CHAPTER 6

DESCRIPTIVE STATISTICS

Thus far we have considered random variables from a theoretical point of view. We have studied two functions, the density and the cumulative distribution function, that enable us to predict the behavior of the variable in a probabilistic sense. We have also considered three parameters that characterize or describe a random variable, namely, μ, σ^2, and σ. In practice, the exact distribution of a random variable is seldom known. Rather, we must determine a reasonable form for the density and appropriate values for the distribution parameters from a data set. In this chapter we consider some simple graphical and analytic methods for doing so.

6.1 RANDOM SAMPLING

We begin by considering a typical problem that calls for a statistical solution. Suppose that we wish to study the performance of the lithium batteries used in a particular model of pocket calculator. The purpose of our study is to determine the mean effective life span of these batteries so that we can place a limited warranty on them in the future. Since this type of battery has not been used in this model before, no one can tell us the distribution of the random variable, X, the life span of a battery. We must attempt to discover its distribution for ourselves. This is inherently a statistical problem. What characteristics identify it as such? Simply the following:

Characteristics of a Statistical Problem

1. Associated with the problem is a large group of objects about which inferences are to be made. This group of objects is called the *population.*
2. There is at least one random variable whose behavior is to be studied relative to the population.
3. The population is too large to study in its entirety, or techniques used in the study are destructive in nature. In either case we must draw conclusions about

the population based on observing only a portion or "sample" of objects drawn from the population.

In our example the population is large and hypothetical in the sense that it consists of all lithium batteries used in this model calculator in the past, present, and future. Since we cannot observe the life span of batteries not yet produced, the population obviously cannot be studied in its entirety! Furthermore, to determine the life span of a battery, it must be used until it fails. That is, the method of study destroys the object being studied. For these reasons, we must devise methods for approximating the characteristics of the life span of a lithium battery based on observing only a sample of these batteries.

To draw inferences about a population using statistical methods, the sample drawn should be "random." To understand what we mean by this term, let us return to our example. Here we have a large population that consists of all lithium batteries produced for a certain model of pocket calculator. Associated with the population is a random variable X. We do not know the form of its density, nor do we know its mean or variance. We want to select a subset of n batteries from the population "at random." That is, we want to select n batteries for study in such a way that the selection of one battery neither ensures nor precludes the selection of any other. In this way the selection of one battery is independent of the selection of any other. This collection of objects can be thought of as a "random sample."

Note that, prior to the actual selection of the batteries to be studied, X_i ($i = 1, 2, 3, \ldots, n$), the life span of the ith battery selected is a random variable. It has the same distribution as X, the life span of batteries in the population. Furthermore, these random variables are independent in the sense that the value assumed by one has no effect on the value assumed by any of the others. The random variables $X_1, X_2, X_3, \ldots, X_n$ and can be thought of as a "random sample."

Once we have actually selected n batteries for study and have observed the life span of each battery, we shall have available n numbers, $x_1, x_2, x_3, \ldots, x_n$. These numbers are the observed values of the random variables $X_1, X_2, X_3, \ldots, X_n$ and can be thought of as a "random sample."

As you can see, the term "random sample" is used in three different but closely related ways in applied statistics. It may refer to the *objects* selected for study, to the *random variables* associated with the objects to be selected, or to the *numerical values* assumed by those variables. It is usually clear from the context of the discussion which is intended. These ideas are illustrated in Fig. 6.1.

Even though the term "random sample" is used in these three ways, the formal definition of the term is mathematical in nature. When we use the term in stating theoretical results, we mean the following:

Definition 6.1.1 (Random sample). A random sample of size n from the distribution of X is a collection of n independent random variables, each with the same distribution as X.

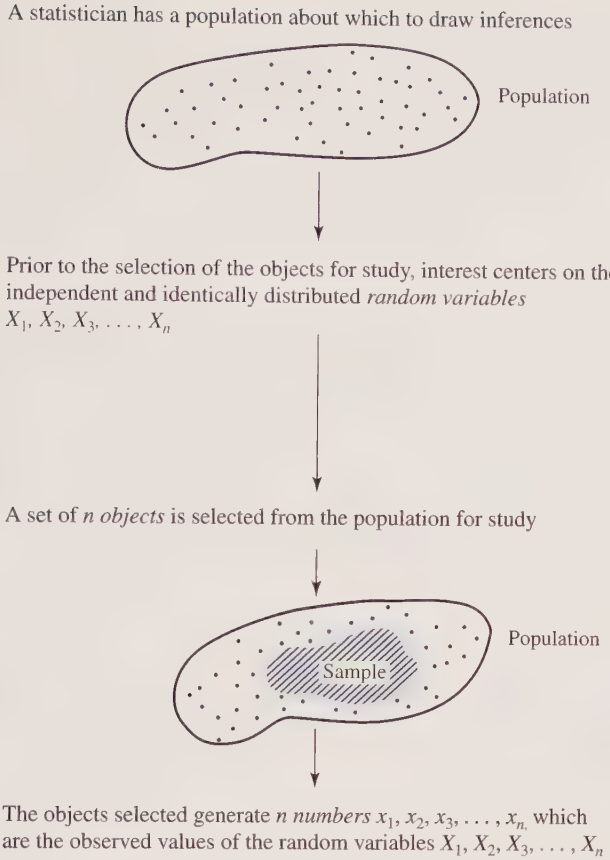

FIGURE 6.1
The objects selected generate *n numbers* $x_1, x_2, x_3, \ldots, x_n$, which are the observed values of the random variables $X_1, X_2, X_3, \ldots, X_n$.

The theorems and definitions presented later use the term "random sample" in the sense just described. When objects are selected from a finite population, this type of sample results only when sampling is done with replacement. That is, an object is drawn, observed, and placed back in the population for possible reselection. This ensures that $X_1, X_2, X_3, \ldots, X_n$ are indeed independent and identically distributed. Usually, sampling from a finite population is done without replacement. This means that the random variables $X_1, X_2, X_3, \ldots, X_n$ are not independent. However, if the sample is small relative to the population itself, then removal of a few items does not drastically alter the composition of the population. A generally accepted guideline is that for all practical purposes we may assume independence whenever the sample constitutes at most 5% of the population. If this is not true, then the techniques used to estimate parameters must be altered to take this into account. We

shall be assuming that for all practical purposes $X_1, X_2, X_3, \ldots, X_n$ are independent in the discussions that follow.

Once a random sample has been drawn, we commonly use the data gathered to evaluate pertinent *statistics*. What is a statistic? Roughly speaking, a statistic is a random variable whose numerical value can be determined from a random sample. That is, a statistic is a random variable that is a function of the elements of a random sample $X_1, X_2, X_3, \ldots, X_n$. Typical statistics of interest to statisticians are $\sum_{i=1}^{n} X_i$, $\sum_{i=1}^{n} X_i^2$, $\sum_{i=1}^{n} X_i/n$, $\max_i\{X_i\}$, and $\min_i\{X_i\}$. These ideas are illustrated in Example 6.1.1.

Example 6.1.1. Consider the random variable X, the number of times per hour that a television signal is interrupted by random interference. Assume that this random variable has a Poisson distribution with unknown mean μ and unknown variance σ^2. To approximate the value of each of these parameters, we intend to observe the signal for ten randomly selected nonoverlapping one-hour periods over a week's time. Let X_i ($i = 1, 2, 3, \ldots, 10$) denote the number of interruptions that occur during the ith observation period. The random variables $X_1, X_2, X_3, \ldots, X_{10}$ constitute a random sample of size 10 from a Poisson distribution with unknown mean μ and unknown variance σ^2. When the experiment is conducted, these data result:

$$\begin{array}{lllll} x_1 = 1 & x_3 = 0 & x_5 = 1 & x_7 = 0 & x_9 = 3 \\ x_2 = 0 & x_4 = 2 & x_6 = 1 & x_8 = 0 & x_{10} = 0 \end{array}$$

The observed values of the statistics ΣX_i, ΣX_i^2, $\Sigma X_i/n$, $\max_i\{X_i\}$, and $\min_i\{X_i\}$ based on this sample are 8, 16, .8, 3, and 0, respectively. Note that the random variable $X_1 - \mu$ is *not* a statistic. Since μ is unknown, we cannot determine its numerical value from a random sample.

6.2 PICTURING THE DISTRIBUTION

When studying a random variable X, one important question to be answered is, "To which family of random variables does X belong?" That is, we need to determine whether X is binomial, Poisson, normal, exponential, or belongs to some other family of variables. In the discrete case it is often possible to determine the appropriate family from the physical description of the experiment. The only job left for the statistician is to approximate the values of the parameters that characterize the distribution. Continuous random variables are more difficult to handle. To determine the family to which such a variable belongs, we must get an idea of the *shape* of its density. For example, if the density appears to be flat, then it is reasonable to suspect that X is uniformly distributed; if it is bell-shaped, then X may be normally distributed.

If the distribution appears to be nonsymmetric with a long tail to the left or the right, then it is called *skewed* left or skewed right, respectively. Distributions such as the exponential, chi-squared, and gamma distributions exhibit this property. For example, see Fig. 4.4. In each case the distribution pictured is skewed to the right.

Stem-and-Leaf Diagram

Here we consider some graphical methods for studying the distribution of a continuous random variable. The first method entails constructing what is called a *stem-and-leaf* diagram. This method was first introduced by John Tukey in 1977 [50].

A stem-and-leaf diagram consists of a series of horizontal rows of numbers. Each row is labeled via a number called its stem; the other numbers in the rows are called *leaves*. There are no rigid rules as to how to construct such a diagram. Basically these steps are followed:

Constructing a Stem-and-Leaf Diagram

1. Choose some convenient numbers to serve as stems. The stems are usually the first one or two digits of the numbers in the data set.
2. Label the rows via the stems selected.
3. Reproduce the data set graphically by recording the digit following the stem as a leaf.
4. Turn the graph on its side to get an idea of the shape of the distribution.

These ideas are illustrated in Example 6.2.1.

Example 6.2.1. To study the random variable X, the life span in hours of the lithium battery in a particular model of pocket calculator, we obtain a random sample of 50 batteries and determine the life span of each we obtain. These data result:

4285	564	1278	205	3920
2066	604	209	602	1379
2584	14	349	3770	99
1009	4152	478	726	510
318	737	3032	3894	582
1429	852	1461	2662	308
981	1560	701	497	3367
1402	1786	1406	35	99
1137	520	261	2778	373
414	396	83	1379	454

To construct a stem-and-leaf diagram for these data, we first choose numbers to serve as "stems." It is often convenient to use the first digit of a number as its stem. If a three-digit number such as 318 is expressed as a four-digit number (0318) by including a leading zero, then this data set entails the use of the five stems 0, 1, 2, 3, 4. We shall use the second digit of a number as its "leaf." The diagram is constructed by listing the stems as a vertical column as shown in Fig. 6.2(a). The first observation, 4285, has a stem of 4 and a leaf of 2. It is represented in the diagram as shown in Fig. 6.2(b). The entire data set, recorded in the order in which the observations appear, is shown in Fig. 6.2(c).

Is it reasonable to assume that X is normally distributed? To answer this question, turn the stem-and-leaf diagram on its side and look for the bell-shape characteristic of

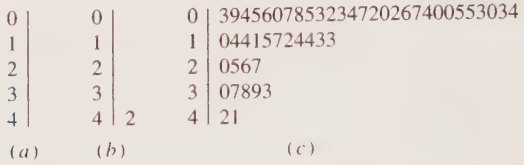

FIGURE 6.2
(*a*) The integers 0, 1, 2, 3, 4 form the stems for a stem-and-leaf diagram; (*b*) the number 4285 has a stem of 4 and a leaf of 2; (*c*) complete stem-and-leaf diagram for the sample of battery life spans of Example 6.2.1.

```
Stem-and-leaf of hours      N   = 50
Leaf Unit = 1.0

  17     0 00000222333334444
 (11)    0 55556677789
  22     1 012334444
  13     1 57
  11     2 0
  10     2 567
   7     3 03
   5     3 789
   2     4 12
```

FIGURE 6.3
A double stem-and-leaf diagram with leaves in order.

a normal density. This bell shape is not present, leading us to suspect that X is *not* a member of the family of normal random variables.

Notice that, in the above example, the first stem has a very large number of leaves. This often occurs when data sets are large or when there is not much variability in the data. In this case it is usually constructive to create what is called a *double* stem-and-leaf diagram. This is done by using each stem twice. We plot the low leaves of 0, 1, 2, 3, 4 on the first stem and the high leaves of 5, 6, 7, 8, 9 on the second. The double stem-and-leaf diagram for the data of Example 6.2.1 is shown in Fig. 6.3. This diagram was produced by MINITAB. This diagram shows even more clearly than that of Fig. 6.2 that the distribution from which this sample was drawn is probably not normal. In fact, it resembles a distribution that is exponential. We know now that a reasonable density for X assumes the general form

$$f(x) = (1/\beta)\exp(-1/\beta) \qquad x > 0 \qquad \beta > 0$$

It is now the job of the researcher to estimate the numerical value of β so that probabilities can be estimated in the future via the exponential density.

Histograms and Ogives

The stem-and-leaf diagram provides a quick look at a data set. It is a useful way to get an idea of the shape of a distribution when the data set is moderate in size. It has

TABLE 6.1
Suggested number of categories to be used in subdividing numeric data as a function of sample size

Sample size	Number of categories
Fewer than 16	Not enough data
16–31	5
32–63	6
64–127	7
128–255	8
256–511	9
512–1023	10
1024–2047	11
2048–4095	12
4096–8190	13

the advantage of preserving, to some extent, the ability to read the actual data values from the diagram. However, the technique does not work well when data sets are large. In this case, we turn to a technique that has been used for many years and that is often seen in data displays in journals, newspapers, corporate reports, and other presentations. This plot, called a *histogram,* is a vertical or horizontal bar graph. The bars or categories are defined in such a way that each observation belongs to one and only one category. We make the width of each bar the same so that the area of the bar is proportional to the number of observations in the respective category. This allows for easy visual comparisons of category frequencies and percentages. It also allows us to get an idea of the family of random variables to which the variable under study belongs by observing the shape of the histogram.

There are many ways to select category boundaries. Statistical packages each use their own algorithm for doing so, and these may differ from package to package. If several different packages are used to plot a given data set via its default technique, then the histograms can vary slightly in terms of number of categories chosen and category boundary values. They will all give the same general impression of shape.

We present here an algorithm for selecting the number of categories and category boundaries. This algorithm will guarantee that each data point falls into exactly one category, that categories are the same width, and that no data point can assume a boundary value. Some computer packages allow the user to select the number of categories or to specify boundary values. If so, then this algorithm can be used to control the construction of the histogram if desired.

Rules for Breaking Data into Categories

1. Decide on the number of categories wanted. The number chosen depends on the number of observations available. Table 6.1 gives suggested guidelines for the number of categories to be used as a function of sample size. It is based on Sturges' rule, a formula developed by H. A. Sturges in 1926.

TABLE 6.2
Units and half units for data reported to the stated degree of accuracy

Data reported to nearest	Unit	1/2 unit
Whole number	1	.5
Tenth (1 decimal place)	.1	.05
Hundredth (2 decimal places)	.01	.005
Thousandth (3 decimal places)	.001	.000 5
Ten thousandth (4 decimal places)	.000 1	.0000 5

2. Locate the largest observation and the smallest observation.
3. Find the difference between the largest and the smallest observations. Subtract in the order of the largest minus the smallest. This difference is called the *range* of the data.
4. Find the minimum length required to cover this range by dividing the range by the number of categories desired. This length is the minimum length required to cover the range if the lower boundary for the first category is taken to be the smallest data point. However, to ensure that no data point falls on a boundary, we shall define boundaries in such a way that they involve one more decimal place than the data. Hence we shall start the first category slightly *below* the first data point. By doing this, the minimum category length required to cover the range is not long enough to trap the largest data point in the last category. For this reason, the actual length used must be a little longer than minimum.
5. The actual category length to be used is found by rounding the minimum length *up* to the same number of decimal places as the data itself. If the minimum length by chance already has the same number of decimal places as the data, we shall round up 1 unit. For example, if we have data reported to one decimal place accuracy and the minimum length required to cover the range is found to be 1.7, we bump this up to 1.8 to obtain the actual category length to be used.
6. The lower boundary for the first category lies 1/2 unit below the smallest observation. Table 6.2 gives units and half units for various types of data sets.
7. The remaining category boundaries are found by adding the category length to the preceding boundary value.

Example 6.2.2. Consider the data of Example 6.2.1. The data set has 50 observations. From Table 6.1 we see that the suggested number of categories to be used is 6. Now we locate the largest data point (4285) and the smallest (14). These are used to find the range, that is, the length of the interval containing all the data points. In this case the data are covered by an interval of length $4285 - 14 = 4271$ units. To find the minimum length required for each category, we divide this number by the number of categories desired. Here the minimum category length is $4271/6 \doteq 711.83$ units. To find the actual category length to be used in splitting the data, we round up the minimum length to the same number of decimal places as the data. Here the data are reported in whole numbers. Thus we round up the minimum length, 711.83, to the nearest whole number, 712. The categories actually used will be of length 712. The

TABLE 6.3

Category	Boundaries	Frequency	Relative frequency
1	13.5 to 725.5	24	24/50 = 48%
2	725.5 to 1437.5	12	12/50 = 24%
3	1437.5 to 2149.5	4	4/50 = 8%
4	2149.5 to 2861.5	3	3/50 = 6%
5	2861.5 to 3573.5	2	2/50 = 4%
6	3573.5 to 4285.5	5	5/50 = 10%

first category starts 1/2 unit below the smallest observation. From Table 6.2 we see that 1/2 unit is .5 in the case of integer data. That is, the lower boundary for the first category is $14 - .5 = 13.5$. The remaining category boundaries are found by successively adding the category length (712) to the preceding boundary until all data points are covered. In this way we obtain the following six finite categories for the battery lives:

13.5 to 725.5	2149.5 to 2861.5
725.5 to 1437.5	2861.5 to 3573.5
1437.5 to 2149.5	3573.5 to 4285.5

Note that since the boundaries have one more decimal place than the data, no data point can fall on a boundary; each data point must fall into exactly one category. The data can be summarized now in table form by recording the number (frequency) and the percentage (relative frequency) of the observations in each category, as shown in Table 6.3. From this table we can construct a histogram of the data. If the frequency per category is plotted along the vertical axis, the resulting bar graph is called a *frequency histogram;* if the vertical axis is used to plot the relative frequency per category, then the diagram is called a *relative frequency histogram.* Both plots provide a visual display of the data that conveys an idea of the shape of the density of the random variable X under study. The relative frequency histogram for the data of Example 6.2.1 is shown in Fig. 6.4. Since the histogram does not exhibit a bell shape, we see once again that these data do not support an assumption of normality. In fact, the distribution suggested by the data is the exponential distribution. In this case it is now the job of the researcher to estimate β, the parameter that describes this distribution. By so doing, we are able to estimate the density for X. This estimated density can then be used to approximate probabilities in the future.

Figure 6.5 shows the histogram produced by MINITAB's default settings. Notice that more categories and different boundaries are chosen by the computer algorithm than is the case with the textbook procedure. We still get the same impression of a distribution that is skewed to the right.

Cumulative Distribution Plots (Ogives)

In addition to the frequency distribution among categories, it is of interest to consider the cumulative frequency distribution of the observations. The cumulative

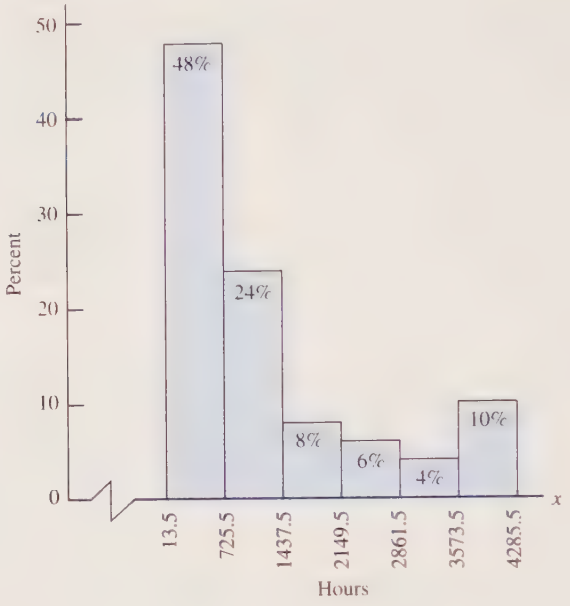

FIGURE 6.4
Relative frequency histogram for the sample of battery life spans of Example 6.2.1.

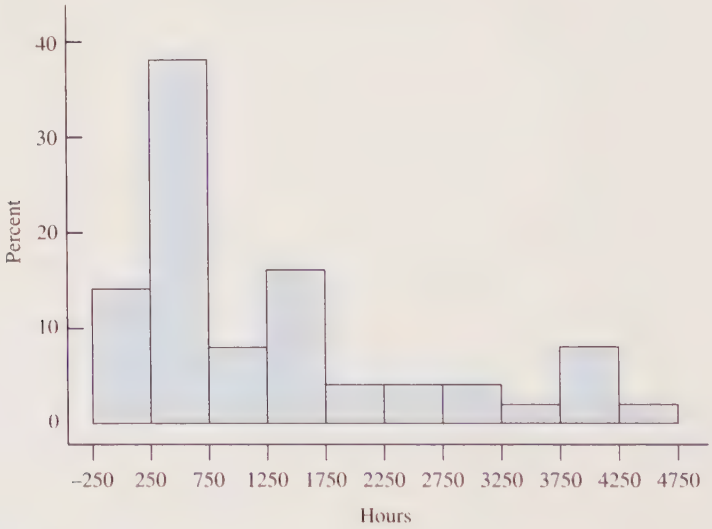

FIGURE 6.5
Histogram produced via MINITAB default settings.

frequency distribution is found by determining for each category the number and percentage of observations falling in or below that category. The cumulative distribution of the data of Example 6.2.1 is shown in Table 6.4.

TABLE 6.4

Category	Boundaries	Frequency	Cumulative frequency	Relative cumulative frequency
1	13.5 to 725.5	24	24	24/50 = 48%
2	725.5 to 1437.5	12	36	36/50 = 72%
3	1437.5 to 2149.5	4	40	40/50 = 80%
4	2149.5 to 2861.5	3	43	43/50 = 86%
5	2861.5 to 3573.5	2	45	45/50 = 90%
6	3573.5 to 4285.5	5	50	50/50 = 100%

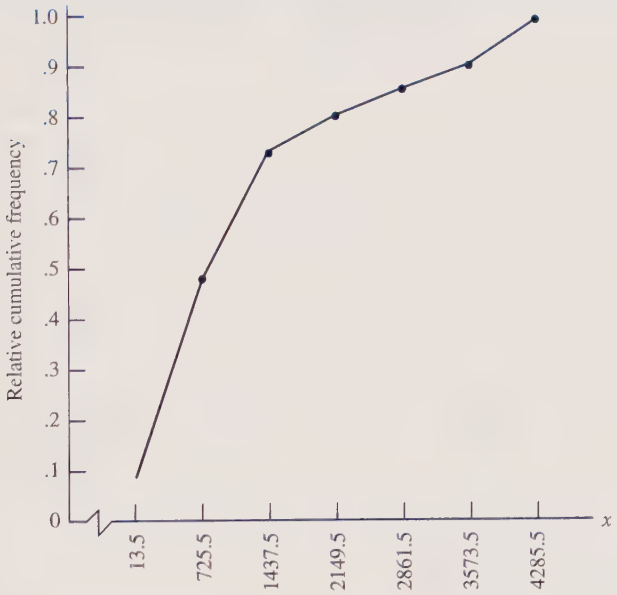

FIGURE 6.6
Relative cumulative frequency ogive for the sample of battery life spans of Example 6.2.1.

When the random variable under study is continuous, the cumulative distribution can be used to construct a graph that approximates its cumulative distribution function F. The graph is a line graph obtained by plotting the upper boundary of each category on the horizontal axis against the relative cumulative frequency. This type of graph is called a *relative cumulative frequency ogive*. The ogive for the data of Example 6.2.1 is shown in Fig. 6.6. From the ogive we can answer questions such as, "Approximately what percentage of batteries fail during the first 1500 hours of operation?" and "What time represents the midway point in the sense that half the batteries fail on or before this time?"

The first question can be answered graphically by locating 1500 on the horizontal axis, projecting a vertical line up to the ogive, and then projecting a horizontal

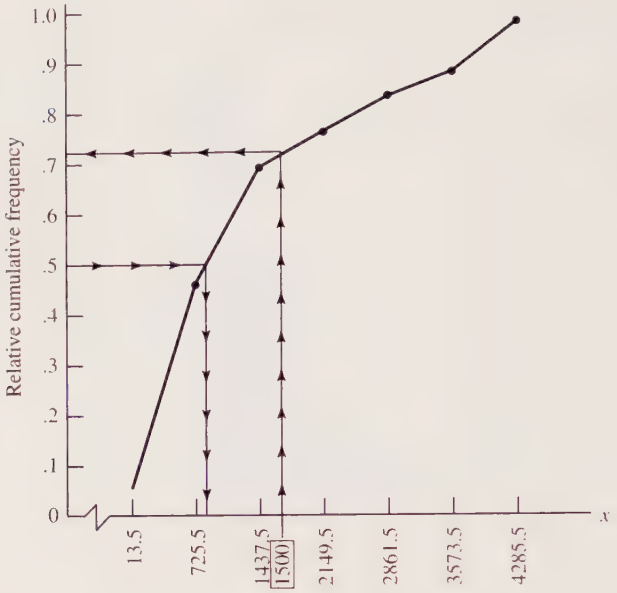

FIGURE 6.7
Projective method of approximating probabilities using a relative cumulative frequency ogive.

line over to the vertical axis, as shown in Fig. 6.7. The desired percentage is seen to be approximately 72%. The second question is answered by locating .5 on the vertical axis and reversing the process. The answer is seen to be a little over 725 hours. (See Fig. 6.7.)

6.3 SAMPLE STATISTICS

We have seen that the behavior of a random variable X is determined by its density. We have also seen that the parameters μ, the theoretical average value of the random variable, and σ^2, its variability about the mean, are helpful in describing X. In the last section we considered some graphical methods for getting an idea of the shape of the density. In this section we consider some statistics that allow us to summarize a data set analytically. Since it is hoped that the data set reflects the population as a whole, these statistics also give us some idea of the values of the parameters that characterize X over the population under study. In particular, we consider two measures of location or central tendency in a data set, the *sample mean* and the *sample median*. We also consider three measures of variability within the data set, the *sample variance*, the *sample standard deviation*, and the *sample range*. The word "sample" is used to emphasize the fact that the data sets presented are based on experiments involving only a small portion of objects that constitute the population being studied. That is, they represent a random sample from the distribution of X.

Location Statistics

The mean or theoretical average value of X is our primary measure of the center of location of X. The primary measure of the center of location of a data set is its arithmetic average. Since we view a data set as a set of observations on X, the arithmetic average for a particular set of observations is just the observed value of the *statistic* $\sum_{i=1}^{n} X_i/n$. This statistic, called the *sample mean,* is defined formally in the next definition.

Definition 6.3.1 (Sample mean). Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample from the distribution of X. The statistic $\sum_{i=1}^{n} X_i/n$ is called the sample mean and is denoted by $\bar{X}$.

Note that μ_X and $\bar{X}$ are *not* the same. The parameter μ_X is the theoretical average value for X over the entire population; $\bar{X}$ is a statistic which, when evaluated over a particular random sample, gives the average value of X *for that sample.* It is hoped, of course, that the observed value of $\bar{X}$ is close to μ_X. In reporting sample means, we shall usually retain one more decimal place than that of the data. Rounding will be used rather than truncation.

Example 6.3.1. A random sample of size 9 yields the following observations on the random variable X, the coal consumption in millions of tons by electric utilities for a given year:

406 395 400 450 390 410 415 401 408

The observed value of the sample mean for these data is

$$\bar{x} = \sum_{i=1}^{n} x_i/n = (406 + 395 + 400 + \cdots + 408)/9$$
$$= 3675/9 \doteq 408.3 \text{ million tons}$$

The average value for X for this sample is 408.3 million tons. What is the average number of tons of coal used by electric utilities across the country in this particular year? That is, What is μ_X? Unfortunately, this question cannot be answered with certainty from this sample. However, the sample leads us to believe that μ_X lies close to 408.3 million tons. Admittedly, the word "close" is a bit vague. In Chap. 8 we shall consider a method for determining how close μ_X is likely to be to 408.3 million tons.

A second measure of the center of location of a random variable X is its *median.* The median of a random variable is its 50th percentile (see Exercise 12). That is, the median for X is that number M such that

$$P[X < M] \leq .50 \qquad \text{and} \qquad P[X \leq M] \geq .50$$

If X is continuous, then its median is the "halfway point" in the sense that an observation on X is just as likely to fall below M as it is to fall above it. We define the median for a sample with this in mind.

Definition 6.3.2. Let $x_1, x_2, \ldots, x_n$ be a sample of observations arranged in order from the smallest to the largest. The sample median is the middle observation if n is odd. It is the average of the two middle observations if n is even. We shall denote the median of a sample by $\tilde{x}$.

If n is small, it is easy to spot the middle of a data set. However, if n is large, it is useful to have a formula that pinpoints the location of the middle observation or observations. The formula is given below, and its use is illustrated in Example 6.3.2.

$$\text{Median location} = \frac{n+1}{2}$$

Example 6.3.2. The nine observations on X, the coal consumption in millions of tons by electric utilities for a given year, arranged in order, are

390 395 400 401 406 408 410 415 450

The median location is $\frac{n+1}{2} = \frac{9+1}{2} = 5$. The median is the fifth data point in the ordered list. In this case, $\tilde{x} = 406$. This observation is the middle value in our ordered list. Note that this is the median for this data set. It gives us a *rough* idea of the median coal consumption across the country during the year.

Measures of Variability

Recall that we are usually concerned not only with the mean of a random variable, but also with its variance. The variance of a random variable, given by

$$\sigma^2 = E[(X-\mu)^2]$$

measures the variability of X about the population mean. We want to develop an analogous measure of variability within a sample. To do so, we parallel the logic used in defining σ^2. We do not know the value of the population mean, but we shall have available an observed value for the sample mean. We cannot observe the differences $(X-\mu)^2$ for all members of the population, but we can observe the difference $(X_i - \bar{X})^2$ for each element X_i of the random sample. Since σ^2 is an expectation, a theoretical average value, logic dictates that we replace this operation by an arithmetic average of sample values. That is, the natural measure of variability within a sample that parallels our definition of variability within the population is

$$\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n}$$

This method of measuring variability within a sample is acceptable. In fact, many electronic calculators with built-in statistical capability utilize this formula to compute the variance of a sample. In most cases we shall be using the variability in the sample to approximate σ^2. However, it can be shown that this statistic tends, on the average, to underestimate σ^2. To improve the situation, we divide $\Sigma_{i=1}^{n}(X_i - \bar{X})^2$ by

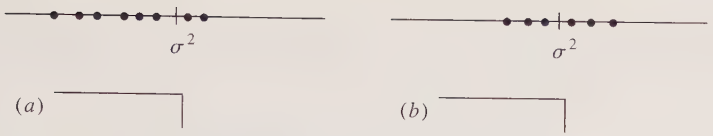

FIGURE 6.8
(*a*) The statistic $\Sigma_{i=1}^{n}(X_i - \bar{X})^2/n$ tends to underestimate σ^2. On the average, it will produce estimates that are a bit too small. It is not an unbiased estimator for σ^2; (*b*) the statistic $\Sigma_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ is unbiased for σ^2. On the average, it will produce estimates that are centered at σ^2.

$n - 1$ rather than by n. In this way we obtain a statistic that is unbiased for σ^2. The term "unbiased" is a technical term. It is defined formally in Sec. 7.1. Basically, it means "centered at the right spot." Since the sample variance is used to estimate σ^2, in this case "the right spot" is σ^2. Successive estimates for σ^2 based on the formula $[\Sigma_{i=1}^{n}(X_i - \bar{X})^2]/(n-1)$ should be centered at σ^2. Figure 6.8 illustrates the expected behavior of the two statistics just discussed. So that the statistic used to estimate σ^2 will be unbiased for σ^2, we choose to define the variance of a sample as given in Definition 6.3.3. The definition of the term "sample standard deviation" follows logically.

Definition 6.3.3 (Sample variance and sample standard deviation). Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample of size n from the distribution of X. Then the statistic

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

is called the *sample variance*. Furthermore, the statistic $S = \sqrt{S^2}$ is called the *sample standard deviation*.

Recall that when we computed the value of σ^2 in Chap. 3, the actual definition of the term "variance" was seldom used; a computational formula was developed that was arithmetically easier to handle than the definition. The same is true here. When S^2 is evaluated from a sample, Definition 6.3.3 is not commonly used. Rather, we use a computational formula.

Theorem 6.3.1 (A computational formula for S^2). Let $X_1, X_2, X_3, \ldots, X_n$ be a random sample of size n from the distribution of X. The sample variance is given by

$$S^2 = \frac{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}{n(n-1)}$$

The above formula was convenient before the advent of calculators with built-in statistical capabilities and statistical computer packages. Since most calculators

will find s for you by simply entering the data in a statistical mode, this formula is not often needed. We present it here because you might encounter it in some other setting and wonder about its validity. You are encouraged to use whatever computing aids you have available to find $\bar{x}$, s^2, and s. However, the use of the formula is illustrated in Example 6.3.3. In reporting s^2, we shall usually retain two more decimal places than that of the data; s will be reported to one more decimal place. Rounding will be used.

Example 6.3.3. These data constitute a sample of observations on X, the coal consumption in millions of tons by electric utilities for a given year:

390 400 406 410 450 395 401 408 415

To compute the sample variance, we must evaluate the statistics $\sum_{i=1}^{n} X_i$ and $\sum_{i=1}^{n} X_i^2$ for this sample. The observed values are

$$\sum_{i=1}^{9} x_i = 3675 \qquad \sum_{i=1}^{9} x_i^2 = 1{,}503{,}051$$

The observed value of S^2 is

$$S^2 = \frac{9\sum_{i=1}^{9} x_i^2 - \left(\sum_{i=1}^{9} x_i\right)^2}{9(8)} = \frac{9(1{,}503{,}051) - (3675)^2}{9(8)} \doteq 303.25$$

Remember that variance is usually considered to be unitless because the physical unit attached to it is often meaningless. The observed value of S is

$$s = \sqrt{s^2} = \sqrt{303.25} \doteq 17.4 \text{ million tons}$$

Notice that the physical measurement unit associated with s matches that of the original data and that 17.4 million tons is the standard deviation for this sample. It is not the standard deviation in coal consumption for all electric utilities across the country for the given year. However, it does indicate that σ probably has a value close to 17.4 million tons.

The last sample statistic to be considered is the *sample range*. This statistic was used in categorizing data in Sec. 6.2.

Definition 6.3.4 (Sample range). The sample range is defined to be the difference between the largest and smallest observations with subtraction in the order largest minus smallest.

The sample range for the data of Example 6.3.3 is $450 - 390 = 60$ million tons.

One word of caution is in order. We have assumed that the data set presented in this section represents a random sample drawn from a larger population because this is the situation most often encountered in practice. Occasionally you will encounter a data set that is *not* a sample. Rather, it represents an observation on X for *every* member of the population. If this is the case, then the population mean is just

the arithmetic average of these observations; that is, $\mu = \bar{x}$. Furthermore, the population variance is given by

Population Variance

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

Be careful! Be sure that you understand the nature of your data set before you begin to summarize its properties.

6.4 BOXPLOTS

In summarizing data, it is useful to report all the statistics considered in Sec. 6.3. This is especially true if the data set contains a value that is unusually large or unusually small. A value that appears to be atypical in that it seems to be far removed from the bulk of the data is called an *outlier* or a "wild" number. It is important to be able to detect such numbers and to understand the effect that they have on the usual sample statistics.

Outliers arise for two reasons: (1) They are legitimate observations whose values are simply unusually large or unusually small, or (2) they are the result of an error in measurement, poor experimental technique, or a mistake in recording or entering the data. In the first case it is suggested that the presence of the outlier be reported and that sample statistics be reported both with and without the outlier. In the second case the data point can be corrected if possible or else dropped from the data set.

Of the statistics presented thus far the sample mean, the variance, the standard deviation, and the range are adversely affected by the presence of an outlier; however, the sample median is not so affected. Thus in the presence of an outlier the sample median may be preferable to the sample mean measure of location. We say that the median is *resistant* to outliers.

Sometimes outliers are so obvious that their presence can be detected by inspection. However, it is useful to have an analytical and graphical technique for identifying values that are truly unusual. One such technique is the *boxplot*. Its construction is based on the interquartile range, a measure of variability that is resistant to outliers. The sample interquartile range, iqr, represents the length of the interval that contains roughly the middle 50% of the data. If the iqr is small, then much of the data lies close to the center of the distribution; if it is large, the data tend to be widely dispersed. These steps are used to calculate the iqr.

Finding the Sample Interquartile Range

1. Find the median location $(n + 1)/2$, where n is the sample size.
2. Truncate the median location by rounding it *down* to the nearest whole number.

3. Find the quartile location q by

$$q = \frac{\text{truncated median location} + 1}{2}$$

4. Find q_1 by counting up from the smallest data point to location q. If q is an integer, then q_1 is the data point in position q. If q is not an integer, then q_1 is the average of the data points in positions $q - .5$ and $q + .5$. Approximately 25% of the data will fall on or below q_1.
5. Find q_3 by counting down from the largest data point to position q as in part 4. Approximately 75% of the data will fall on or below q_3.
6. Define iqr by iqr $= q_3 - q_1$.

Example 6.4.1. A study of the type of sediment found at two different deep-sea drilling sites is conducted. The random variable of interest is the percentage by volume of cement found in core samples. By cement we mean dissolved and reprecipitated carbonate material. The following data are obtained:

Site I, % cement				Site II, % cement		
10	21	12	12	1	10	14
20	13	24	36	9	21	19
31	18	17	16	15	17	13
37	16	32	13	25	22	20
14	49	25	19	24	12	23
13	32	27		15	20	18

The double stem-and-leaf diagram for the data of site I is shown in Fig. 6.9. The sample is size $n = 23$. The median location is $(n + 1)/2 = 12$. The quartile location is $q = (12 + 1)/2 = 6.5$. To find q_1, we use the stem-and-leaf diagram to locate the sixth and seventh data points, counting from the smaller numbers up. These values are 13 and 14, respectively. Hence $q_1 = (13 + 14)/2 = 13.5$. To find q_3, we find the sixth and seventh data points counting from the higher numbers down. These points are 31 and 27, respectively, yielding $q_3 = (31 + 27)/2 = 29$. The sample interquartile range is $q_3 - q_1 = 29 - 13.5 = 15.5$. For site II you can verify that $q_1 = 13$ and $q_3 = 21$.

A word of caution is in order. All computer software and statistical calculators calculate the median as we have done. However, different algorithms are sometimes used to find the quartiles; some will agree with our values, but others will not. All produce good estimates of the population quartiles. For example, if the TI83 calculator is used to find q_1 and q_3 for the data of Example 6.4.1, site I, it reports $q_1 = 13$ and $q_3 = 31$. These values differ slightly from those that we found previously. That calculator's answers will agree with ours for the data of site II. MINITAB reports $q_1 = 13$ and $q_3 = 31$ for site I and thus agrees with the TI83 calculator. However, it yields $q_1 = 12.75$ and $q_3 = 21.25$ for the quartiles of site II. These do not agree with our estimates or those of the TI83. Just be aware that different technologies can yield slightly different quartiles and therefore will produce slightly different boxplots when applied to the same set of data.

1	0433223
1	86769
2	014
2	57
3	122
3	76
4	
4	9

FIGURE 6.9
Double stem-and-leaf diagram for the percentage by volume of cement in core samples taken at deep-sea drilling site I.

Once the interquartile range has been found, it can be used to construct a boxplot. The *boxplot* is a graphical representation of a data set that gives a visual impression of location, spread, and the degree and direction of skewness. For an approximately bell-shaped distribution the *boxplot* also allows us to identify outliers. It is especially useful when we want to compare two or more data sets.

Constructing a Boxplot

1. A horizontal or vertical reference scale is constructed.
2. Find the sample median, q_1, q_3, and iqr.
3. Find two points f_1 and f_3, called *inner fences,* by

$$f_1 = q_1 - 1.5(\text{iqr})$$
$$f_3 = q_3 + 1.5(\text{iqr})$$

   These points will be used to identify outliers.

   They are *not* a visible part of the boxplot.
4. Find two points a_1 and a_3, called *adjacent values.* The point a_1 is the data point that is closest to f_1 without lying below f_1 in value. The point a_3 is the data point that is closest to f_3 without lying above f_3 in value.
5. Find two points F_1 and F_3, called *outer fences,* by

$$F_1 = q_1 - 2(1.5)(\text{iqr})$$
$$F_3 = q_3 + 2(1.5)(\text{iqr})$$

   These fences, as with inner fences, are not visible on the boxplot.
6. Locate the points found thus far on the horizontal or vertical scale. Their relative positions are shown in Fig. 6.10(*a*).
7. Construct a box with ends at q_1 and q_3 with an interior line drawn at the median, as shown in Fig. 6.10(*b*).
8. Indicate adjacent values by *x,* and connect them to the box with dashed lines. Locate any data points falling between the inner and outer fences, and denote these by open circles. These points are considered to be mild outliers. Indicate data points that fall beyond the outer fences with asterisks. These points are considered to be extreme outliers [see Fig. 6.10(*c*)].

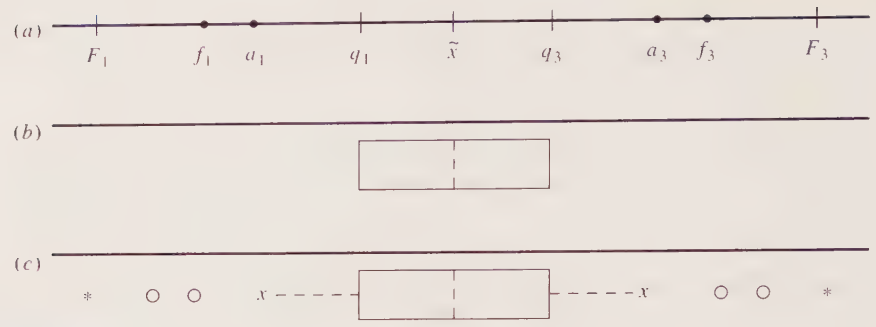

FIGURE 6.10
(*a*) Relative positions of median ($\tilde{x}$), quartiles (q_1 and q_3), adjacent values (a_1 and a_3), inner fences (f_1 and f_3), and outer fences (F_1 and F_3); (*b*) a box is drawn with ends at q_1 and q_3 and interior line at $\tilde{x}$; (*c*) adjacent values are indicated by *x*. Mild outliers are indicated by open circles; extreme outliers are given by asterisks.

The location of the midline of the box is an indication of the shape of the distribution. If the line is badly off center, then we know that the distribution is skewed in the direction of the longer end of the box.

Before we illustrate this technique, the notion of fences needs to be clarified. It can be shown that when sampling from a normal distribution, only about 7 values in every 1000 fall beyond the inner fences. You are asked to verify this result in Exercises 26 and 27. Since these values are very unusual, they are deemed to be outliers. Outliers must be treated with care since, as you have already seen, their presence can have a dramatic impact on $\bar{x}$, s^2, and s, the usual measures of location and variation. When an outlier is found, we should consider its source. Is it a legitimate data point whose value is simply unusually large or small? Is it a misrecorded value? Is it the result of some error or accident in experimentation? In the last two instances the point can be deleted from the data set and the analysis completed on the remaining data. In the first case we suggest that the presence of the outlier be made known and that statistics be reported both with and without the outlier. In this way the decision of whether or not to include the outlier in future analyses can be made by the researcher who is the subject matter expert.

Example 6.4.2. A study of posttraumatic amnesia after a closed head injury is conducted. One variable studied is the length of hospitalization in days. The stem-and-leaf diagram for the data is shown in Fig. 6.11. (Based on information found in Jerry Mysia et al., "Prospective Assessment of Posttraumatic Amnesia: A Comparison of GOAT and the OGMS," *Journal of Head Trauma Rehabilitation*, March 1990, pp. 65–77.) For these data the median location is $(n + 1)/2 = 11$ and the median is 40 days. Quartile location is $q =$ (truncated median location $+ 1)/2 = 6$. The points q_1 and q_3 are 32 and 47, respectively. The interquartile range is iqr $= q_3 - q_1 = 15$. The inner fences are

$$f_1 = q_1 - 1.5(\text{iqr}) \qquad f_3 = q_3 + 1.5(\text{iqr})$$
$$= 32 - 22.5 \qquad = 47 + 22.5$$
$$= 9.5 \qquad = 69.5$$

The adjacent values are $a_1 = 12$ and $a_3 = 61$. The outer fences are

$$F_1 = q_1 - 2(1.5)(\text{iqr}) \qquad F_3 = q_3 + 2(1.5)(\text{iqr})$$
$$= 32 - 45 \qquad = 47 + 45$$
$$= -13 \qquad = 92$$

The data set contains two points, 8 and 89, that qualify as mild outliers. The point 108 qualifies as an extreme outlier. Notice that since F_1 is negative, it is physically impossible to see an extreme outlier on the lower end of the scale. The boxplot is shown in Fig. 6.12. Notice that the midline of the box is near its center, indicating a nearly symmetric distribution. Are the outliers real observations that must be taken into account, or are they the result of errors in data collection? In this case it would be easy to check patient records to find the answer, and this should be done before proceeding with any further analysis of the data.

As with any other statistical technique, the method given here for detecting outliers must be used with care. Since the location of the fences is chosen to detect unusual values when sampling from a normal distribution, this fact must be kept in mind when interpreting the boxplot. If the data set is large enough so that a histogram or a stem-and-leaf plot exhibits the bell characteristic of a normal curve, then legitimate data points that are flagged as outliers are unusual enough to warrant investigation. If the data set is small or appears to be drawn from a distribution that is not normal, then no real conclusions concerning outliers can be drawn. For example, the exponential distribution is far from symmetric and by nature has a long tail. In this case it is quite likely that the technique demonstrated in this section would flag the largest data point as an outlier. In fact, the point might not be unusual

0	8
1	2
2	07
3	0256
4	00001257
5	02
6	1
7	
8	9
9	
10	8

FIGURE 6.11
Stem-and-leaf diagram for the data of Example 6.4.2. Data represent length of hospitalization in days of posttraumatic amnesia patients ($n = 21$).

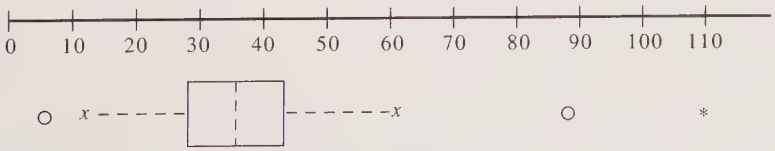

FIGURE 6.12
Boxplot for the data of Example 6.4.2.

at all. David Hoaglin and John Tukey [50] have a nice discussion of the use of boxplots and outliers for distributions that are not normal.

CHAPTER SUMMARY

This chapter is a link between the study of probability in its own right and the use of probability in the study of applied statistics. We began by defining exactly what we mean by the term "random sample." In particular, we noted that the term is used in three ways. It can denote the objects sampled, the random variables associated with those objects, or the numerical values assumed by these random variables. We noted also that in this text we are assuming that either sampling is from an infinite population, sampling is done with replacement from a finite population, or sampling without replacement from a finite population is done in such a way that the sample constitutes at most 5% of the population. This ensures that it is reasonable to assume that the random variables $X_1, X_2, \ldots, X_n$ are, for all practical purposes, independent. We introduced three graphical methods for picturing the distribution of a data set. These methods, the stem-and-leaf chart, histograms, and boxplots, help to determine the type of random variable with which we are dealing. That is, they help us get an idea of the shape of the density f associated with the random variable. The relative cumulative frequency ogive was introduced as a means of approximating the cumulative distribution function, F, of a continuous random variable. We introduced some summary statistics that serve two purposes. They describe the data set at hand, and they help approximate the value of corresponding parameters associated with the population from which the sample was drawn. We introduced and defined important terms that you should know. These are:

Population
Percentile
Median
Statistic
Decile
Sample variance
Frequency histogram
Relative frequency histogram
Relative cumulative frequency ogive
Inner fences
Outer fences
Adjacent values
Resistant statistic
Sample mean
Random sample
Quartile
Sample median
Stem and leaf
Interquartile range
Sample standard deviation
Sample range
Outlier
Mild outliers
Extreme outliers
Boxplots

EXERCISES

Section 6.1

In Exercises 1 through 5 a problem is described. In each case, decide whether a statistical study is appropriate. If so, explain why you think this is the case and identify the population(s) of interest.

1. A bridge is to be built across a deep canyon. An engineer is interested in determining the distribution of the random variable X, the maximum wind speed per day at the site, so that the bridge can be designed to withstand potential stresses that will be placed upon it from this source.
2. A botanist thinks that indoleacetic acid is effective in stimulating the formation of roots in cuttings from lemon trees. In an experiment to verify this contention two groups of cuttings are to be used. One group is to be treated with a dilute solution of indoleacetic acid; the other is given only water. Later a comparison of the root systems of the two groups will be made.
3. An architectural firm is to sublet a contract for a wiring project. Seven electrical contractors are available for the job. We want to determine the average estimated cost of the job and the average projected time required to complete the job for these seven contractors.
4. A computer system has a number of remote terminals attached to it. To decide whether or not to increase this number, it is necessary to study the random variable X, the length of time expended per session by users of the terminals currently in place.
5. Prior to changing from the traditional 8-hour-a-day, 5-day-a-week work schedule to a 10-hour-a-day, 4-day-a-week schedule, the opinion of the 50,000 workers who would be affected is to be sought.
6. Air quality is of concern to everyone. It is judged by the number of micrograms of particulate present per cubic meter of air. Assume that this variable is normally distributed with unknown mean and unknown variance. Monitoring stations sample air by sucking it through a thin fiberglass sheet that collects the fine particles suspended in the air. In a particular locality this is done for five randomly selected 24-hour periods each month. Thus each month a random sample of size $n = 5$ from a normal distribution is available.
   (*a*) Consider the random variable X_1, the particulate level for the first 24-hour period studied during a given month. What is the distribution of this random variable?
   (*b*) For a given month, these readings result:

   $$x_1 = 45 \qquad x_2 = 50 \qquad x_3 = 62 \qquad x_4 = 57 \qquad x_5 = 70$$

   For these data, evaluate the statistics ΣX_i, ΣX_i^2, $\Sigma X_i/n$, $\max_i\{X_i\}$, $\min_i\{X_i\}$.
   (*c*) Is the random variable $X_5 - \mu$ a statistic? Is the random variable $(X_5 - \mu)/\sigma$ a statistic? Explain.

Section 6.2

7. A data set containing 70 observations, each reported to one decimal place, is to be split into seven categories. The largest observation is 75.1, and the smallest is 16.3.
   (*a*) These data are covered by an interval of what length?
   (*b*) Using the method outlined in this section, each category will be of what length?
   (*c*) What is the lower boundary for the first category?
   (*d*) What are the boundaries for each of the seven categories?

8. Acute exposure to cadmium produces respiratory distress and kidney and liver damage, and may even result in death. For this reason, the level of airborne cadmium dust and cadmium oxide fume in the air is monitored. This level is measured in milligrams cadmium per cubic meter of air. A sample of 35 readings yields the following data:

.044	.030	.052	.044	.046
.020	.066	.052	.049	.030
.040	.045	.039	.039	.039
.057	.050	.056	.061	.042
.055	.037	.062	.062	.070
.061	.061	.058	.053	.060
.047	.051	.054	.042	.051

(*a*) Construct a stem-and-leaf diagram for these data. Use the numbers 02, 03, 04, 05, 06, and 07 as stems.

(*b*) Would you be surprised to hear someone claim that the random variable X, the cadmium level in the air, is normally distributed? Explain.

(*c*) Use the method outlined in this section to break these data into six categories. (Here a unit is .001 and a half unit is .0005.)

(*d*) Construct a frequency table and a relative frequency histogram for these data. Does the histogram exhibit the bell-shape characteristic of a normal density?

(*e*) Construct a cumulative frequency table and a relative cumulative frequency ogive for these data. Use the ogive to approximate that point above which 50% of the readings should fall.

9. Let X denote the time in minutes that a vehicle must wait to get through a traffic light at a busy intersection. The following data are obtained from a random sample of 36 vehicles:

.2	.5	.7	1.1	1.2	1.2	1.3	1.4	1.4	1.4
1.5	1.5	1.6	1.6	1.7	1.9	2.0	2.1	2.1	2.2
2.3	2.5	2.6	2.9	2.8	3.0	3.1	3.0	3.7	3.7
4.0	4.1	4.5	5.1	5.8	1.4				

(*a*) Construct a double stem-and-leaf diagram for these data.

(*b*) Do the data suggest that the distribution of X is skewed? If so, what is the direction of the skew?

10. Liquid products were first obtained from coal in England during the 1700s. Lamp oil was produced from coal in the United States as early as 1850, but the domestic coal chemicals industry did not develop until World War I. A modern coal-for-recovery system uses a battery of coke ovens to produce liquid products from the coal feed. These observations are obtained on the random variable X, the number of gallons of liquid product obtained per ton of coal feed:

7.6	8.2	7.1	10.0	6.5	9.6
6.1	6.2	7.6	6.2	9.5	6.7
7.4	9.5	9.2	8.0	8.5	9.3
8.8	9.6	9.7	6.8	7.1	7.7
8.7	7.8	8.7	8.2	8.2	7.4
9.0	8.8	7.3	7.9	7.1	7.9
7.6	6.7	8.1	6.2	5.3	7.4
7.7	9.1	7.9	8.7	8.4	8.1

(*a*) Construct a stem-and-leaf diagram for these data. Use the numbers 5, 6, 7, 8, 9, 10 as stems.
(*b*) Is the assumption that X is normally distributed justifiable? Explain.
(*c*) Use the method outlined in this section to break these data into six categories.
(*d*) Construct a frequency table and a relative frequency histogram for these data. Does the histogram exhibit the bell-shape characteristic of a normal density?
(*e*) Construct a cumulative frequency table and a relative cumulative frequency ogive for these data. Use the ogive to approximate the probability that a randomly selected ton of coal will yield less than 7 gallons of liquid product.

11. Some efforts are currently being made to make textile fibers out of peat fibers. This would provide a source of cheap feedstock for the textile and paper industries. One variable being studied is X, the percentage ash content of a particular variety of peat moss. Assume that a random sample of 50 mosses yields these observations:

.5	1.8	4.0	1.0	2.0
1.1	1.6	2.3	3.5	2.2
2.0	3.8	3.0	2.3	1.8
3.6	2.4	.8	3.4	1.4
1.9	2.3	1.2	1.9	2.3
2.6	3.1	2.5	1.7	5.0
1.3	3.0	2.7	1.2	1.5
3.2	2.4	2.5	1.9	3.1
2.4	2.8	2.7	4.5	2.1
1.5	.7	3.7	1.8	1.7

(*a*) Construct a stem-and-leaf diagram for these data. Use the numbers 0, 1, 2, 3, 4, 5 as stems.
(*b*) Is there any reason to suspect that X is not normally distributed? Explain.
(*c*) Use the method outlined in this section to break these data into six categories.
(*d*) Construct a frequency table and a relative frequency histogram for these data. Does the histogram suggest that X might not be normally distributed? If so, what distribution might be appropriate?
(*e*) Construct a cumulative frequency table and a relative cumulative frequency ogive for these data. Use the ogive to approximate the probability that a randomly selected specimen of this variety of moss will have an ash content that exceeds 2%.

12. *(Percentiles.)* Let X be a random variable. The point $p_{k/100}$ ($k = 1, 2, 3, \ldots, 100$) such that

$$P[X < p_{k/100}] \leq k/100 \qquad \text{and} \qquad P[X \leq p_{k/100}] \geq k/100$$

is called the *kth percentile* for X. For example, let X be binomial with $n = 20$ and $p = .5$. The 25th percentile for X is the point $p_{25/100} = 8$ since, from Table I of App. A, we see that

$$P[X < 8] = .1316 \leq .25 \qquad \text{and} \qquad P[X \leq 8] = .2517 \geq .25$$

(*a*) Let X be binomial with $n = 20$ and $p = .5$. Find the 60th percentile for X.
(*b*) Let X be Poisson with $\lambda s = 10$. Find the 30th percentile for X.
(*c*) Argue that in the case of a continuous random variable the kth percentile is that point such that $P[X \leq p_{k/100}] = k/100$.
(*d*) Let X be exponentially distributed with $\beta = 1$. Show that the 20th percentile for X is $-\ln .80$. *Hint:* Find the point p such that

$$\int_0^p e^{-x}\,dx = .20$$

13. *(Quartiles.)* The 25th, 50th, 75th, and 100th percentiles for X are called its first, second, third, and fourth *quartiles,* respectively.
(*a*) State the definition of the first quartile in terms of probabilities.
(*b*) Let X be binomial with $n = 20$ and $p = .5$. Find the first quartile for X.
(*c*) Let X be exponentially distributed with $\beta = 1$. Find the first quartile for X.

14. *(Deciles.)* The 10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, 90th, and 100th percentiles for X are called its *deciles.*
(*a*) State the definition of the 4th decile for X in terms of probabilities.
(*b*) Let X be Poisson with $\lambda s = 10$. Find the 6th decile for X.
(*c*) Let X be exponentially distributed with $\beta = 1$. Find the third decile for X.

15. The percentiles, quartiles, and deciles for a continuous random variable can be approximated from a relative cumulative frequency ogive using the projective method. For instance, in Fig. 6.5 we approximated the 50th percentile for X, the life span of a lithium battery, to be a little over 725 hours.
(*a*) Approximate the first quartile for X, the cadmium level in the air, using the data of Exercise 8.
(*b*) Approximate the fourth decile for X, the number of gallons of liquid product obtained per ton of coal fuel, using the data of Exercise 10.
(*c*) Approximate the 50th percentile for X, the percentage ash content for a particular variety of moss, using the data of Exercise 11.

16. In running computer programs on a time-sharing basis, the costs vary from session to session. These observations are obtained on the random variable X, the cost per session to the user:

$1.08	.84	1.41	.99	.82
.89	.38	1.05	1.19	.65
1.09	1.03	.81	.55	.71
1.89	.47	.59	1.22	1.27
1.02	1.09	1.02	.86	1.23
1.23	.85	1.02	1.25	.80

Construct a relative cumulative frequency ogive for these data. Use the ogive to approximate the 50th percentile; the first quartile; the third quartile.

Section 6.3

17. Consider these data sets:

I			II			
1	3	2	1	2	4	1
2	5	4	2	5	2	5
4	3	3	1	5	5	3

(*a*) Find the sample mean and sample median for each data set.
(*b*) Find the sample range for each data set.
(*c*) Find the sample variance and sample standard deviation for each data set.
(*d*) Would you be surprised to hear someone claim that these data were drawn from the same population? Explain. *Hint:* Consider the shape of the distribution as well as the observed values of the sample statistics.

18. The observed values of the statistics $\Sigma_{i=1}^{50} X_i$ and $\Sigma_{i=1}^{50} X_i^2$ for the data of Example 6.2.1 are $\Sigma_{i=1}^{50} x_i = 63{,}707$ and $\Sigma_{i=1}^{50} x_i^2 = 154{,}924{,}261$.
(*a*) Would you be surprised to hear someone claim that the mean lifespan of the lithium batteries used in this model calculator is 1270 hours? Explain.
(*b*) Find the sample variance and sample standard deviation for these data.

19. Use the data of Example 6.1.1 to approximate the mean and variance of the random variable X, the number of times per hour that a television signal is interrupted by random interference.

20. Use the data of Exercise 8 to approximate the mean, variance, and standard deviation of the random variable X, the level of airborne cadmium dust and cadmium oxide fumes. Assume that these approximations are fairly accurate. Between what two values would you expect approximately 95% of the readings to fall? Explain.

21. Use the data of Exercise 10 to approximate the mean, variance, and standard deviation of the random variable X, the number of gallons of liquid product obtained per ton of coal feed.

22. Use the data of Exercise 11 to approximate the mean, variance, and standard deviation of the random variable X, the percentage ash content of a particular variety of peat moss.

23. Consider the data of Exercise 9.
(*a*) Find the mean and median for these data.
(*b*) Find the standard deviation and variance for these data.
(*c*) What physical measurement unit is associated with each of the statistics in parts (*a*) and (*b*)?

24. There have been many improvements made in lighting in the last 10 years. One new bulb, the Philips' Earth Light, uses a compact screw-in fluorescent bulb with an electronic ballast incorporated in its base. It is thought to last 10 to 13 times longer than household bulbs used in the past. These data are obtained on the life span of a sample of these new bulbs (time is in thousands of hours):

9.1	10.1	9.0	11.4
10.5	9.5	12.0	9.1
12.2	13.1	10.0	9.3
9.0	9.6	11.1	9.1
13.3	10.7	9.1	9.0
9.0	11.0	9.2	11.6

(Based on information found in "Lighting Comes of Age with New Technology," *Research and Development,* November 1992, pp. 30–31.)
(*a*) Construct a stem-and-leaf diagram for these data, and suggest a distribution from which these data might have been drawn.

(*b*) Based on these data, approximate the value of μ, the average life span of these bulbs.

(*c*) Approximate the median life span of these bulbs, and explain exactly what this value means.

(*d*) Find the sample variance and sample standard deviation for these data.

(*e*) Criticize the following statement:
"Based on the normal probability rule, it is estimated that approximately 95% of all bulbs have a life span between 7,530 and 13,050 hours."

(*f*) Based on Chebyshev's inequality, what can be said about the proportion of bulbs whose life span is expected to fall between 7,530 and 13,050 hours?

25. *(Approximating σ via the range.)* The range can play an important role in the design of statistical studies. To obtain a prespecified degree of accuracy when estimating population parameters, an adequate sized sample must be drawn. Most formulas used to determine sample size require knowledge of σ, the population standard deviation. Often the researcher will not have an estimate of σ available but will have an idea of the expected range of his or her data. In Sec. 4.5 we saw that when sampling from a normal distribution,

$$P[-2\sigma < X - \mu < 2\sigma] \doteq .95$$

If X is not normally distributed, then Chebyshev's inequality can be applied to conclude that

$$P[-3\sigma < X - \mu < 3\sigma] \geq .89$$

That is, X always lies within at most 3 standard deviations of its mean with high probability. From this it can be concluded that the estimated range covers an interval of roughly 4σ for normally distributed random variables and 6σ otherwise. In the normal case an estimate of σ can be obtained by solving the equation

$$4\sigma \doteq \text{estimated range}$$

for σ. Thus we see that

$$\sigma \doteq \text{(estimated range)}/4$$

when X is normally distributed. If X is not normally distributed, then

$$\sigma \doteq \text{(estimated range)}/6$$

These data are obtained on the random variable X, the cpu time in seconds required to run a program using a statistical package:

6.2	5.8	4.6	4.9	7.1	5.2
8.1	.2	3.4	4.5	8.0	7.9
6.1	5.6	5.5	3.1	6.8	4.6
3.8	2.6	4.5	4.6	7.7	3.8
4.1	6.1	4.1	4.4	5.2	1.5

(*a*) Construct a stem-and-leaf diagram for these data. Is the assumption justified that X is normally distributed?

(*b*) Approximate σ via the sample standard deviation s.

(*c*) Find the sample range for these data, and use it to approximate σ. Compare your result to that obtained in part (*b*).

Section 6.4

26. Consider the standard normal distribution.
(*a*) Use the Z table to verify that q_1 is approximately $-.67$ and q_3 is approximately .67.
(*b*) Find the interquartile range for Z, and explain what this means.
(*c*) Verify that the inner fences for Z are $f_1 = -2.68$ and $f_3 = 2.68$.
(*d*) Verify that the probability that a standard normal random variable will fall beyond the inner fences is approximately .007.
(*e*) Find the outer fences for Z.
(*f*) Find the probability that a standard normal random variable will fall beyond the outer fences.

27. Let X be normally distributed with mean μ and variance σ^2.
(*a*) Verify that $q_3 = \mu + .67\sigma$ and that $q_1 = \mu - .67\sigma$.
(*b*) Find the interquartile range for X.
(*c*) Verify that the inner fences for X are $f_1 = \mu - 2.68\sigma$ and $f_3 = \mu + 2.68\sigma$.
(*d*) Verify that the probability that X will fall beyond the inner fences is approximately .007.

28. Temperature differences between the warm upper surface of the ocean and the colder deeper levels can be utilized to convert thermal energy to mechanical energy. This mechanical energy can in turn be used to produce electrical power using a vapor turbine. Let X denote the difference in temperature between the surface of the water and the water at a depth of 1 kilometer. Measurements are taken at 15 randomly selected sites in the Gulf of Mexico. These data result in the following temperatures:

22.5	23.8	23.2	22.8	10.1*
23.5	24.0	23.2	24.2	24.3
23.3	23.4	23.0	23.5	22.8

(*a*) Construct a double stem-and-leaf diagram for these data.
(*b*) Find the sample mean, sample median, and sample standard deviation for these data.
(*c*) Note that the starred observation in the data set is very different from the others. It is a potential outlier. Construct a boxplot for these data to verify that the value 10.1 does, in fact, qualify as an outlier.
(*d*) To see the effect of this outlier, drop it from the data set and calculate the sample mean, median, and standard deviation for the remaining 14 observations. Which measure is least affected by the presence of the outlier? Do you see why it is desirable to report both the mean and median of a data set?

29. Most homes utilize a variety of electronic equipment and appliances. For this reason, both suppliers and consumers of these products have become interested in product reliability. One aspect of reliability is the ability of the appliance to withstand power surges. In a study of this phenomena the following data are obtained on the strength of a surge in kilovolts required to damage or upset the appliance (based on figures found in "The Effects of Surges on Electronic Appliances," Stephen B. Smith and Ronald B. Standler, *IEEE Power Engineering Review,* July 1992, p. 50):

Clocks

1.1	3.5	3.2	4.0
3.6	1.5	2.3	3.0
4.7	4.0	4.9	2.9
2.6	2.5	2.7	4.2
2.4	3.7	3.8	6.0
5.0	1.8	5.6	5.1
3.9	3.8	3.7	3.5

Television receivers

2.0	5.0	4.5	5.4
4.2	4.3	5.1	5.6
5.2	4.7	4.4	5.8
5.2	7.8	5.4	4.9
4.6	4.6	5.0	4.8
5.3	5.2	5.3	5.9

dc power supplies

4.2	4.4	5.1	4.1	6.1
4.5	5.9	3.9	5.0	58
5.0	4.8	4.3	5.4	
4.7	5.1	5.2	5.6	
4.9	4.7	4.6	4.8	

(*a*) Sketch a double stem-and-leaf diagram for the clock data. Based on this diagram, would you be surprised to hear a claim that these data are drawn from an exponential distribution? Explain.

(*b*) Use the boxplot technique to check for outliers in the clock data. Based on your results, which measure of location, the sample mean or the sample median, is probably the better measure of the location of the bulk of the data for these data?

(*c*) Sketch a stem-and-leaf diagram for the television data. Use the stem 4 five times and the stem 5 five times. Based on this diagram, does there appear to be at least one outlier in the data set?

(*d*) Use the boxplot technique on the television data to test the suspicious points. Do you think that they are truly outliers? If so, are they mild outliers or extreme outliers? Which measure of variability, the sample variance or the iqr, is probably a better measure of the variability of the bulk of the data?

(*e*) Sketch a double stem-and-leaf diagram for the dc power supply data. These data contain an outlier due to a misplaced decimal point. Do you see it? Calculate the mean for the data using the bad data point as written. Now correct the data point and recalculate the sample mean. In light of this, explain what it means to say that $\bar{x}$ is not resistant to outliers.

REVIEW EXERCISES

30. Bricks are produced in lots of size 1000. Before shipping a lot, a sample of 25 bricks is selected and inspected for quality. Two random variables are of interest.

These are X, the number of chips per brick, and Y, the hardness of the brick. Assume that hardness is measured on a continuous scale from 1 to 10 with larger numbers indicating a harder brick:

x					y				
2	5	0	1	2	3.2	6.3	6.4	6.7	7.3
0	3	0	0	2	7.1	5.4	4.6	5.8	9.1
1	1	0	1	3	7.7	6.1	8.1	5.9	6.2
2	1	1	7	4	6.0	6.8	7.2	6.3	8.2
0	2	3	5	1	5.1	4.2	6.9	4.5	5.0

(*a*) What is the name of the family of random variables to which X belongs?

(*b*) Approximate the mean, variance, standard deviation, and median of X based on these data.

(*c*) Construct a stem-and-leaf diagram for the hardness measurements. Based on this diagram, would it be unrealistic to assume that Y is approximately normally distributed?

(*d*) Approximate the mean, variance, standard deviation, and median of Y.

31. In an attempt to study the problem of failure in field-installed computer equipment, data is collected on fifty field trips made to repair equipment. The random variables studied are X, the time in hours required to locate and rectify the problem, and Y, the cause of the failure. We define Y by

$$Y = \begin{cases} 1 & \text{if the failure is due to a faulty microprocessor chip} \\ 0 & \text{otherwise} \end{cases}$$

These data are obtained:

x							y						
1.52	1.83	2.25	4.73	2.89	1.49	1.34	0	0	0	0	0	0	1
2.15	2.66	2.79	1.35	1.54	4.59	4.27	0	0	0	0	0	0	0
3.91	2.76	3.03	3.52	5.97	1.45		0	0	0	0	0	0	
3.07	2.18	1.38	2.04	1.49	1.11		1	0	0	0	0	0	
1.24	4.84	2.82	3.16	4.58	3.28		0	0	0	0	0	0	
1.30	3.01	1.20	3.42	1.86	3.49		0	0	0	0	0	0	
3.93	2.56	2.63	5.60	4.60	5.34		0	0	0	0	0	0	
1.62	2.82	4.88	2.04	1.62	.24		0	0	1	0	0	0	

(*a*) Construct a relative frequency histogram for the data on the time required to locate and rectify the problem. Use six categories. Based on this histogram, would you be surprised to hear someone claim that X is approximately normally distributed? Explain.

(*b*) Approximate the mean, variance, and standard deviation for X.

(*c*) Construct a relative cumulative frequency ogive. Use this ogive to approximate the median for X. Approximately what percentage of problems can be located and rectified in 1.5 hours or less?

(*d*) Let p denote the probability that the failure is due to a faulty microprocessor chip. Assume that even though p is unknown its value is the same for each chip. Theoretically, Y follows a point binomial distribution with parameter p. What is the theoretical mean for Y? Approximate this mean based on these data. If asked to approximate the probability that a future failure is due to the failure of a microprocessor chip, what would you say?

(*e*) What is the theoretical variance for Y? Use your answer to part (*d*) to approximate the variance of Y. Use the sample variance to approximate σ_Y^2. Did you get the same result? Which answer is unbiased for σ_Y^2?

(*f*) Use the technique of Exercise 25 to estimate σ_X. Compare your answer to that of part (*b*).

(*g*) Construct a boxplot for the data on x.

32. Most people are familiar with sparklers burned to celebrate New Year's Day and the Fourth of July. Two random variables are of interest. These are X, the length of the chemical coating that covers the tip of the sparkler, and Y, the burn time of the sparkler in seconds. These data are obtained on these random variables (based on data gathered in 1993–1994 by students at Radford University and Virginia Polytechnic Institute and State University):

x (in)	y (s)	x (in)	y (s)
4.5	29	4.3	22
3.6	26	3.5	21
4.0	25	4.5	30
3.7	25	4.6	22
4.0	27	4.6	25
3.7	27	3.9	20
4.0	28	3.5	13
4.0	25	3.8	19
3.8	25	3.9	23
4.0	28	3.6	25
3.8	24	3.6	27
4.1	15	3.6	18
3.9	22	3.3	11
4.1	25	3.7	24
3.9	24	3.7	23
4.2	26	4.3	26
3.8	24	3.9	27

(*a*) Construct a stem-and-leaf diagram for the burn time data. Use each stem 5 times so that each stem will involve two leaves.

(*b*) Construct a boxplot for the burn time data. Are any data points flagged as outliers?

(*c*) Take a good look at the shape of the distribution as indicated by the stem-and-leaf diagram. Does the distribution appear to be skewed? If so, what is the direction of the skew? If we assume in this case that all data points are legitimate and not due to poor technique or recording errors, then from what family of random variables might these data have been drawn? Do you think that the "outliers" should be treated as such? Explain.

(*d*) Construct a stem-and-leaf diagram for the length data. Again, use each stem 5 times. Does the normality assumption appear fairly reasonable here?

(*e*) Construct a boxplot for the length data, and comment on any outliers that might be identified.

33. Let X denote the gasoline mileage obtained in tests on a newly designed SUV (sport utility vehicle). A sample of 21 simulated test runs yields these data:

15	16	17	18	17	20
16	17	18	20	18	
18	19	19	17	21	
17	19	18	17	22	

(*a*) Construct a stem-and-leaf diagram for these data. Do the data suggest that X is normally distributed?
(*b*) Calculate the mean and median for this sample.
(*c*) Calculate the standard deviation and variance for this sample.
(*d*) Find the values of q_1 and q_3 and the iqr for the sample. Compare these values to those obtained via a TI83 calculator or any other technology tool that you have at your disposal.

34. In designing airplanes and airplane seats it is important to consider such variables as height and weight of passengers. A random sample of 100 adult male passengers yielded these weights:

212.8	256.3	278.1	298.3
213.7	257.0	278.2	298.4
214.2	258.6	279.1	299.3
217.7	259.1	279.6	300.8
219.8	259.2	279.9	300.9
220.0	261.6	283.0	301.1
224.5	262.5	283.1	301.7
225.3	265.2	283.6	302.5
227.8	267.0	284.9	304.8
230.8	267.9	285.0	306.6
232.7	268.1	286.0	306.8
233.8	268.3	286.3	310.5
236.1	269.1	286.6	310.6
237.7	269.5	286.6	310.9
239.7	270.1	286.8	310.9
241.0	271.2	286.9	312.4
243.3	271.8	289.3	313.8
244.7	272.7	290.4	316.0
246.1	273.3	291.0	316.9
249.8	274.8	291.2	320.2
250.9	275.1	293.8	321.5
252.0	275.2	296.1	332.2
252.7	275.8	296.1	335.7
254.9	276.8	297.1	342.4
255.4	277.6	298.2	353.6

(*a*) Calculate $\bar{x}$ and s.
(*b*) Use whatever technology tools you have available to construct a histogram for these data.
(*c*) It is thought that adult male weight is normally distributed with $\mu = 273$ pounds and $\sigma = 30$ pounds. Do your findings tend to support this notion?
(*d*) Figure 6.13 shows the ogive, the graph of the relative cumulative frequency distribution, for these data. Use it to estimate q_1, q_3, and the median.

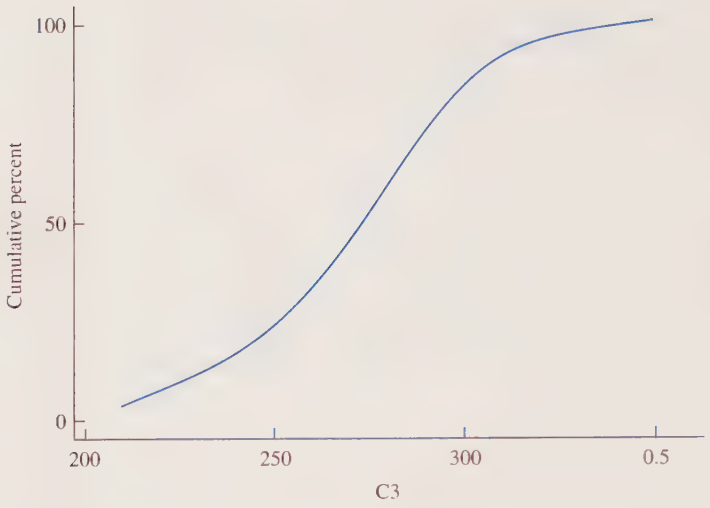

FIGURE 6.13 Ogive for the data of Exercise 6.34.

(*e*) Use the textbook method or any other technology tool to estimate q_1, q_3, and the median, and compare these values to your graphical estimates.

35. A study of the lights used at railroad-highway grade crossings is conducted. The purpose of the study is to compare two types of lamps. These are 25-watt lamps to which a low warming voltage is applied during the off portion of the flashing cycle and 25-watt standard lamps. The data obtained in the study are found on the website. Variables are observation number, type of lamp with 1 = warmed lamp and 2 = standard lamp, and life span in thousands of hours.

(*a*) Plot a histogram for each type of lamp, and discuss the shape of the distribution from which each sample was drawn.

(*b*) Find the mean, median, standard deviation, and variance for each sample. Compare the values of these statistics. Do the samples seem similar in any way?

(*c*) Construct a boxplot for each sample, and note any outliers that are identified.

(*d*) If outliers are found, delete them and recompute the statistics requested in part(*b*) to see the effect that these outliers have on each statistic.

36. It is known that power surges or line spikes can damage sensitive electronic equipment. A study of these surges is conducted. The purpose of the study is to ascertain whether or not there are differences in the frequency of these surges among the seven days of the week. Date for the study is found on the website. Variables are observation number; day, with *m* = Monday, *t* = Tuesday, *w* = Wednesday, *th* = Thursday, *f* = Friday, *s* = Saturday, and *sn* = Sunday; and number of spikes per day.

(*a*) Obtain descriptive statistics on the number of spikes per day for each day of the week. Discuss any differences among days that appear to exist.

(*b*) Construct boxplots for each day, and use the boxplots for a visual comparison of days.