^2 + 2ab + b^2$, and let $a = x_i$, $b = -\bar{x}$. Then $(x_i - \bar{x})^2$ can be written in the form $x_i^2 - 2x_i\bar{x} + \bar{x}^2$. Thus, s^2 can be rewritten as follows:

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

Since $-2\bar{x}$ and $\bar{x}^2$ are constants, $\sum_{i=1}^{n}(-2x_i\bar{x})$ can be written as $-2\bar{x}\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} \bar{x}^2$ as $n\bar{x}^2$, and the following expression can be obtained:

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

It is sometimes more convenient to represent s^2 in terms of $\sum_{i=1}^{n} x_i$ rather than $\bar{x}$. To accomplish this, substitute $\left(\sum_{i=1}^{n} x_i/n\right)$ for $\bar{x}$ and obtain

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\left(\sum_{i=1}^{n} x_i/n\right)^2}{n-1}$$
$$= \frac{\sum_{i=1}^{n} x_i^2 - n\left(\sum_{i=1}^{n} x_i\right)^2/n^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2/n}{n-1}$$

The first form is useful when the sample mean has already been computed, whereas the second form is useful if the sum and sum of squares of the observations have been computed, but the sample mean has not. Thus, the sample variance can be computed directly from the sum and the sum of squares of the individual observations. The second form is actually preferable from the standpoint of computational accuracy, since rounding error is often introduced in the computation of $\bar{x}$.

Similarly, the two short forms for the standard deviation can be written as follows:

2.6

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2/n}{n-1}}$$

EXAMPLE 2.20 Compute the variance and standard deviation for the Autoanalyzer and Microenzymatic data in Figure 2.4 using the alternative computational forms.

SOLUTION **Autoanalyzer method**

$$\sum_{i=1}^{5} x_i = 177 + 193 + 195 + 209 + 226 = 1000$$

$$\sum_{i=1}^{5} x_i^2 = 177^2 + 193^2 + 195^2 + 209^2 + 226^2 = 201{,}360$$

$$s^2 = [201{,}360 - 1000^2/5]/4 = (201{,}360 - 200{,}000)/4 = 1360/4 = 340$$

$$s = \sqrt{340} = 18.4$$

Microenzymatic method

$$\sum_{i=1}^{5} x_i = 192 + 197 + 200 + 202 + 209 = 1000$$

$$\sum_{i=1}^{5} x_i^2 = 192^2 + 197^2 + 200^2 + 202^2 + 209^2 = 200{,}158$$

$$s^2 = [200{,}158 - 1000^2/5]/4 = (200{,}158 - 200{,}000)/4 = 158/4 = 39.5$$

$$s = \sqrt{39.5} = 6.3$$

Alternatively, s^2 could be represented in terms of $\bar{x}$ (= 200) by writing

$$s^2 = \frac{200{,}158 - 5(200)^2}{4}$$

$$= \frac{200{,}158 - 200{,}000}{4} = \frac{158}{4} = 39.5$$

$$s = \sqrt{39.5} = 6.3$$

■■■

SECTION 2.5 Some Properties of the Variance and Standard Deviation

The same questions can be asked of the variance and standard deviation as of the arithmetic mean: namely, how are the variance and standard deviation affected by a change in origin or a change in the units being worked with? Suppose there is a sample $x_1, \ldots, x_n$ and all data points in the sample are shifted by a constant c; that is, a new sample $y_1, \ldots, y_n$ is created such that $y_i = x_i + c$, $i = 1, \ldots, n$.

In Figure 2.5, we would clearly expect that the variance and standard deviation would remain the same, since the relationship of the points in the sample relative to one another remains the same. This property is stated on the following page.

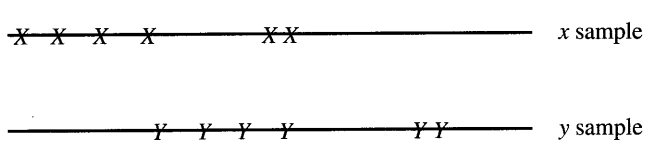

FIGURE 2.5 Comparison of the variances of two samples, where one sample has an origin shifted relative to the other

2.7 Suppose there are two samples

$$x_1, \ldots, x_n \quad \text{and} \quad y_1, \ldots, y_n$$

where $y_i = x_i + c, \quad i = 1, \ldots, n$

If the respective sample variances of the two samples are denoted by

$$s_x^2 \quad \text{and} \quad s_y^2$$

then $s_y^2 = s_x^2$

EXAMPLE 2.21 Compare the variances and standard deviations for the menstrual period data in Tables 2.3 and 2.5.

SOLUTION The variance and standard deviation of the two samples are the same, since the second sample was obtained from the first by subtracting 28 days from each data value; that is,

$$y_i = x_i - 28$$

■■■

Suppose the units are now changed so that a new sample $y_1, \ldots, y_n$ is created such that $y_i = cx_i$, $i = 1, \ldots, n$. The following relationship holds between the variances of the two samples.

2.8 Suppose there are two samples

$$x_1, \ldots, x_n \quad \text{and} \quad y_1, \ldots, y_n$$

where $y_i = cx_i, \quad i = 1, \ldots, n, \quad c > 0$

Then $s_y^2 = c^2 s_x^2 \quad s_y = cs_x$

This can be shown by noting that

$$s_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^n (cx_i - c\bar{x})^2}{n-1}$$
$$= \frac{\sum_{i=1}^n [c(x_i - \bar{x})]^2}{n-1} = \frac{\sum_{i=1}^n c^2(x_i - \bar{x})^2}{n-1}$$
$$= \frac{c^2 \sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = c^2 s_x^2$$
$$s_y = \sqrt{c^2 s_x^2} = cs_x$$

EXAMPLE 2.22 Compute the variance and standard deviation of the birthweight data in Table 2.1 in both grams and ounces.

SOLUTION The original data are given in grams, so first compute the variance and standard deviation in these units.

$$\sum_{i=1}^{20} x_i = 63{,}338 \qquad \sum_{i=1}^{20} x_i^2 = 204{,}353{,}260$$

$$s^2 = [204{,}353{,}260 - 63{,}338^2/20]/19 = 3{,}768{,}147.8/19 = 198{,}323.6 \text{ g}^2$$
$$s = 445.3 \text{ g}$$

To compute the variance and standard deviation in ounces, note that

$$1 \text{ oz} = 28.35 \text{ g} \quad \text{or} \quad y_i = \frac{1}{28.35} x_i$$

Thus
$$s^2 \text{ (oz)} = \frac{1}{28.35^2} s^2\text{(g)} = 246.8 \text{ oz}^2$$
$$s \text{ (oz)} = \frac{1}{28.35} s \text{ (g)} = 15.7 \text{ oz}$$ ■■■

Thus, if the sample points change in scale by a factor of c, the variance changes by a factor of c^2 and the standard deviation changes by a factor of c. This relationship is the main reason why the standard deviation is more often used than the variance as a measure of spread, since the standard deviation and the arithmetic mean are in the same units, whereas the variance and the arithmetic mean are not. Thus, as illustrated in Examples 2.12 and 2.22, both the mean and the standard deviation change by a factor of 28.35 in the birthweight data of Table 2.1 when the units are expressed in terms of ounces rather than grams.

The mean and standard deviation are the most widely used measures of location and spread in the literature. One of the principal reasons for this is that the normal (or bell-shaped) distribution is defined explicitly in terms of these two parameters, and this distribution has wide applicability in many biological and medical settings. The normal distribution is discussed extensively in Chapter 5.

SECTION 2.6 The Coefficient of Variation

It is useful to relate the arithmetic mean and the standard deviation together, since, for example, a standard deviation of 10 would mean something different conceptually if the arithmetic mean were 10 than if it were 1000. A special measure, called the coefficient of variation, is often used for this purpose.

DEFINITION 2.9

The **coefficient of variation** (CV) is defined by

$$100\% \times (s/\bar{x})$$ ■

This measure remains the same regardless of what units are used, because if the units are changed by a factor c, both the mean and standard deviation change by the factor c; the CV, which is the ratio between them, remains unchanged.

EXAMPLE 2.23 Compute the coefficient of variation for the data in Table 2.1 when the birthweights are expressed in either grams or ounces.

SOLUTION
$$CV = 100\% \times (s/\bar{x}) = 100\% \times (445.3 \text{ g}/3166.9 \text{ g}) = 14.1\%$$

If the data were expressed in ounces, then

$$CV = 100\% \times (15.7 \text{ oz}/111.71 \text{ oz}) = 14.1\%$$ ■■■

The coefficient of variation is most useful in comparing the variability of several different samples, each with different arithmetic means. This is because a higher variability is usually expected when the mean increases, and the *CV* is a measure that accounts for this variability. Thus, if we are conducting a study where air pollution is measured at several sites and we wish to compare day-to-day variability at the different sites, we might expect a higher variability for the more highly polluted sites. A more accurate comparison could be made by comparing the *CV*'s at different sites than by comparing the standard deviations.

The coefficient of variation is also useful for comparing the reproducibility of different variables. Consider, for example, data from the Bogalusa Heart Study, a large study of cardiovascular risk factors in children [3].

TABLE 2.6 Reproducibility of cardiovascular risk factors in children, Bogalusa Heart Study, 1978–1979

	n	Mean	sd	*CV* (%)
Height (cm)	364	142.6	0.31	0.2
Weight (kg)	365	39.5	0.77	1.9
Triceps skin fold (mm)	362	15.2	0.51	3.4
Systolic blood pressure (mm Hg)	337	104.0	4.97	4.8
Diastolic blood pressure (mm Hg)	337	64.0	4.57	7.1
Total cholesterol (mg/dL)	395	160.4	3.44	2.1
HDL cholesterol (mg/dL)	349	56.9	5.89	10.4

Children in the study were seen at approximately 3-year intervals. Every 3 years, a subset of the children had replicate measurements a short time apart of cardiovascular risk factors. In Table 2.6 we present reproducibility data on a selected subset of cardiovascular risk factors. We note that the coefficient of variation ranges from 0.2% for height to 10.4% for HDL cholesterol. The standard deviations reported here are within-subject standard deviations based on the repeated assessments on the same child. Details on how within- and between-subject variation is computed will be covered at length in Chapter 9 when we discuss the random-effects analysis of variance model.

SECTION 2.7 Grouped Data

Sometimes the sample size is prohibitively large to display all the raw data. Also, data are frequently collected in grouped form, since the required degree of accuracy to specify a measured quantity exactly is often lacking, because of either measurement error or imprecise patient recall. For example, systolic blood-pressure measurements taken with a standard cuff are usually specified to the nearest 2 mm Hg, since assessing them with any more precision is difficult using this instrument. Thus, a stated measurement of 120 mm Hg may actually imply that the reading is some number ≥ 119 mm Hg and <121 mm Hg. Similarly, because dietary recall is generally not very accurate, the most precise estimate of fish consumption might take the following form: 2–3 servings per day, 1 serving per day, 5–6 servings per week, 2–4 servings per week, 1 serving per week, <1 serving per week and ≥ 1 serving per month, never.

TABLE 2.7
Sample of birthweights from 100 consecutive deliveries (oz)

58	118	92	108	132	32	140	138	96	161
120	86	115	118	95	83	112	128	127	124
123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113
121	68	107	122	126	88	89	108	115	85
111	121	124	104	125	102	122	137	110	101
91	122	138	99	115	104	98	89	119	109
104	115	138	105	144	87	88	103	108	109
128	106	125	108	98	133	104	122	124	110
133	115	127	135	89	121	112	135	115	64

Consider the data set in Table 2.7, which represents the birthweights from 100 consecutive deliveries at a Boston hospital. Suppose we wish to display these data for publication purposes. How can we do this? If the data are on a computer, then the simplest way to display the data would be to generate a frequency distribution using one of the common statistical packages.

DEFINITION 2.10

A **frequency distribution** is an ordered display of each value in a data set together with its **frequency**, that is, the number of times that value occurs in the data set. In addition, the percentage of sample points that take on a particular value is also typically given. ■

A frequency distribution of the sample of 100 birthweights in Table 2.7 was generated using the Statistical Analysis System (SAS) package and is displayed in Table 2.8.

The SAS frequency-distribution program provides the frequency, cumulative frequency (CUM FREQ), relative frequency (PERCENT), and cumulative percent (CUM PERCENT) for each birthweight present in the sample. For any particular birthweight b, the cumulative frequency, or CUM FREQ, is the number of birthweights in the sample that are less than or equal to b. The PERCENT = 100 × FREQUENCY/n, while the cumulative percent (CUM PERCENT) = 100 × CUM FREQ/n = the percentage of birthweights less than or equal to b.

If the number of unique sample values is large, then a frequency distribution may still be too detailed a summary for publication purposes. Instead, the data could be grouped into broader categories. Some general instructions for categorizing the data are provided in the following guidelines:

1. Subdivide the data into k intervals, starting at some lower bound y_1 and ending at some upper bound y_{k+1}.

2. The first interval is from y_1 inclusive to y_2 exclusive; the second interval is from y_2 inclusive to y_3 exclusive; . . . ; the kth and last interval is from y_k inclusive to y_{k+1} exclusive. The rationale for this representation is to make certain that the group intervals include all possible values *and* do not overlap. These errors are common in the presentation of grouped data.

3. The group intervals are generally chosen to be equal, although the appropriateness of equal group sizes should be dictated more by subject-matter considerations. Thus, equal intervals might be appropriate for the blood-pressure or birthweight data but not for the dietary-recall data, where the nature of the data dictates unequal group sizes corresponding to how most people remember what they eat.

TABLE 2.8
Frequency, distribution of birthweight data in Table 2.7 using the Statistical Analysis System (SAS)

SAMPLE OF BIRTHWEIGHTS FROM 100 CONSECUTIVE DELIVERIES (OZ.)

BIRTHWT	FREQUENCY	CUM FREQ	PERCENT	CUM PERCENT
32	1	1	1.000	1.000
58	1	2	1.000	2.000
64	1	3	1.000	3.000
67	1	4	1.000	4.000
68	1	5	1.000	5.000
83	1	6	1.000	6.000
85	2	8	2.000	8.000
86	1	9	1.000	9.000
87	1	10	1.000	10.000
88	2	12	2.000	12.000
89	3	15	3.000	15.000
91	1	16	1.000	16.000
92	1	17	1.000	17.000
93	1	18	1.000	18.000
94	2	20	2.000	20.000
95	1	21	1.000	21.000
96	1	22	1.000	22.000
98	3	25	3.000	25.000
99	1	26	1.000	26.000
100	1	27	1.000	27.000
101	1	28	1.000	28.000
102	1	29	1.000	29.000
103	1	30	1.000	30.000
104	5	35	5.000	35.000
105	2	37	2.000	37.000
106	1	38	1.000	38.000
107	1	39	1.000	39.000
108	4	43	4.000	43.000
109	2	45	2.000	45.000
110	2	47	2.000	47.000
111	1	48	1.000	48.000
112	3	51	3.000	51.000
113	1	52	1.000	52.000
115	6	58	6.000	58.000
116	1	59	1.000	59.000
118	2	61	2.000	61.000
119	1	62	1.000	62.000
120	1	63	1.000	63.000
121	3	66	3.000	66.000
122	4	70	4.000	70.000
123	1	71	1.000	71.000
124	4	75	4.000	75.000
125	2	77	2.000	77.000
126	1	78	1.000	78.000
127	2	80	2.000	80.000
128	2	82	2.000	82.000
132	3	85	3.000	85.000
133	2	87	2.000	87.000
134	1	88	1.000	88.000
135	2	90	2.000	90.000
137	1	91	1.000	91.000
138	3	94	3.000	94.000
140	1	95	1.000	95.000
141	1	96	1.000	96.000
144	1	97	1.000	97.000
146	1	98	1.000	98.000
155	1	99	1.000	99.000
161	1	100	1.000	100.000

4. A count is made of the number of units that fall in each interval, which is denoted by the frequency within that interval.

5. The midpoint of each group interval is computed for calculation of descriptive statistics. The midpoint of the first interval is denoted by

$$m_1 = \frac{y_1 + y_2}{2}$$

the midpoint of the second interval by

$$m_2 = \frac{y_2 + y_3}{2}, \ldots$$

and the midpoint of the last interval by

$$m_k = \frac{y_k + y_{k+1}}{2}$$

The intervals and their midpoints are depicted in Figure 2.6.

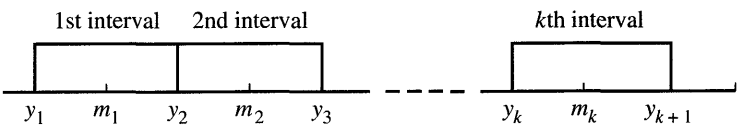

FIGURE 2.6 Subdivision of the real line for the purpose of forming group intervals

6. Finally, for the purpose of computing descriptive statistics, the group intervals and their midpoints, m_i, and frequencies, f_i, are then displayed concisely in a table such as Table 2.9.

TABLE 2.9 General layout of grouped data

Group interval	Midpoint of group interval	Frequency
$\geqslant y_1, < y_2$	m_1	f_1
$\geqslant y_2, < y_3$	m_2	f_2
⋮	⋮	⋮
$\geqslant y_i, < y_{i+1}$	m_i	f_i
⋮	⋮	⋮
$\geqslant y_k, < y_{k+1}$	m_k	f_k

For example, the raw data in Table 2.7 might be displayed according to the format in Table 2.10.

TABLE 2.10 Grouped frequency distribution of birthweight (oz) from 100 consecutive deliveries

Group interval	Midpoint	Frequency
$\geq 29.5, < 69.5$	49.5	5
$\geq 69.5, < 89.5$	79.5	10
$\geq 89.5, < 99.5$	94.5	11
$\geq 99.5, < 109.5$	104.5	19
$\geq 109.5, < 119.5$	114.5	17
$\geq 119.5, < 129.5$	124.5	20
$\geq 129.5, < 139.5$	134.5	12
$\geq 139.5, < 169.5$	154.5	6
		100

If we are confronted with grouped data either in the form of published data from a secondary source or from our own data, then we want to be able to compute grouped means and variances that are analogous to the arithmetic mean and variance. Suppose that f_i observations fall in the ith group interval, $i = 1, \ldots, k$, and that the midpoint of the ith interval is m_i, $i = 1, \ldots, k$, where $n = \sum_{i=1}^{k} f_i$ = total number of observations over all groups. The grouped mean is then defined as follows:

DEFINITION 2.11

The **grouped mean** is defined by

$$\bar{x}_g \equiv \frac{\sum_{i=1}^{k} f_i m_i}{\sum_{i=1}^{k} f_i}$$

EXAMPLE 2.24

Compute the grouped mean for the data in Table 2.10.

SOLUTION

$$\bar{x}_g = \frac{\sum_{i=1}^{k} f_i m_i}{\sum_{i=1}^{k} f_i}$$

$$= [5(49.5) + 10(79.5) + 11(94.5) + \cdots + 6(154.5)]/100$$

$$= 11{,}045/100 = 110.45 \text{ oz}$$

DEFINITION 2.12

The **grouped variance** is defined by

$$s_g^2 = \frac{\sum_{i=1}^{k} f_i (m_i - \bar{x}_g)^2}{\left(\sum_{i=1}^{k} f_i\right) - 1}$$

As for the ungrouped variance, the expression for the grouped variance can be simplified, yielding the following two short forms:

2.9 Short Forms for the Grouped Variance

$$\text{Grouped variance} = \frac{\sum_{i=1}^{k} f_i m_i^2 - n\bar{x}_g^2}{n-1} = \frac{\sum_{i=1}^{k} f_i m_i^2 - \left(\sum_{i=1}^{k} f_i m_i\right)^2 / n}{n-1}$$

EXAMPLE 2.25 Compute the grouped variance for the data in Table 2.10.

SOLUTION
$$\sum_{i=1}^{k} f_i m_i^2 = 5(49.5)^2 + 10(79.5)^2 + \cdots + 6(154.5)^2 = 1{,}274{,}355$$

Thus,
$$s_g^2 = (1{,}274{,}355 - 11{,}045^2/100)/99$$
$$= 54{,}434.75/99 = 549.85$$

and
$$s_g = \sqrt{549.85} = 23.45 \text{ oz}$$ ■■■

SECTION 2.8 Graphic Methods for Grouped Data

In Section 2.7 we concentrated on methods for presenting grouped data in tabular form and on numerical measures for describing such data. In this section these techniques are supplemented by presenting certain commonly used graphic methods for displaying grouped data. The purpose of using graphic displays is to give a quick overall impression of the data, which is sometimes difficult to obtain with numerical measures.

2.8.1 Bar Graphs

One of the most widely used methods for displaying grouped data is the bar graph.

A **bar graph** can be constructed as follows:

(1) The data are divided in a number of groups using the guidelines provided in Section 2.7.

(2) For each group a rectangle is constructed with a base of a constant width and a height proportional to the frequency within that group.

(3) The rectangles are generally not contiguous and are equally spaced from each other.

A bar graph of daily vitamin-A consumption among 200 cancer cases and 200 age- and sex-matched controls is presented in Figure 2.1.

2.8.2 Histograms

The bar graph tends to work well with grouped data when the groups are characterized by nonnumerical attributes, such as {current smoker/ex-smoker/never smoker} or {patient gets worse/patient gets better/patient stays the same}. If the groups are characterized by a numerical attribute, such as systolic blood pressure or birthweight, then a histogram is preferable. For a histogram, the position of the rectangle will correspond to the location of the group interval along the x-axis, and the size of the rectangle will correspond to the frequency within the group.

A **histogram** is constructed as follows:

(1) The data are divided into groups as described in Section 2.7.

(2) A rectangle is constructed for each group. The location of the base of the rectangle corresponds to the position of the ends of the group interval along the x-axis, and the **area** of the rectangle is proportional to the frequency within the group.

(3) The scale used along either axis should allow all the rectangles to fit into the space allotted for the graph.

Note that the area, rather than the height, is proportional to the frequency. If the length of each group interval is the same, then the area and the height are in the same proportions and the height will be proportional to the frequency as well. However, if one group interval is 5 times as long as another and the two group intervals have the same frequency, then the first group interval should have a height $\frac{1}{5}$ as high as the second group interval so that the areas will be the same. A common mistake in the literature is to construct histograms with group intervals of different lengths but with the height proportional to the frequency. This representation gives a misleading impression of the data. A histogram for the birthweight data in Table 2.10 is given in Figure 2.7.

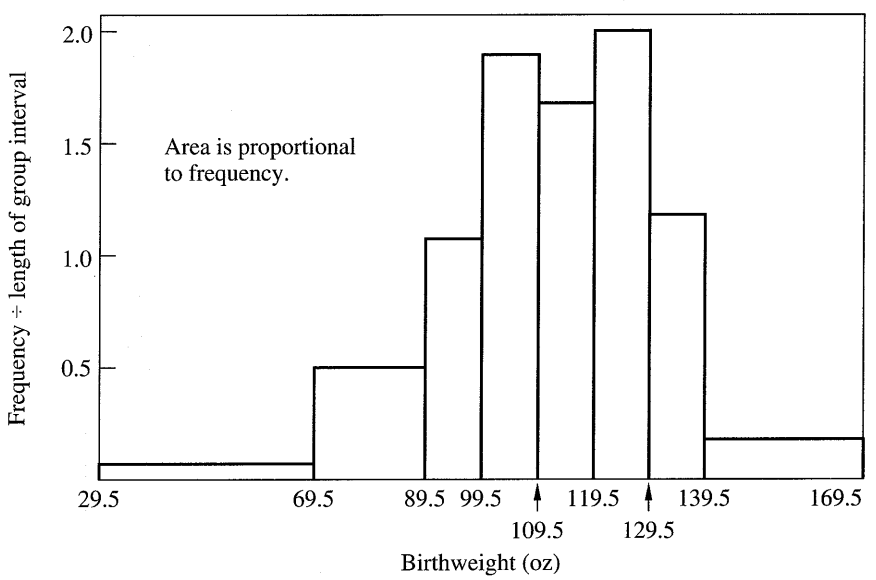

FIGURE 2.7 Histogram for the birthweight data in Table 2.10

2.8.3 Stem-and-Leaf Plots

Two problems with histograms are that (1) they are somewhat difficult to construct and (2) the sense of what the actual sample points are within the respective groups is lost. One type of graphic display that overcomes these problems is the stem-and-leaf plot.

A **stem-and-leaf** plot can be constructed as follows:

(1) Separate each data point into a stem component and a leaf component, respectively, where the stem component consists of the number formed by all but the rightmost digit of the number, and the leaf component consists of the rightmost digit. Thus, the stem of the number 483 is 48, and the leaf is 3.

(2) Write the smallest stem in the data set in the upper-left-hand corner of the plot.

(3) Write the second stem, which equals the first stem + 1, below the first stem.

(4) Continue with step 3 until you reach the highest stem in the data set.

(5) Draw a vertical bar to the right of the column of stems.

(6) For each number in the data set, find the appropriate stem and write the leaf to the right of the vertical bar.

The collection of leaves thus formed will take on the general shape of the distribution of the sample points. Furthermore, the actual sample values are preserved and yet there is a grouped display for the data, which is a distinct advantage over a histogram. Finally, a stem-and-leaf plot can usually be constructed more quickly than a histogram from raw data, since the number of data points in each group interval do not have to be counted. It is also easy to compute the median and the range from a stem-and-leaf plot. A stem-and-leaf plot is given in Figure 2.8 for the birthweight data in Table 2.7. Thus, the point 5|8 represents 58, 11|8 represents 118, and so forth. Notice how this plot gives an overall feel for the distribution without losing the individual values.

FIGURE 2.8
Stem-and-leaf plot for the birthweight data (oz) in Table 2.7

3	2
4	
5	8
6	7 8 4
7	
8	6 3 5 8 9 5 9 7 8 9
9	2 6 5 4 8 3 4 1 9 8 8
10	8 5 0 4 7 8 4 2 1 4 9 4 5 3 8 9 6 8 4
11	8 5 8 2 2 6 3 5 1 0 5 9 5 0 5 2 5
12	0 8 7 4 3 4 1 2 6 1 4 5 2 2 8 5 2 4 7 1
13	2 8 4 2 2 7 8 8 3 3 5 5
14	0 1 6 4
15	5
16	1

There are variations of stem-and-leaf plots where the leaf can consist of more than one digit. This variation might be appropriate for the birthweight data in Table 2.1, since the number of three-digit stems required would be very large relative to the

number of data points. In this case, the leaf would consist of the rightmost two digits and the stem the leftmost two digits, and the pairs of digits to the right of the vertical bar would be underlined to distinguish between two different leaves. The stem-and-leaf display for the data in Table 2.1 is presented in Figure 2.9.

FIGURE 2.9
Stem-and-leaf plot for the birthweight data (g) in Table 2.1

20	69
21	
22	
23	
24	
25	81
26	
27	59
28	41 38 34
29	
30	31
31	01
32	65 60 45 00 48
33	23 14
34	84
35	41
36	49 09
37	
38	
39	
40	
41	46

Another common variation on the ordinary stem-and-leaf plot if the number of leaves is large is to allow more than one line for each stem. Similarly, one can position the largest stem at the top of the plot and the smallest stem at the bottom of the plot. In Figure 2.10 some graphic displays using the SAS UNIVARIATE procedure are given to illustrate this technique.

Notice that each stem is allowed two lines, with the leaves from 5 to 9 on the upper line and the leaves from 0 to 4 on the lower line. Furthermore, the leaves are ordered on each line, and a count of the number of leaves on each line is provided under the # column to allow easy computation of the median and other quantiles. Thus, the number 7 in the # column on the upper line for stem 12 indicates that there are 7 birthweights from 125 to 129 oz in the sample, whereas the number 13 indicates that there are 13 birthweights from 120 to 124 oz. Finally, a multiplication factor is

FIGURE 2.10
Stem-and-leaf and box plots for the birthweight data (oz) in Table 2.7 as generated by the SAS UNIVARIATE procedure

```
STEM LEAF                  #        BOXPLOT
  16 1                     1           |
  15 5                     1           |
  15                                   |
  14 6                     1           |
  14 014                   3           |
  13 557888                6           |
  13 222334                6           |
  12 5567788               7           |
  12 0111222234444        13     +-----+
  11 5555556889           10     |     |
  11 0012223               7     +--+--+
  10 5567888899           10     |     |
  10 012344444             9     |     |
   9 568889                6     +-----+
   9 12344                 5           |
   8 556788999             9           |
   8 3                     1           |
   7                                   |
   7                                   |
   6 78                    2           |
   6 4                     1           |
   5 8                     1           0
   5
   4
   4
   3
   3 2                     1           0
     ----+----+----+----+
     MULTIPLY STEM.LEAF BY 10**+01
```

given in the bottom of the display to allow for the representation of decimal numbers in stem-and-leaf form. In particular, if no multiplication factor (m) is present, then it is assumed that all numbers have actual value stem.leaf; whereas if m is present, then the actual value of the number is assumed to be stem.leaf $\times 10^m$. Thus, for example, since the multiplication factor is 10^1, the value 6 4 on the stem-and-leaf plot represents the number $6.4 \times 10^1 = 64$ oz.

2.8.4 Box Plots

In Section 2.2.3 the comparison of the arithmetic mean and the median was discussed as a method for looking at the skewness of a distribution. This goal can also be accomplished by a graphic technique known as the **box plot**. To describe a box plot, the concept of the **hinges** of a sample must be introduced. To accomplish this, it is first necessary to understand the notion of the depth of the median.

DEFINITION 2.13

The **depth** (m) of the median for a sample of size n is

(1) $\frac{n}{2}$ if n is even

(2) $\frac{n+1}{2}$ if n is odd ■

The upper and lower hinges can be thought of conceptually as the approximate 75th and 25th percentiles of the sample, that is, the points $\frac{3}{4}$ and $\frac{1}{4}$ along the way in the ordered sample.

DEFINITION 2.14

The **upper hinge** of a sample is

(1) The $\frac{m+1}{2}$th largest point if m is odd

(2) The average of the $\frac{m}{2}$th and $\left(\frac{m}{2}+1\right)$th largest points if m is even

where m = depth of the median. The **lower hinge** is defined similarly, starting from the smallest point in the sample. ■

EXAMPLE 2.26

Compute the upper and lower hinges for the birthweight data in Table 2.7.

SOLUTION

Since $n = 100$, it follows that the depth of the median $(m) = 50$. Since m is even, the upper hinge is given by the average of the $\frac{50}{2}$th and $\left(\frac{50}{2}+1\right)$th largest sample values or the average of the 25th- and 26th-largest points in the sample. In the stem-and-leaf plot in Figure 2.10, counting down from the top, $1 + 1 + 1 + 3 + 6 + 6 + 7 = 25$ points are in the upper 12 row or above. Thus, the 25th-largest point is the smallest number in the upper 12 row, which equals 125 oz. Also, the 26th-largest point = largest number in the lower 12 row = 124 oz. Thus, the upper hinge = $(125 + 124)/2 = 124.5$ oz.

Similarly, the lower hinge = the average of the 25th and 26th smallest points in the sample. Counting up from the bottom, $1 + 1 + 1 + 2 + 1 + 9 + 5 = 20$ points are in the lower 9 row or below, and 26 points are in the upper 9 row or below. Thus, the 25th smallest point = the 2nd largest value in the upper 9 row = 98; the 26th smallest point = the largest value in the upper 9 row = 99. Therefore, the lower hinge = $(98 + 99)/2 = 98.5$ oz. ■■■

How can the hinges be used to judge the symmetry of a distribution?

(1) If the distribution is symmetric, then the upper and lower hinges should be approximately equally spaced from the median.

(2) If the upper hinge is farther from the median than the lower hinge, then the distribution is positively skewed.

(3) If the lower hinge is farther from the median than the upper hinge, then the distribution is negatively skewed.

These relationships are illustrated graphically in a box plot. In Figure 2.10 the top of the box corresponds to the upper hinge, whereas the bottom of the box corresponds to the lower hinge. A horizontal line is also drawn at the median value. Furthermore, in the SAS implementation of the box plot, the sample mean is indicated by a + sign.

EXAMPLE 2.27

What can be learned about the symmetry properties of the distribution of birthweights from the box plot in Figure 2.10?

SOLUTION

In Figure 2.10, because the lower hinge is farther from the median than the upper hinge, the distribution is slightly negatively skewed. This pattern is true of many birthweight distributions. ■■■

In addition to displaying the symmetry properties of a sample, a box plot can also be used to give a feel for the spread of a sample and can help identify possible outlying values, that is, values that seem inconsistent with the rest of the points in the sample. In the context of box plots, outlying values are defined as follows:

DEFINITION 2.15

An **outlying value** is a value x such that either

(1) $x >$ upper hinge $+ 1.5 \times$ (upper hinge $-$ lower hinge) or

(2) $x <$ lower hinge $- 1.5 \times$ (upper hinge $-$ lower hinge) ■

DEFINITION 2.16

An **extreme outlying value** is a value x such that either

(1) $x >$ upper hinge $+ 3.0 \times$ (upper hinge $-$ lower hinge) or

(2) $x <$ lower hinge $- 3.0 \times$ (upper hinge $-$ lower hinge)

The box plot is then completed by

(1) Drawing a vertical bar from the upper hinge to the largest nonoutlying value in the sample

(2) Drawing a vertical bar from the lower hinge to the smallest nonoutlying value in the sample

(3) Individually identifying the outlying and extreme outlying values in the sample by 0's and *'s, respectively ■

EXAMPLE 2.28 Using the box plot in Figure 2.10, comment on the spread of the sample in Table 2.7 and the presence of outlying values.

SOLUTION Since the upper and lower hinges are 124.5 and 98.5 oz, respectively, an outlying value x must satisfy the following relations:

$$x > 124.5 + 1.5 \times (124.5 - 98.5) = 124.5 + 39.0 = 163.5$$

or

$$x < 98.5 - 1.5 \times (124.5 - 98.5) = 98.5 - 39.0 = 59.5$$

Similarly, an extreme outlying value x must satisfy the following relations:

$$x > 124.5 + 3.0 \times (124.5 - 98.5) = 124.5 + 78.0 = 202.5$$

or

$$x < 98.5 - 3.0 \times (124.5 - 98.5) = 98.5 - 78.0 = 20.5$$

Thus, the values 32 and 58 oz are outlying values but not extreme outlying values. These values are identified by 0's on the box plot. A vertical bar extends from 64 oz (the smallest nonoutlying value) to the lower hinge and from 161 oz (the largest nonoutlying value = the largest value in the sample) to the upper hinge. The accuracy of the two identified outlying values should probably be checked. ■■■

The methods used to identify outlying values in Definitions 2.15 and 2.16 are descriptive. Alternative methods for identifying outliers based on a hypothesis-testing framework are given in Chapter 8.

Many more details on stem-and-leaf plots, box plots, and other exploratory data methods are given in Tukey [4].

SECTION 2.9 Case Study: Effects of Lead Exposure on Neurological and Psychological Function in Children

A study was performed [5] of the effects of exposure to lead on the psychological and neurological well-being of children. The complete raw data for this study are provided in Data Set LEAD.DAT, and the documentation for this file is given in Data Set LEAD.DOC. All Data Sets are on the data disk.

In summary, a group of children who lived near a lead smelter in El Paso, Texas, were identified and their blood levels of lead were measured. An exposed group of 46 children were identified who had blood-lead levels ≥ 40 μg/ml in 1972 (or in a few cases in 1973). This group is defined by the variable GROUP = 2 or 3. A control group of 78 children were also identified who had blood-lead levels < 40 μg/ml in both 1972 and 1973. This group is defined by the variable GROUP = 1. All children lived in close proximity to the lead smelter.

Two key outcome variables studied were (1) the number of finger-wrist taps in the dominant hand (a measure of neurological function) and (2) the Wechsler full-scale IQ score. To explore the relationship of lead exposure to the outcome variables, we used the SAS UNIVARIATE procedure to obtain box plots for these two variables for children in the exposed and control groups, respectively. These are given in Figures 2.11 and 2.12, respectively. For this purpose, since the dominant hand was not identified in the data base, we used the larger of the finger-wrist tapping scores for the right and left hand as a proxy for the number of finger-wrist taps in the dominant hand.

We note that although there is considerable spread within each group, both finger-wrist tapping scores (MAXFWT) and full-scale IQ scores (IQF) seem to be lower in

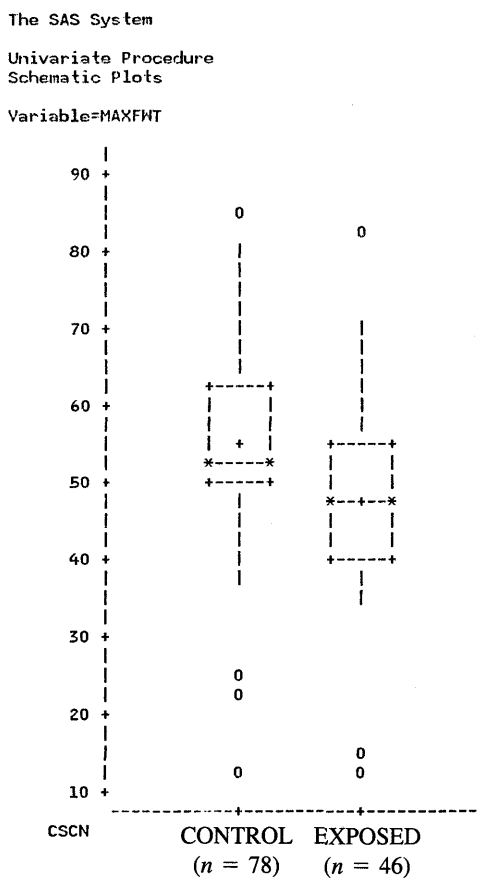

FIGURE 2.11
Number of finger-wrist taps in the dominant hand for the exposed and control groups, El Paso Lead Study

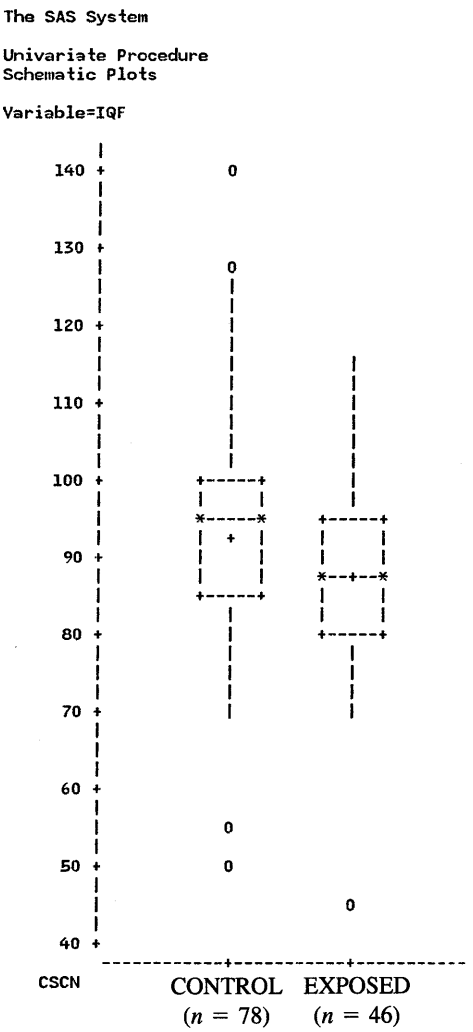

FIGURE 2.12
Wechsler full-scale IQ scores for the exposed and control groups, El Paso Lead Study

the exposed group than the control group. We will be analyzing these data in more detail in subsequent chapters, using t tests, analysis of variance, and regression methods.

SECTION 2.10 Summary

In this chapter several **numeric and graphic methods for describing data** for the purpose of

(1) quickly summarizing a data set and for

(2) presenting results to others

were presented.

In general, a data set can be described numerically in terms of a **measure of location** and a **measure of spread**. Several alternatives were introduced for each of these measures, including the **arithmetic mean**, **median**, **mode**, and **geometric mean** as possible choices for measures of location, and the **standard deviation**, **quantiles**, and **range** as possible choices for measures of spread. Criteria were discussed for choosing the appropriate measures in particular circumstances. Several graphic techniques for summarizing data, including traditional methods, such as the **bar graph** and **histogram**, and some more modern methods characteristic of exploratory data analysis (EDA), such as the **stem-and-leaf plot** and **box plot**, were introduced.

How do the descriptive methods in this chapter fit in with the methods of statistical inference discussed later in this book? Specifically, if, based on some prespecified hypotheses, some interesting trends using descriptive methods can be found, then we need some method to judge how "significant" these trends are. For this purpose several commonly used **probability models** are introduced in Chapters 3 through 5 and approaches for testing the validity of these models using the methods of **statistical inference** are explored in Chapters 6 through 13.

PROBLEMS

Infectious Disease

The data in Table 2.11 are a sample from a larger data set collected on persons discharged from a selected Pennsylvania hospital as part of a retrospective chart review of antibiotic usage in hospitals [6]. The data are also given in Data Set HOSPITAL.DAT with documentation in HOSPITAL.DOC on the data disk.

2.1 Compute the mean and median for the duration of hospitalization for the 25 patients.

2.2 Compute the standard deviation and range for the duration of hospitalization for the 25 patients.

2.3 It is of clinical interest to know if the duration of hospitalization is affected by whether or not a patient has received antibiotics. Can you answer this question using either numeric or graphic methods?

Suppose the scale for a data set is changed by multiplying each observation by a positive constant.

* **2.4** What is the effect on the median?

* **2.5** What is the effect on the mode?

* **2.6** What is the effect on the geometric mean?

* **2.7** What is the effect on the range?

* Asterisk indicates that the answer to the problem is given in the Answer Section at the back of the book.

Ophthalmology

Table 2.12 comes from a paper giving the distribution of astigmatism in 1033 young men, aged 18–22, who were accepted for military service in Great Britain [7]. Assume that astigmatism is rounded to the nearest 10th of a diopter.

TABLE 2.12 Distribution of astigmatism in 1033 young men aged 18–22

Degree of astigmatism (diopters)	Frequency
0.0 or less than 0.2	458
0.2–0.3	268
0.4–0.5	151
0.6–1.0	79
1.1–2.0	44
2.1–3.0	19
3.1–4.0	9
4.1–5.0	3
5.1–6.0	2
	1033

Source: Reprinted with permission of the Editor, the authors and the Journal from the *British Medical Journal*, May 7, 1394–1398, 1960.

TABLE 2.11 Hospital-stay data

ID no.	Duration of hospital stay	Age	Sex (1 = M, 2 = F)	First temp. following admission	First WBC ($\times 10^3$) following admission	Received anti-biotic (1 = yes, 2 = no)	Received bacterial culture (1 = yes, 2 = no)	Service (1 = med., 2 = surg.)
1	5	30	2	99.0	8	2	2	1
2	10	73	2	98.0	5	2	1	1
3	6	40	2	99.0	12	2	2	2
4	11	47	2	98.2	4	2	2	2
5	5	25	2	98.5	11	2	2	2
6	14	82	1	96.8	6	1	2	2
7	30	60	1	99.5	8	1	1	1
8	11	56	2	98.6	7	2	2	1
9	17	43	2	98.0	7	2	2	1
10	3	50	1	98.0	12	2	1	2
11	9	59	2	97.6	7	2	1	1
12	3	4	1	97.8	3	2	2	2
13	8	22	2	99.5	11	1	2	2
14	8	33	2	98.4	14	1	1	2
15	5	20	2	98.4	11	2	1	2
16	5	32	1	99.0	9	2	2	2
17	7	36	1	99.2	6	1	2	2
18	4	69	1	98.0	6	2	2	2
19	3	47	1	97.0	5	1	2	1
20	7	22	1	98.2	6	2	2	2
21	9	11	1	98.2	10	2	2	2
22	11	19	1	98.6	14	1	2	2
23	11	67	2	97.6	4	2	2	1
24	9	43	2	98.6	5	2	2	2
25	4	41	2	98.0	5	2	2	1

2.8 Compute the grouped mean.

2.9 Compute the grouped standard deviation.

2.10 Plot a histogram to properly illustrate these data.

Cardiovascular Disease

The data in Table 2.13 are a sample of cholesterol levels taken from 24 hospital employees who were on a standard American diet and who agreed to adopt a vegetarian diet for 1 month. Serum-cholesterol measurements were made before adopting the diet and 1 month after.

* **2.11** Compute the mean change in cholesterol.

* **2.12** Compute the standard deviation of the change in cholesterol levels.

2.13 Construct a stem-and-leaf plot of the cholesterol changes.

* **2.14** Compute the median change in cholesterol.

2.15 Construct a box plot of the cholesterol changes to the right of the stem-and-leaf plot.

2.16 Comment on the symmetry of the distribution of change scores based on your answers to Problems 2.11 through 2.15.

2.17 Some investigators feel that the effects of diet on cholesterol are more evident in people with high rather than low cholesterol levels. If you split the data in Table 2.13 according to whether baseline cholesterol is above or below the median, can you comment on this issue?

TABLE 2.13 Serum-cholesterol levels before and after adopting a vegetarian diet

Subject	Before	After	Before–after
1	195	146	49
2	145	155	−10
3	205	178	27
4	159	146	13
5	244	208	36
6	166	147	19
7	250	202	48
8	236	215	21
9	192	184	8
10	224	208	16
11	238	206	32
12	197	169	28
13	169	182	−13
14	158	127	31
15	151	149	2
16	197	178	19
17	180	161	19
18	222	187	35
19	168	176	−8
20	168	145	23
21	167	154	13
22	161	153	8
23	178	137	41
24	137	125	12

Hypertension

An experiment was performed to look at the effect of position on level of blood pressure [8]. In the experiment 32 subjects had their blood pressures measured while lying down with their arms at their sides and again standing with their arms supported at heart level. The data are given in Table 2.14.

2.18 Compute the arithmetic mean and median for the difference in systolic and diastolic blood pressure, respectively, between the positions (recumbent and standing).

2.19 Construct stem-and-leaf and box plots for the difference scores for each type of blood pressure.

2.20 Based on your answers to Problems 2.18 and 2.19, comment on the effect of position on the levels of systolic and diastolic blood pressure.

TABLE 2.14 Effect of position on blood pressure

	Blood pressure (mm Hg)			
	Recumbent, arm at side		Standing, arm at heart level	
Subject				
B. R. A.	99[a]	71[b]	105[a]	79[b]
J. A. B.	126	74	124	76
F. L. B.	108	72	102	68
V. P. B.	122	68	114	72
M. F. B.	104	64	96	62
E. H. B.	108	60	96	56
G. C.	116	70	106	70
M. M. C.	106	74	106	76
T. J. F.	118	82	120	90
R. R. F.	92	58	88	60
C. R. F.	110	78	102	80
E. W. G.	138	80	124	76
T. F. H.	120	70	118	84
E. J. H.	142	88	136	90
H. B. H.	118	58	92	58
R. T. K.	134	76	126	68
W. E. L.	118	72	108	68
R. L. L.	126	78	114	76
H. S. M.	108	78	94	70
V. J. M.	136	86	144	88
R. H. P.	110	78	100	64
R. C. R.	120	74	106	70
J. A. R.	108	74	94	74
A. K. R.	132	92	128	88
T. H. S.	102	68	96	64
O. E. S.	118	70	102	68
R. E. S.	116	76	88	60
E. C. T.	118	80	100	84
J. H. T.	110	74	96	70
F. P. V.	122	72	118	78
P. F. W.	106	62	94	56
W. J. W.	146	90	138	94

[a]Systolic blood pressure
[b]Diastolic blood pressure
Source: Reprinted with permission of the *American Journal of Medicine*.

Pulmonary Disease

FEV (forced expiratory volume) is an index of pulmonary function that measures the volume of air expelled after one second of constant effort. The Data Set FEV.DAT (on the data disk) contains determinations of FEV in 1980 on 654 children ages 3–19 who were seen in the Childhood Respiratory Disease Study (CRD Study) in East Boston, Massachusetts. These data are part of a longitudinal study

to follow the change in pulmonary function over time in children [9].

The data in Table 2.15 are available for each child.

2.21 For each variable (other than ID), obtain appropriate descriptive statistics (both numeric and graphic).

2.22 Use both numeric and graphic measures to assess the relationship of FEV to age, height, and smoking status (Do this separately for boys and girls.)

2.23 Compare the pattern of growth of FEV by age for boys and girls. Are there any similarities? Any differences?

2.24 Answer Problem 2.23 for height rather than FEV.

TABLE 2.15 Format for FEV.DAT

Column	Variable	Format or code
1–5	ID number	
7–8	Age (years)	
10–15	FEV (liters)	X.XXX
17–20	Height (inches)	XX.X
22	Sex	0 = female/1 = male
24	Smoking status	0 = noncurrent smoker/1 = current smoker

Nutrition

The food-frequency questionnaire (FFQ) is an instrument that is often used in dietary epidemiology to assess consumption of specific foods. A person is asked to write down the number of servings per day typically eaten in the past year of over 100 individual food items. A food-composition table is then used to compute nutrient intakes (e.g., protein, fat, etc.), based on aggregating responses for individual foods. The FFQ is inexpensive to administer but is considered less accurate than the diet record (DR) (the gold standard of dietary epidemiology). For the diet record, a participant writes down the amount of each specific food eaten over the past week in a food diary and a nutritionist using a special computer program computes nutrient intakes from the food diaries. This is a much more expensive method of dietary recording. To validate the FFQ, 173 nurses participating in the Nurses Health Study completed 4 weeks of diet recording about equally spaced over a 12-month period and an FFQ at the end of diet recording [10]. Data are presented in the Data Set VALID.DAT (on the data disk) for saturated fat, total fat, total alcohol consumption, and total caloric intake for both the DR and FFQ. For the DR, average nutrient intakes were computed over the 4 weeks of diet recording. The format of this file is shown in Table 2.16.

2.25 Compute appropriate descriptive statistics for each nutrient for both DR and FFQ using both numeric and graphic measures.

TABLE 2.16 Format for VALID.DAT

Column	Variable	Format or code
1–6	ID number	XXXXX.XX
8–15	Saturated fat–DR	XXXXX.XX
17–24	Saturated fat–FFQ	XXXXX.XX
26–33	Total fat–DR	XXXXX.XX
35–42	Total fat–FFQ	XXXXX.XX
44–51	Alcohol consumption–DR	XXXXX.XX
53–60	Alcohol consumption–FFQ	XXXXX.XX
62–70	Total calories–DR	XXXXXX.XX
72–80	Total calories–FFQ	XXXXXX.XX

2.26 Use descriptive statistics to relate nutrient intake for the DR and FFQ. Do you think that the FFQ is a reasonably accurate approximation to the DR? Why or why not?

2.27 A frequently used method for quantifying dietary intake is in the form of quintiles. Compute quintiles for each nutrient and each method of recording and relate the nutrient composition for DR and FFQ using the quintile scale. (That is, how does the quintile category based on DR relate to the quintile category based on FFQ for the same individual?) Do you get the same impression about the concordance between DR and FFQ using quintiles as in Problem 2.26, where raw (ungrouped) nutrient intake is considered?

Environmental Health, Pediatrics

In Section 2.9, we described Data Set LEAD.DAT (on the data disk) concerning the effect of lead exposure on neurological and psychological function in children.

2.28 Compare the exposed and control groups on age and gender, using appropriate numeric and graphic descriptive measures.

2.29 Compare the exposed and control groups on verbal and performance IQ, using appropriate numeric and graphic descriptive measures.

2.30 Did your answer to problem 2.28 influence how you approached Problem 2.29? If so, in what way?

References

[1] White, J. R., & Froeb, H. E. (1980). Small-airways dysfunction in nonsmokers chronically exposed to tobacco smoke. *New England Journal of Medicine, 302*(33), 720–723.

[2] Pedersen, A., Wiesner, P., Holmes, K., Johnson, C., & Turck, M. (1972). Spectinomycin and Penicillin G in the treatment of gonorrhea. *JAMA, 220*(2), 205–208.

[3] Foster, T. A., & Berenson, G. (1987). Measurement error and reliability in four pediatric cross-sectional surveys of cardiovascular disease risk factor variables—the Bogalusa Heart Study. *Journal of Chronic Disease, 40*(1): 13–21.

[4] Tukey, J. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

[5] Landrigan, P. J., Whitworth, R. H., Baloh, R. W., Staehling, N. W., Barthel, W. F., & Rosenblum, B. F. (1975, March 29). Neuropsychological dysfunction in children with chronic low-level lead absorption. *Lancet* 1, 708–715.

[6] Townsend, T. R., Shapiro, M., Rosner, B., & Kass, E. H. (1979). Use of antimicrobial drugs in general hospitals I. Description of population and definition of methods. *Journal of Infectious Diseases, 139*(6), 688–697.

[7] Sorsby, A., Sheridan, M., Leary, G. A., & Benjamin, B. (1960). Vision, visual acuity and ocular refraction of young men in a sample of 1033 subjects. *British Medical Journal*, 1394–1398.

[8] Kossmann, C. E. (1946). Relative importance of certain variables in clinical determination of blood pressure. *American Journal of Medicine, 1*, 464–467.

[9] Tager, I. B., Weiss, S. T., Rosner, B., & Speizer, F. E. (1979). Effect of parental cigarette smoking on pulmonary function in children. *American Journal of Epidemiology, 110*, 15–26.

[10] Willett, W. C., Sampson, L., Stampfer, M. J., Rosner, B., Bain, C., Witschi, J., Hennekens, C. H., & Speizer, F. E. (1985). Reproducibility and validity of a semi-quantitative food frequency questionnaire. *American Journal of Epidemiology, 122*, 51–65.