Exercises

1.7 Consider the drying time data for Exercise 1.1 on page 13. Compute the sample variance and sample standard deviation.

1.8 Compute the sample variance and standard deviation for the water absorbency data of Exercise 1.2 on page 13.

1.9 Exercise 1.3 on page 13 showed tensile strength data for two samples, one in which specimens were exposed to an aging process and one in which there was no aging of the specimens.

(a) Calculate the sample variance as well as standard deviation in tensile strength for both samples.

(b) Does there appear to be any evidence that aging affects the variability in tensile strength? (See also the plot for Exercise 1.3 on page 13.)

1.10 For the data of Exercise 1.4 on page 13, compute both the mean and the variance in "flexibility" for both company A and company B. Does there appear to be a difference in flexibility between company A and company B?

1.11 Consider the data in Exercise 1.5 on page 13. Compute the sample variance and the sample standard deviation for both control and treatment groups.

1.12 For Exercise 1.6 on page 13, compute the sample standard deviation in tensile strength for the samples separately for the two temperatures. Does it appear as if an increase in temperature influences the variability in tensile strength? Explain.

1.5 Discrete and Continuous Data

Statistical inference through the analysis of observational studies or designed experiments is used in many scientific areas. The data gathered may be **discrete** or **continuous**, depending on the area of application. For example, a chemical engineer may be interested in conducting an experiment that will lead to conditions where yield is maximized. Here, of course, the yield may be in percent or grams/pound, measured on a continuum. On the other hand, a toxicologist conducting a combination drug experiment may encounter data that are binary in nature (i.e., the patient either responds or does not).

Great distinctions are made between discrete and continuous data in the probability theory that allow us to draw statistical inferences. Often applications of statistical inference are found when the data are *count data*. For example, an engineer may be interested in studying the number of radioactive particles passing through a counter in, say, 1 millisecond. Personnel responsible for the efficiency of a port facility may be interested in the properties of the number of oil tankers arriving each day at a certain port city. In Chapter 5, several distinct scenarios, leading to different ways of handling data, are discussed for situations with count data.

Special attention even at this early stage of the textbook should be paid to some details associated with binary data. Applications requiring statistical analysis of binary data are voluminous. Often the measure that is used in the analysis is the *sample proportion*. Obviously the binary situation involves two categories. If there are n units involved in the data and x is defined as the number that fall into category 1, then $n - x$ fall into category 2. Thus, x/n is the sample proportion in category 1, and $1 - x/n$ is the sample proportion in category 2. In the biomedical application, 50 patients may represent the sample units, and if 20 out of 50 experienced an improvement in a stomach ailment (common to all 50) after all were given the drug, then $\frac{20}{50} = 0.4$ is the sample proportion for which the drug was a success and $1 - 0.4 = 0.6$ is the sample proportion for which the drug was not successful. Actually the basic numerical measurement for binary data is generally denoted by either 0 or 1. For example, in our medical example, a successful result is denoted by a 1 and a nonsuccess a 0. As a result, the sample proportion is actually a sample mean of the ones and zeros. For the successful category,

$$\frac{x_1 + x_2 + \cdots + x_{50}}{50} = \frac{1 + 1 + 0 + \cdots + 0 + 1}{50} = \frac{20}{50} = 0.4.$$

What Kinds of Problems Are Solved in Binary Data Situations?

The kinds of problems facing scientists and engineers dealing in binary data are not a great deal unlike those seen where continuous measurements are of interest. However, different techniques are used since the statistical properties of sample proportions are quite different from those of the sample means that result from averages taken from continuous populations. Consider the example data in Exercise 1.6 on page 13. The statistical problem underlying this illustration focuses on whether an intervention, say, an increase in curing temperature, will alter the population mean tensile strength associated with the silicone rubber process. On the other hand, in a quality control area, suppose an automobile tire manufacturer reports that a shipment of 5000 tires selected randomly from the process results in 100 of them showing blemishes. Here the sample proportion is $\frac{100}{5000} = 0.02$. Following a change in the process designed to reduce blemishes, a second sample of 5000 is taken and 90 tires are blemished. The sample proportion has been reduced to $\frac{90}{5000} = 0.018$. The question arises, "Is the decrease in the sample proportion from 0.02 to 0.018 substantial enough to suggest a real improvement in the population proportion?" Both of these illustrations require the use of the statistical properties of sample averages—one from samples from a continuous population, and the other from samples from a discrete (binary) population. In both cases, the sample mean is an **estimate** of a population parameter, a population mean in the first illustration (i.e., mean tensile strength), and a population proportion in the second case (i.e., proportion of blemished tires in the population). So here we have sample estimates used to draw scientific conclusions regarding population parameters. As we indicated in Section 1.3, this is the general theme in many practical problems using statistical inference.

1.6 Statistical Modeling, Scientific Inspection, and Graphical Diagnostics

Often the end result of a statistical analysis is the estimation of parameters of a **postulated model**. This is natural for scientists and engineers since they often deal in modeling. A statistical model is not deterministic but, rather, must entail some probabilistic aspects. A model form is often the foundation of **assumptions** that are made by the analyst. For example, in Example 1.2 the scientist may wish to draw some level of distinction between the nitrogen and no-nitrogen populations through the sample information. The analysis may require a certain model for the data, for example, that the two samples come from **normal** or **Gaussian distributions**. See Chapter 6 for a discussion of the normal distribution.

Obviously, the user of statistical methods cannot generate sufficient information or experimental data to characterize the population totally. But sets of data are often used to learn about certain properties of the population. Scientists and engineers are accustomed to dealing with data sets. The importance of characterizing or *summarizing* the nature of collections of data should be obvious. Often a summary of a collection of data via a graphical display can provide insight regarding the system from which the data were taken. For instance, in Sections 1.1 and 1.3, we have shown dot plots.

In this section, the role of sampling and the display of data for enhancement of **statistical inference** is explored in detail. We merely introduce some simple but often effective displays that complement the study of statistical populations.

Scatter Plot

At times the model postulated may take on a somewhat complicated form. Consider, for example, a textile manufacturer who designs an experiment where cloth specimen that contain various percentages of cotton are produced. Consider the data in Table 1.3.

Table 1.3: Tensile Strength

Cotton Percentage	**Tensile Strength**
15	7, 7, 9, 8, 10
20	19, 20, 21, 20, 22
25	21, 21, 17, 19, 20
30	8, 7, 8, 9, 10

Five cloth specimens are manufactured for each of the four cotton percentages. In this case, both the model for the experiment and the type of analysis used should take into account the goal of the experiment and important input from the textile scientist. Some simple graphics can shed important light on the clear distinction between the samples. See Figure 1.5; the sample means and variability are depicted nicely in the scatter plot. One possible goal of this experiment is simply to determine which cotton percentages are truly distinct from the others. In other words, as in the case of the nitrogen/no-nitrogen data, for which cotton percentages are there clear distinctions between the populations or, more specifically, between the population means? In this case, perhaps a reasonable model is that each sample comes from a normal distribution. Here the goal is very much like that of the nitrogen/no-nitrogen data except that more samples are involved. The formalism of the analysis involves notions of hypothesis testing discussed in Chapter 10. Incidentally, this formality is perhaps not necessary in light of the diagnostic plot. But does this describe the real goal of the experiment and hence the proper approach to data analysis? It is likely that the scientist anticipates the existence of a *maximum population mean tensile strength* in the range of cotton concentration in the experiment. Here the analysis of the data should revolve around a different type of model, one that postulates a type of structure relating the population mean tensile strength to the cotton concentration. In other words, a model may be written

$$\mu_{t,c} = \beta_0 + \beta_1 C + \beta_2 C^2,$$

where $\mu_{t,c}$ is the population mean tensile strength, which varies with the amount of cotton in the product C. The implication of this model is that for a fixed cotton level, there is a population of tensile strength measurements and the population mean is $\mu_{t,c}$. This type of model, called a **regression model**, is discussed in Chapters 11 and 12. The functional form is chosen by the scientist. At times the data analysis may suggest that the model be changed. Then the data analyst "entertains" a model that may be altered after some analysis is done. The use of an empirical model is accompanied by **estimation theory**, where β_0, β_1, and β_2 are estimated by the data. Further, statistical inference can then be used to determine model adequacy.

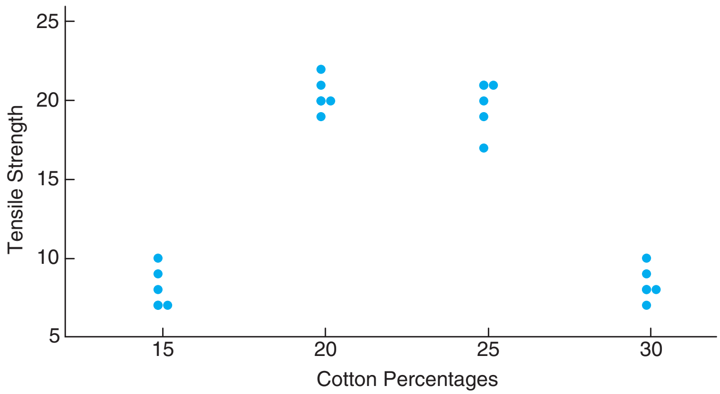

Figure 1.5: Scatter plot of tensile strength and cotton percentages.

Two points become evident from the two data illustrations here: (1) The type of model used to describe the data often depends on the goal of the experiment; and (2) the structure of the model should take advantage of nonstatistical scientific input. A selection of a model represents a **fundamental assumption** upon which the resulting statistical inference is based. It will become apparent throughout the book how important graphics can be. Often, plots can illustrate information that allows the results of the formal statistical inference to be better communicated to the scientist or engineer. At times, plots or **exploratory data analysis** can teach the analyst something not retrieved from the formal analysis. Almost any formal analysis requires assumptions that evolve from the model of the data. Graphics can nicely highlight **violation of assumptions** that would otherwise go unnoticed. Throughout the book, graphics are used extensively to supplement formal data analysis. The following sections reveal some graphical tools that are useful in exploratory or descriptive data analysis.

Stem-and-Leaf Plot

Statistical data, generated in large masses, can be very useful for studying the behavior of the distribution if presented in a combined tabular and graphic display called a **stem-and-leaf plot**.

To illustrate the construction of a stem-and-leaf plot, consider the data of Table 1.4, which specifies the "life" of 40 similar car batteries recorded to the nearest tenth of a year. The batteries are guaranteed to last 3 years. First, split each observation into two parts consisting of a stem and a leaf such that the stem represents the digit preceding the decimal and the leaf corresponds to the decimal part of the number. In other words, for the number 3.7, the digit 3 is designated the stem and the digit 7 is the leaf. The four stems 1, 2, 3, and 4 for our data are listed vertically on the left side in Table 1.5; the leaves are recorded on the right side opposite the appropriate stem value. Thus, the leaf 6 of the number 1.6 is recorded opposite the stem 1; the leaf 5 of the number 2.5 is recorded opposite the stem 2; and so forth. The number of leaves recorded opposite each stem is summarized under the frequency column.

Table 1.4: Car Battery Life

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

Table 1.5: Stem-and-Leaf Plot of Battery Life

Stem	**Leaf**	**Frequency**
1	69	2
2	25669	5
3	0011112223334445567778899	25
4	11234577	8

The stem-and-leaf plot of Table 1.5 contains only four stems and consequently does not provide an adequate picture of the distribution. To remedy this problem, we need to increase the number of stems in our plot. One simple way to accomplish this is to write each stem value twice and then record the leaves 0, 1, 2, 3, and 4 opposite the appropriate stem value where it appears for the first time, and the leaves 5, 6, 7, 8, and 9 opposite this same stem value where it appears for the second time. This modified double-stem-and-leaf plot is illustrated in Table 1.6, where the stems corresponding to leaves 0 through 4 have been coded by the symbol $\star$ and the stems corresponding to leaves 5 through 9 by the symbol $\cdot$.

In any given problem, we must decide on the appropriate stem values. This decision is made somewhat arbitrarily, although we are guided by the size of our sample. Usually, we choose between 5 and 20 stems. The smaller the number of data available, the smaller is our choice for the number of stems. For example, if the data consist of numbers from 1 to 21 representing the number of people in a cafeteria line on 40 randomly selected workdays and we choose a double-stem-and-leaf plot, the stems will be $0\star$, $0\cdot$, $1\star$, $1\cdot$, and $2\star$ so that the smallest observation 1 has stem $0\star$ and leaf 1, the number 18 has stem $1\cdot$ and leaf 8, and the largest observation 21 has stem $2\star$ and leaf 1. On the other hand, if the data consist of numbers from \$18,800 to \$19,600 representing the best possible deals on 100 new automobiles from a certain dealership and we choose a single-stem-and-leaf plot, the stems will be 188, 189, 190, ..., 196 and the leaves will now each contain two digits. A car that sold for \$19,385 would have a stem value of 193 and the two-digit leaf 85. Multiple-digit leaves belonging to the same stem are usually separated by commas in the stem-and-leaf plot. Decimal points in the data are generally ignored when all the digits to the right of the decimal represent the leaf. Such was the case in Tables 1.5 and 1.6. However, if the data consist of numbers ranging from 21.8 to 74.9, we might choose the digits 2, 3, 4, 5, 6, and 7 as our stems so that a number such as 48.3 would have a stem value of 4 and a leaf of 8.3.

Table 1.6: Double-Stem-and-Leaf Plot of Battery Life

Stem	**Leaf**	**Frequency**
$1\cdot$	69	2
$2\star$	2	1
$2\cdot$	5669	4
$3\star$	001111222333444	15
$3\cdot$	5567778899	10
$4\star$	11234	5
$4\cdot$	577	3

The stem-and-leaf plot represents an effective way to summarize data. Another way is through the use of the **frequency distribution**, where the data, grouped into different classes or intervals, can be constructed by counting the leaves belonging to each stem and noting that each stem defines a class interval. In Table 1.5, the stem 1 with 2 leaves defines the interval 1.0–1.9 containing 2 observations; the stem 2 with 5 leaves defines the interval 2.0–2.9 containing 5 observations; the stem 3 with 25 leaves defines the interval 3.0–3.9 with 25 observations; and the stem 4 with 8 leaves defines the interval 4.0–4.9 containing 8 observations. For the double-stem-and-leaf plot of Table 1.6, the stems define the seven class intervals 1.5–1.9, 2.0–2.4, 2.5–2.9, 3.0–3.4, 3.5–3.9, 4.0–4.4, and 4.5–4.9 with frequencies 2, 1, 4, 15, 10, 5, and 3, respectively.

Histogram

Dividing each class frequency by the total number of observations, we obtain the proportion of the set of observations in each of the classes. A table listing relative frequencies is called a **relative frequency distribution**. The relative frequency distribution for the data of Table 1.4, showing the midpoint of each class interval, is given in Table 1.7.

The information provided by a relative frequency distribution in tabular form is easier to grasp if presented graphically. Using the midpoint of each interval and the

Table 1.7: Relative Frequency Distribution of Battery Life

Class Interval	Class Midpoint	Frequency, f	Relative Frequency
1.5–1.9	1.7	2	0.050
2.0–2.4	2.2	1	0.025
2.5–2.9	2.7	4	0.100
3.0–3.4	3.2	15	0.375
3.5–3.9	3.7	10	0.250
4.0–4.4	4.2	5	0.125
4.5–4.9	4.7	3	0.075

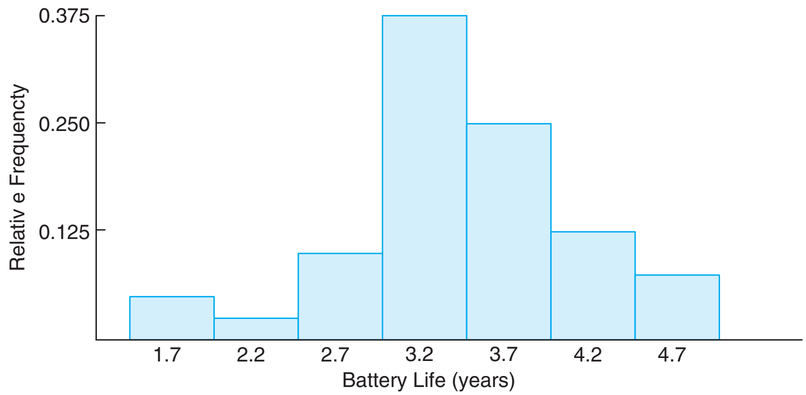

Figure 1.6: Relative frequency histogram.

corresponding relative frequency, we construct a **relative frequency histogram** (Figure 1.6).

Many continuous frequency distributions can be represented graphically by the characteristic bell-shaped curve of Figure 1.7. Graphical tools such as what we see in Figures 1.6 and 1.7 aid in the characterization of the nature of the population. In Chapters 5 and 6 we discuss a property of the population called its **distribution**. While a more rigorous definition of a distribution or **probability distribution** will be given later in the text, at this point one can view it as what would be seen in Figure 1.7 in the limit as the size of the sample becomes larger.

A distribution is said to be **symmetric** if it can be folded along a vertical axis so that the two sides coincide. A distribution that lacks symmetry with respect to a vertical axis is said to be **skewed**. The distribution illustrated in Figure 1.8(a) is said to be skewed to the right since it has a long right tail and a much shorter left tail. In Figure 1.8(b) we see that the distribution is symmetric, while in Figure 1.8(c) it is skewed to the left.

If we rotate a stem-and-leaf plot counterclockwise through an angle of 90°, we observe that the resulting columns of leaves form a picture that is similar to a histogram. Consequently, if our primary purpose in looking at the data is to determine the general shape or form of the distribution, it will seldom be necessary

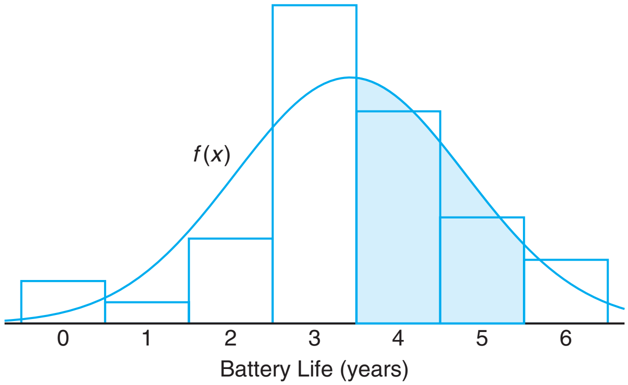

Figure 1.7: Estimating frequency distribution.

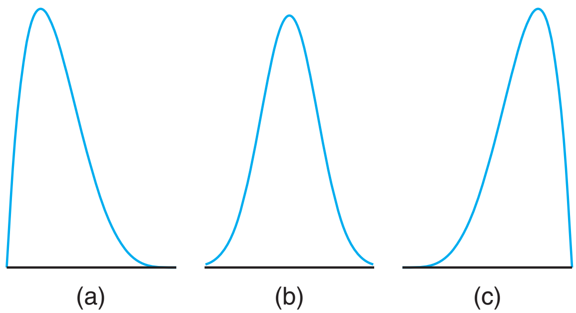

Figure 1.8: Skewness of data.

to construct a relative frequency histogram.

Box-and-Whisker Plot or Box Plot

Another display that is helpful for reflecting properties of a sample is the **box-and-whisker plot**. This plot encloses the *interquartile range* of the data in a box that has the median displayed within. The interquartile range has as its extremes the 75th percentile (upper quartile) and the 25th percentile (lower quartile). In addition to the box, "whiskers" extend, showing extreme observations in the sample. For reasonably large samples, the display shows center of location, variability, and the degree of asymmetry.

In addition, a variation called a **box plot** can provide the viewer with information regarding which observations may be **outliers**. Outliers are observations that are considered to be unusually far from the bulk of the data. There are many statistical tests that are designed to detect outliers. Technically, one may view an outlier as being an observation that represents a "rare event" (there is a small probability of obtaining a value that far from the bulk of the data). The concept of outliers resurfaces in Chapter 12 in the context of regression analysis.

The visual information in the box-and-whisker plot or box plot is not intended to be a formal test for outliers. Rather, it is viewed as a diagnostic tool. While the determination of which observations are outliers varies with the type of software that is used, one common procedure is to use a **multiple of the interquartile range**. For example, if the distance from the box exceeds 1.5 times the interquartile range (in either direction), the observation may be labeled an outlier.

Example 1.5: Nicotine content was measured in a random sample of 40 cigarettes. The data are displayed in Table 1.8.

Table 1.8: Nicotine Data for Example 1.5

1.09	1.92	2.31	1.79	2.28	1.74	1.47	1.97
0.85	1.24	1.58	2.03	1.70	2.17	2.55	2.11
1.86	1.90	1.68	1.51	1.64	0.72	1.69	1.85
1.82	1.79	2.46	1.88	2.08	1.67	1.37	1.93
1.40	1.64	2.09	1.75	1.63	2.37	1.75	1.69

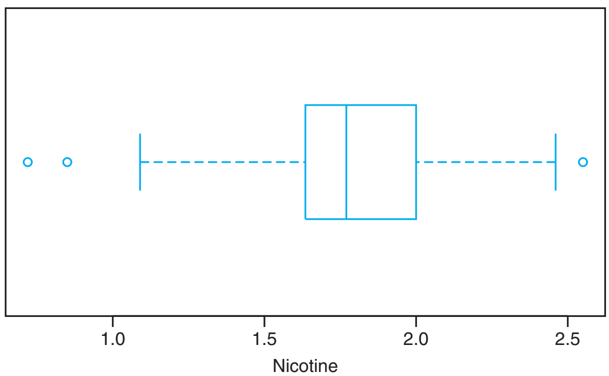

Figure 1.9: Box-and-whisker plot for Example 1.5.

Figure 1.9 shows the box-and-whisker plot of the data, depicting the observations 0.72 and 0.85 as mild outliers in the lower tail, whereas the observation 2.55 is a mild outlier in the upper tail. In this example, the interquartile range is 0.365, and 1.5 times the interquartile range is 0.5475. Figure 1.10, on the other hand, provides a stem-and-leaf plot.

Example 1.6: Consider the data in Table 1.9, consisting of 30 samples measuring the thickness of paint can "ears" (see the work by Hogg and Ledolter, 1992, in the Bibliography). Figure 1.11 depicts a box-and-whisker plot for this asymmetric set of data. Notice that the left block is considerably larger than the block on the right. The median is 35. The lower quartile is 31, while the upper quartile is 36. Notice also that the extreme observation on the right is farther away from the box than the extreme observation on the left. There are no outliers in this data set.

```
The decimal point is 1 digit(s) to the left of the |
 7 | 2
 8 | 5
 9 |
10 | 9
11 |
12 | 4
13 | 7
14 | 07
15 | 18
16 | 3447899
17 | 045599
18 | 2568
19 | 0237
20 | 389
21 | 17
22 | 8
23 | 17
24 | 6
25 | 5
```

Figure 1.10: Stem-and-leaf plot for the nicotine data.

Table 1.9: Data for Example 1.6

Sample	**Measurements**	**Sample**	**Measurements**
1	29 36 39 34 34	16	35 30 35 29 37
2	29 29 28 32 31	17	40 31 38 35 31
3	34 34 39 38 37	18	35 36 30 33 32
4	35 37 33 38 41	19	35 34 35 30 36
5	30 29 31 38 29	20	35 35 31 38 36
6	34 31 37 39 36	21	32 36 36 32 36
7	30 35 33 40 36	22	36 37 32 34 34
8	28 28 31 34 30	23	29 34 33 37 35
9	32 36 38 38 35	24	36 36 35 37 37
10	35 30 37 35 31	25	36 30 35 33 31
11	35 30 35 38 35	26	35 30 29 38 35
12	38 34 35 35 31	27	35 36 30 34 36
13	34 35 33 30 34	28	35 30 36 29 35
14	40 35 34 33 35	29	38 36 35 31 31
15	34 35 38 35 30	30	30 34 40 28 30

There are additional ways that box-and-whisker plots and other graphical displays can aid the analyst. Multiple samples can be compared graphically. Plots of data can suggest relationships between variables. Graphs can aid in the detection of anomalies or outlying observations in samples.

There are other types of graphical tools and plots that are used. These are discussed in Chapter 8 after we introduce additional theoretical details.

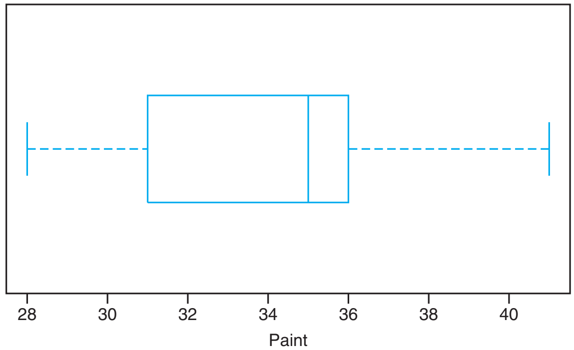

Figure 1.11: Box-and-whisker plot for thickness of paint can "ears."

Other Distinguishing Features of a Sample

There are features of the distribution or sample other than measures of center of location and variability that further define its nature. For example, while the median divides the data (or distribution) into two parts, there are other measures that divide parts or pieces of the distribution that can be very useful. Separation is made into four parts by *quartiles*, with the third quartile separating the upper quarter of the data from the rest, the second quartile being the median, and the first quartile separating the lower quarter of the data from the rest. The distribution can be even more finely divided by computing percentiles of the distribution. These quantities give the analyst a sense of the so-called *tails* of the distribution (i.e., values that are relatively extreme, either small or large). For example, the 95th percentile separates the highest 5% from the bottom 95%. Similar definitions prevail for extremes on the lower side or *lower tail* of the distribution. The 1st percentile separates the bottom 1% from the rest of the distribution. The concept of percentiles will play a major role in much that will be covered in future chapters.

1.7 General Types of Statistical Studies: Designed Experiment, Observational Study, and Retrospective Study

In the foregoing sections we have emphasized the notion of sampling from a population and the use of statistical methods to learn or perhaps affirm important information about the population. The information sought and learned through the use of these statistical methods can often be influential in decision making and problem solving in many important scientific and engineering areas. As an illustration, Example 1.3 describes a simple experiment in which the results may provide an aid in determining the kinds of conditions under which it is not advisable to use a particular aluminum alloy that may have a dangerous vulnerability to corrosion. The results may be of use not only to those who produce the alloy, but also to the customer who may consider using it. This illustration, as well as many more that appear in Chapters 13 through 15, highlights the concept of designing or controlling experimental conditions (combinations of coating conditions and humidity) of interest to learn about some characteristic or measurement (level of corrosion) that results from these conditions. Statistical methods that make use of measures of central tendency in the corrosion measure, as well as measures of variability, are employed. As the reader will observe later in the text, these methods often lead to a statistical model like that discussed in Section 1.6. In this case, the model may be used to estimate (or predict) the corrosion measure as a function of humidity and the type of coating employed. Again, in developing this kind of model, descriptive statistics that highlight central tendency and variability become very useful.

The information supplied in Example 1.3 illustrates nicely the types of engineering questions asked and answered by the use of statistical methods that are employed through a designed experiment and presented in this text. They are

(i) What is the nature of the impact of relative humidity on the corrosion of the aluminum alloy within the range of relative humidity in this experiment?

(ii) Does the chemical corrosion coating reduce corrosion levels and can the effect be quantified in some fashion?

(iii) Is there **interaction** between coating type and relative humidity that impacts their influence on corrosion of the alloy? If so, what is its interpretation?

What Is Interaction?

The importance of questions (i) and (ii) should be clear to the reader, as they deal with issues important to both producers and users of the alloy. But what about question (iii)? The concept of *interaction* will be discussed at length in Chapters 14 and 15. Consider the plot in Figure 1.3. This is an illustration of the detection of interaction between two **factors** in a simple designed experiment. Note that the lines connecting the sample means are not parallel. **Parallelism** would have indicated that the effect (seen as a result of the slope of the lines) of relative humidity is the same, namely a negative effect, for both an uncoated condition and the chemical corrosion coating. Recall that the negative slope implies that corrosion becomes more pronounced as humidity rises. Lack of parallelism implies an interaction between coating type and relative humidity. The nearly "flat" line for the corrosion coating as opposed to a steeper slope for the uncoated condition suggests that *not only is the chemical corrosion coating beneficial (note the displacement between the lines), but the presence of the coating renders the effect of humidity negligible.* Clearly all these questions are very important to the effect of the two individual factors and to the interpretation of the interaction, if it is present.

Statistical models are extremely useful in answering questions such as those listed in (i), (ii), and (iii), where the data come from a designed experiment. But one does not always have the luxury or resources to employ a designed experiment. For example, there are many instances in which the conditions of interest to the scientist or engineer cannot be implemented simply because the *important factors cannot be controlled.* In Example 1.3, the relative humidity and coating type (or lack of coating) are quite easy to control. This of course is the defining feature of a designed experiment. In many fields, factors that need to be studied cannot be controlled for any one of various reasons. Tight control as in Example 1.3 allows the analyst to be confident that any differences found (for example, in corrosion levels) are due to the factors under control. As a second illustration, consider Exercise 1.6 on page 13. Suppose in this case 24 specimens of silicone rubber are selected and 12 assigned to each of the curing temperature levels. The temperatures are controlled carefully, and thus this is an example of a designed experiment with a **single factor** being curing temperature. Differences found in the mean tensile strength would be assumed to be attributed to the different curing temperatures.

What If Factors Are Not Controlled?

Suppose there are no factors controlled and *no random assignment* of fixed treatments to experimental units and yet there is a need to glean information from a data set. As an illustration, consider a study in which interest centers around the relationship between blood cholesterol levels and the amount of sodium measured in the blood. A group of individuals were monitored over time for both blood cholesterol and sodium. Certainly some useful information can be gathered from such a data set. However, it should be clear that there certainly can be no strict control of blood sodium levels. Ideally, the subjects should be divided randomly into two groups, with one group assigned a specific high level of blood sodium and the other a specific low level of blood sodium. Obviously this cannot be done. Clearly changes in cholesterol can be experienced because of changes in one of a number of other factors that were not controlled. This kind of study, without factor control, is called an **observational study**. Much of the time it involves a situation in which subjects are observed across time.

Biological and biomedical studies are often by necessity observational studies. However, observational studies are not confined to those areas. For example, consider a study that is designed to determine the influence of ambient temperature on the electric power consumed by a chemical plant. Clearly, levels of ambient temperature cannot be controlled, and thus the data structure can only be a monitoring of the data from the plant over time.

It should be apparent that the striking difference between a well-designed experiment and observational studies is the difficulty in determination of true cause and effect with the latter. Also, differences found in the fundamental response (e.g., corrosion levels, blood cholesterol, plant electric power consumption) may be due to other underlying factors that were not controlled. Ideally, in a designed experiment the *nuisance factors* would be equalized via the randomization process. Certainly changes in blood cholesterol could be due to fat intake, exercise activity, and so on. Electric power consumption could be affected by the amount of product produced or even the purity of the product produced.

Another often ignored disadvantage of an observational study when compared to carefully designed experiments is that, unlike the latter, observational studies are at the mercy of nature, environmental or other uncontrolled circumstances that impact the ranges of factors of interest. For example, in the biomedical study regarding the influence of blood sodium levels on blood cholesterol, it is possible that there is indeed a strong influence but the particular data set used did not involve enough observed variation in sodium levels because of the nature of the subjects chosen. Of course, in a designed experiment, the analyst chooses and controls ranges of factors.

A third type of statistical study which can be very useful but has clear disadvantages when compared to a designed experiment is a **retrospective study**. This type of study uses strictly **historical data**, data taken over a specific period of time. One obvious advantage of retrospective data is that there is reduced cost in collecting the data. However, as one might expect, there are clear disadvantages.

(i) Validity and reliability of historical data are often in doubt.

(ii) If time is an important aspect of the structure of the data, there may be data missing.

(iii) There may be errors in collection of the data that are not known.

(iv) Again, as in the case of observational data, there is no control on the ranges of the measured variables (the factors in a study). Indeed, the ranges found in historical data may not be relevant for current studies.

In Section 1.6, some attention was given to modeling of relationships among variables. We introduced the notion of regression analysis, which is covered in Chapters 11 and 12 and is illustrated as a form of data analysis for designed experiments discussed in Chapters 14 and 15. In Section 1.6, a model relating population mean tensile strength of cloth to percentages of cotton was used for illustration, where 20 specimens of cloth represented the experimental units. In that case, the data came from a simple designed experiment where the individual cotton percentages were selected by the scientist.

Often both observational data and retrospective data are used for the purpose of observing relationships among variables through model-building procedures discussed in Chapters 11 and 12. While the advantages of designed experiments certainly apply when the goal is statistical model building, there are many areas in which designing of experiments is not possible. Thus, *observational or historical data must be used*. We refer here to a historical data set that is found in Exercise 12.5 on page 450. The goal is to build a model that will result in an equation or relationship that relates monthly electric power consumed to average ambient temperature x_1, the number of days in the month x_2, the average product purity x_3, and the tons of product produced x_4. The data are the past year's historical data.

Exercises

1.13 A manufacturer of electronic components is interested in determining the lifetime of a certain type of battery. A sample, in hours of life, is as follows:

$$123, 116, 122, 110, 175, 126, 125, 111, 118, 117.$$

(a) Find the sample mean and median.

(b) What feature in this data set is responsible for the substantial difference between the two?

1.14 A tire manufacturer wants to determine the inner diameter of a certain grade of tire. Ideally, the diameter would be 570 mm. The data are as follows:

$$572, 572, 573, 568, 569, 575, 565, 570.$$

(a) Find the sample mean and median.

(b) Find the sample variance, standard deviation, and range.

(c) Using the calculated statistics in parts (a) and (b), can you comment on the quality of the tires?

1.15 Five independent coin tosses result in $HHHHH$. It turns out that if the coin is fair the probability of this outcome is $(1/2)^5 = 0.03125$. Does this produce strong evidence that the coin is not fair? Comment and use the concept of P-value discussed in Section 1.1.

1.16 Show that the n pieces of information in $\sum_{i=1}^{n}(x_i - \bar{x})^2$ are not independent; that is, show that

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

1.17 A study of the effects of smoking on sleep patterns is conducted. The measure observed is the time, in minutes, that it takes to fall asleep. These data are obtained:

Smokers:	69.3	56.0	22.1	47.6
	53.2	48.1	52.7	34.4
	60.2	43.8	23.2	13.8
Nonsmokers:	28.6	25.1	26.4	34.9
	29.8	28.4	38.5	30.2
	30.6	31.8	41.6	21.1
	36.0	37.9	13.9	

(a) Find the sample mean for each group.

(b) Find the sample standard deviation for each group.

(c) Make a dot plot of the data sets A and B on the same line.

(d) Comment on what kind of impact smoking appears to have on the time required to fall asleep.

1.18 The following scores represent the final examination grades for an elementary statistics course:

23	60	79	32	57	74	52	70	82
36	80	77	81	95	41	65	92	85
55	76	52	10	64	75	78	25	80
98	81	67	41	71	83	54	64	72
88	62	74	43	60	78	89	76	84
48	84	90	15	79	34	67	17	82
69	74	63	80	85	61			

(a) Construct a stem-and-leaf plot for the examination grades in which the stems are $1, 2, 3, \ldots, 9$.

(b) Construct a relative frequency histogram, draw an estimate of the graph of the distribution, and discuss the skewness of the distribution.

(c) Compute the sample mean, sample median, and sample standard deviation.

1.19 The following data represent the length of life in years, measured to the nearest tenth, of 30 similar fuel pumps:

2.0	3.0	0.3	3.3	1.3	0.4
0.2	6.0	5.5	6.5	0.2	2.3
1.5	4.0	5.9	1.8	4.7	0.7
4.5	0.3	1.5	0.5	2.5	5.0
1.0	6.0	5.6	6.0	1.2	0.2

(a) Construct a stem-and-leaf plot for the life in years of the fuel pumps, using the digit to the left of the decimal point as the stem for each observation.

(b) Set up a relative frequency distribution.

(c) Compute the sample mean, sample range, and sample standard deviation.

1.20 The following data represent the length of life, in seconds, of 50 fruit flies subject to a new spray in a controlled laboratory experiment:

17	20	10	9	23	13	12	19	18	24
12	14	6	9	13	6	7	10	13	7
16	18	8	13	3	32	9	7	10	11
13	7	18	7	10	4	27	19	16	8
7	10	5	14	15	10	9	6	7	15

(a) Construct a double-stem-and-leaf plot for the life span of the fruit flies using the stems $0\star$, $0\cdot$, $1\star$, $1\cdot$, $2\star$, $2\cdot$, and $3\star$ such that stems coded by the symbols $\star$ and $\cdot$ are associated, respectively, with leaves 0 through 4 and 5 through 9.

(b) Set up a relative frequency distribution.

(c) Construct a relative frequency histogram.

(d) Find the median.

1.21 The lengths of power failures, in minutes, are recorded in the following table.

22	18	135	15	90	78	69	98	102
83	55	28	121	120	13	22	124	112
70	66	74	89	103	24	21	112	21
40	98	87	132	115	21	28	43	37
50	96	118	158	74	78	83	93	95

(a) Find the sample mean and sample median of the power-failure times.

(b) Find the sample standard deviation of the power-failure times.

1.22 The following data are the measures of the diameters of 36 rivet heads in 1/100 of an inch.

6.72	6.77	6.82	6.70	6.78	6.70	6.62	6.75
6.66	6.66	6.64	6.76	6.73	6.80	6.72	6.76
6.76	6.68	6.66	6.62	6.72	6.76	6.70	6.78
6.76	6.67	6.70	6.72	6.74	6.81	6.79	6.78
6.66	6.76	6.76	6.72				

(a) Compute the sample mean and sample standard deviation.

(b) Construct a relative frequency histogram of the data.

(c) Comment on whether or not there is any clear indication that the sample came from a population that has a bell-shaped distribution.

1.23 The hydrocarbon emissions at idling speed in parts per million (ppm) for automobiles of 1980 and 1990 model years are given for 20 randomly selected cars.

1980 models:

141	359	247	940	882	494	306	210	105	880
200	223	188	940	241	190	300	435	241	380

1990 models:

140	160	20	20	223	60	20	95	360	70
220	400	217	58	235	380	200	175	85	65

(a) Construct a dot plot as in Figure 1.1.

(b) Compute the sample means for the two years and superimpose the two means on the plots.

(c) Comment on what the dot plot indicates regarding whether or not the population emissions changed from 1980 to 1990. Use the concept of variability in your comments.

1.24 The following are historical data on staff salaries (dollars per pupil) for 30 schools sampled in the eastern part of the United States in the early 1970s.

3.79	2.99	2.77	2.91	3.10	1.84	2.52	3.22
2.45	2.14	2.67	2.52	2.71	2.75	3.57	3.85
3.36	2.05	2.89	2.83	3.13	2.44	2.10	3.71
3.14	3.54	2.37	2.68	3.51	3.37		

(a) Compute the sample mean and sample standard deviation.

(b) Construct a relative frequency histogram of the data.

(c) Construct a stem-and-leaf display of the data.

1.25 The following data set is related to that in Exercise 1.24. It gives the percentages of the families that are in the upper income level, for the same individual schools in the same order as in Exercise 1.24.

72.2	31.9	26.5	29.1	27.3	8.6	22.3	26.5
20.4	12.8	25.1	19.2	24.1	58.2	68.1	89.2
55.1	9.4	14.5	13.9	20.7	17.9	8.5	55.4
38.1	54.2	21.5	26.2	59.1	43.3		

(a) Calculate the sample mean.

(b) Calculate the sample median.

(c) Construct a relative frequency histogram of the data.

(d) Compute the 10% trimmed mean. Compare with the results in (a) and (b) and comment.

1.26 Suppose it is of interest to use the data sets in Exercises 1.24 and 1.25 to derive a model that would predict staff salaries as a function of percentage of families in a high income level for current school systems. Comment on any disadvantage in carrying out this type of analysis.

1.27 A study is done to determine the influence of the wear, y, of a bearing as a function of the load, x, on the bearing. A designed experiment is used for this study. Three levels of load were used, 700 lb, 1000 lb, and 1300 lb. Four specimens were used at each level, and the sample means were, respectively, 210, 325, and 375.

(a) Plot average wear against load.

(b) From the plot in (a), does it appear as if a relationship exists between wear and load?

(c) Suppose we look at the individual wear values for each of the four specimens at each load level (see the data that follow). Plot the wear results for all specimens against the three load values.

(d) From your plot in (c), does it appear as if a clear relationship exists? If your answer is different from that in (b), explain why.

	x		
	700	1000	1300
y_1	145	250	150
y_2	105	195	180
y_3	260	375	420
y_4	330	480	750
	$\bar{y}_1 = 210$	$\bar{y}_2 = 325$	$\bar{y}_3 = 375$

1.28 Many manufacturing companies in the United States and abroad use molded parts as components of a process. Shrinkage is often a major problem. Thus, a molded die for a part is built larger than nominal size to allow for part shrinkage. In an injection molding study it is known that the shrinkage is influenced by many factors, among which are the injection velocity in ft/sec and mold temperature in °C. The following two data sets show the results of a designed experiment in which injection velocity was held at two levels (low and high) and mold temperature was held constant at a low level. The shrinkage is measured in cm $\times$ 10^4.

Shrinkage values at low injection velocity:

72.68	72.62	72.58	72.48	73.07
72.55	72.42	72.84	72.58	72.92

Shrinkage values at high injection velocity:

71.62	71.68	71.74	71.48	71.55
71.52	71.71	71.56	71.70	71.50

(a) Construct a dot plot of both data sets on the same graph. Indicate on the plot both shrinkage means, that for low injection velocity and high injection velocity.

(b) Based on the graphical results in (a), using the location of the two means and your sense of variability, what do you conclude regarding the effect of injection velocity on shrinkage at low mold temperature?

1.29 Use the data in Exercise 1.24 to construct a box plot.

1.30 Below are the lifetimes, in hours, of fifty 40-watt, 110-volt internally frosted incandescent lamps, taken from forced life tests:

919	1196	785	1126	936	918
1156	920	948	1067	1092	1162
1170	929	950	905	972	1035
1045	855	1195	1195	1340	1122
938	970	1237	956	1102	1157
978	832	1009	1157	1151	1009
765	958	902	1022	1333	811
1217	1085	896	958	1311	1037
702	923				

Construct a box plot for these data.

1.31 Consider the situation of Exercise 1.28. But now use the following data set, in which shrinkage is measured once again at low injection velocity and high injection velocity. However, this time the mold temperature is raised to a high level and held constant.

Shrinkage values at low injection velocity:

76.20	76.09	75.98	76.15	76.17
75.94	76.12	76.18	76.25	75.82

Shrinkage values at high injection velocity:

93.25	93.19	92.87	93.29	93.37
92.98	93.47	93.75	93.89	91.62

(a) As in Exercise 1.28, construct a dot plot with both data sets on the same graph and identify both means (i.e., mean shrinkage for low injection velocity and for high injection velocity).

(b) As in Exercise 1.28, comment on the influence of injection velocity on shrinkage for high mold temperature. Take into account the position of the two means and the variability around each mean.

(c) Compare your conclusion in (b) with that in (b) of Exercise 1.28 in which mold temperature was held at a low level. Would you say that there is an interaction between injection velocity and mold temperature? Explain.

1.32 Use the results of Exercises 1.28 and 1.31 to create a plot that illustrates the interaction evident from the data. Use the plot in Figure 1.3 in Example 1.3 as a guide. Could the type of information found in Exercises 1.28 and 1.31 have been found in an observational study in which there was no control on injection velocity and mold temperature by the analyst? Explain why or why not.

1.33 **Group Project:** Collect the shoe size of everyone in the class. Use the sample means and variances and the types of plots presented in this chapter to summarize any features that draw a distinction between the distributions of shoe sizes for males and females. Do the same for the height of everyone in the class.