CHAPTER 3

Describing Data: Summary Measures

PREDICTORS OF SUCCESSFUL MOVIES

The movie industry is a high-profile industry with a highly variable revenue stream. In 1998, U.S. moviegoers spent close to $7 billion at the box office alone. With this much money at stake, it is not surprising that movie studios are interested in knowing what variables are useful for predicting a movie’s financial success. The article by Simonoff and Sparrow (2000) examines this issue for 311 movies released in 1998 and late 1997. (They obtained their data from a public Web site at http://www.imdb.com.) Although it is preferable to examine movie profits, the costs of making movies are virtually impossible to obtain. Therefore, the authors focused instead on revenues—specifically, the total U.S. domestic gross revenue for each film.

Simonoff and Sparrow obtained prerelease information on a number of variables that were thought to be possible predictors of gross revenue. (By “prerelease,” we mean that this information is known about a film before the
These variables include: (1) the genre of the film, categorized as action, children's, comedy, documentary, drama, horror, science fiction, or thriller; (2) the Motion Picture Association of America (MPAA) rating of the film, categorized as G (general audiences), PG (parental guidance suggested), PG-13 (possibly unsuitable for children under 13), R (children not admitted unless accompanied by an adult), NC-17 (no one under 17 admitted), or U (unrated); (3) the country of origin of the movie, categorized as United States, English-speaking but non–United States, or non–English-speaking; (4) number of actors and actresses in the movie who were listed in Entertainment Weekly's lists of the 25 Best Actors and 25 Best Actresses, as of 1998; (5) number of actors and actresses in the movie who were among the top 20 actors and top 20 actresses in average box office gross per movie in their careers; (6) whether the movie was a sequel; (7) whether the movie was released before a holiday weekend; (8) whether the movie was released during the Christmas season; and (9) whether the movie was released during the summer season.

To get a sense of whether these variables are related to gross revenue, we could calculate a lot of summary measures and create numerous tables. However, we agree with Simonoff and Sparrow that the information is best presented in a series of side-by-side boxplots, a type of graph we will introduce in this chapter. (See Figure 3.1.) These boxplots are slightly different from the versions we describe in this book, but they accomplish exactly the same purpose. (There are two differences: First, their boxplots are vertical; ours are horizontal. Second, their boxplots capture an extra piece of information—the widths of their boxes are proportional to the square roots of the sample sizes, so that wide boxes correspond to categories with more movies. In contrast, the heights of our boxes carry no information about sample size.) Basically, each box and the lines and points extending above and below it indicate the distribution of gross revenues for any category. The box itself, from bottom to top, captures the middle 50% of the revenues in the category, the line in the middle of the box represents the median revenue, and the lines and dots indicate possible skewness and outliers.

These particular boxplots indicate some interesting and possibly surprising information about the movie business. First, almost all of the boxplots indicate a high degree of variability and positive skewness, where there are a few movies that gross extremely large amounts compared to the “typical” movies in the category. Second, genre certainly makes a difference. There are more comedies and dramas (wider boxes), but they typically gross considerably less than action, children's, and science fiction films. Third, the same is true of R-rated movies compared to movies rated G, PG, or PG-13—there are more of them, but they typically gross much less. Fourth, U.S. movies do considerably better than foreign movies. Fifth, it helps to have stars, although there are quite a few “sleepers” that succeed without having big-name stars. Sixth, sequels do better, presumably reflecting the success of the earlier films. Finally, the release date makes a big difference. Movies released before holidays, during the Christmas season, or during the summer season tend to have larger gross revenues. Indeed, as Simonoff and Sparrow discuss, movie studios compete fiercely for the best release dates.

Are these prerelease variables sufficient to predict gross revenues accurately? As we might expect from the amount of variability in most of the boxplots in Figure 3.1, the answer is “no.” Many intangible factors evidently determine the ultimate success of a movie, so that some, such as There's Something About Mary, do much better than expected, and others, such as Godzilla, do worse than expected. We revisit this movie data set in the chapter opener to Chapter 12. There, we see how Simonoff and Sparrow use “multiple regression” to predict gross revenue—with only limited success.
3.1 INTRODUCTION

In Chapter 2 we summarize data mainly with tables and graphs. It is often useful to summarize data even further with a few well-chosen summary measures. In this chapter we learn the most frequently used numerical summary measures. These include measures for describing a single variable, such as the mean, median, and standard deviation, plus a
Chapter 3 Describing Data: Summary Measures

couple of measures, correlation and covariance, for describing the potential relationship between two variables. Using these numerical summary measures, we then discuss an additional graph called a boxplot. Boxplots are useful for describing a single variable or comparing two or more related variables.

We conclude this chapter by examining three relatively complex examples. Here we are able to put all of the descriptive tools we have learned to good use. These examples are typical of the large-scale data sets business managers face on a continual basis. Only by looking at the data from a number of points of view can we discover the information and patterns hidden in the data.

3.2 MEASURES OF CENTRAL LOCATION

Most of the numerical summary measures we consider here are for a single variable. Each describes some aspect of the distribution of the variable. That is, each number describes one feature of the distribution that we see graphically in a histogram. We begin with measures of central location. The three most commonly used measures are the mean, median, and mode, each of which gives a slightly different interpretation to the term central location.

3.2.1 The Mean

The mean is the average of all values of a variable. If the data represent a sample from some larger population, we call this measure the sample mean and denote it by $\bar{X}$ (pronounced “x-bar”). If the data represent the entire population, we call it the population mean and denote it by $\mu$. This distinction is not important in this chapter, but it will become relevant in later chapters when we discuss statistical inference. In either case the formula for the mean is given by equation (3.1).

\[
\text{Mean} = \frac{\sum_{i=1}^{n} X_i}{n} \quad (3.1)
\]

Here $n$ is the number of observations and $X_i$ is the value of observation $i$. Equation (3.1) simply says to add all the observations and divide by $n$, the number of observations.

To obtain the mean in Excel, we use the AVERAGE function on the appropriate range, as illustrated in the following example.

3.1 Summarizing Starting Salaries for Business Undergraduates

The file Salary.xls lists starting salaries for 190 graduates from an undergraduate school of business. The data are in the range named Salary on a sheet called Data. Find the average of all salaries.

Objective To use Excel’s AVERAGE function to summarize salaries.

Solution

Figure 3.2 includes a number of summary measures produced by Excel’s built-in functions. In particular, we calculate the mean salary by entering the formula

= AVERAGE(Salary)
in cell B6. It is nearly $30,000. (The other summary measures in Figure 3.2 are obtained using other Excel functions, as is discussed subsequently.)

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Summary measures using Excel functions</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Count</td>
</tr>
<tr>
<td>4</td>
<td>Minimum</td>
</tr>
<tr>
<td>5</td>
<td>Maximum</td>
</tr>
<tr>
<td>6</td>
<td>Average</td>
</tr>
<tr>
<td>7</td>
<td>Median</td>
</tr>
<tr>
<td>8</td>
<td>Lower quartile</td>
</tr>
<tr>
<td>9</td>
<td>Upper quartile</td>
</tr>
<tr>
<td>10</td>
<td>5-percentile</td>
</tr>
<tr>
<td>11</td>
<td>95-percentile</td>
</tr>
<tr>
<td>12</td>
<td>Range</td>
</tr>
<tr>
<td>13</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>14</td>
<td>Variance</td>
</tr>
</tbody>
</table>

The mean in Example 3.1 is a “representative” measure of central location because the distribution of salaries is nearly symmetric. (You can check this statement by constructing a histogram of salaries.) However, the mean is often misleading because of skewness. For example, if a few of the undergraduates got abnormally high salaries (over $100,000, say), these large values would tend to inflate the mean and make it unrepresentative of the majority of the salaries. In such cases, the median is often a more appropriate measure of central location.

### 3.2.2 The Median

The median is the “middle” observation when the data are arranged from smallest to largest. If there is an odd number of observations, the median is the middle observation. For example, if there are nine observations, the median is the fifth smallest (or fifth largest) observation. If there is an even number of observations, the median is usually defined to be the average of the two middle observations. For example, if there are 10 observations, the median is usually defined to be the average of the fifth and sixth smallest values.

We calculate the median salary in Example 3.1 by entering the formula

$$=\text{MEDIAN}(\text{Salary})$$

in cell B7. (See Figure 3.2.) Its value is again approximately $30,000, almost the same as the mean. This is typical of symmetric distributions, but it is not true for skewed distributions. For example, if a few graduates received abnormally large salaries, the mean would be affected by them, but the median would not be affected at all. It would still represent the “middle” of the distribution.

The median is the middle observation (for an odd number of observations) or the average of the middle two observations (for an even number of observations) after the observations have been sorted from low to high.
### 3.2.3 The Mode

The **mode** is the most frequently occurring value. If the values are essentially continuous, as with the salaries in Example 3.1, then the mode is essentially irrelevant. There is typically no *single* value that occurs more than once, or there are at best a few ties for the most frequently occurring value. In either case the mode is not likely to provide much information. However, the following example illustrates where the mode can be useful.

The **mode** is the most frequently occurring observation.

---

#### Example 3.2: Shoe Size Purchased at a Shoe Store

The file `Shoes.xls` lists shoe sizes purchased at a shoe store. What is the store’s best-seller?

**Objective**

To use Excel’s MODE function to find the best-selling shoe size.

**Solution**

Shoe sizes come in discrete increments, rather than a continuum, so it makes sense to find the mode, the size that is requested most often. This can be done with Excel’s MODE function, which shows that size 11 is the most frequently purchased size. This is also apparent from the histogram in Figure 3.3, where the category for size 11 corresponds to the highest bar. (We deliberately overrode the default StatTools bins for this histogram. The trick is to ask StatTools for 15 bins, with minimum 6.25 and maximum 13.75. Then a typical bin has length 0.5, such as 10.25 to 10.75. Its midpoint, 10.5, is then the shoe size.)

---

**Figure 3.3** Distribution of Shoe Sizes

![Histogram of Shoe Size](image)
3.3 QUARTILES AND PERCENTILES

The median splits the data in half. It is sometimes called the 50th percentile because (approximately) half of the data are below the median. It is also called the second quartile, because if we divide the data into four parts, then the median separates the lower two parts from the upper two. We can also find other percentiles and quartiles. Some of these appear in Figure 3.2, which summarizes the salary data from Example 3.1.

For example, the 5th and 95th percentiles appear in cells B10 and B11. They are calculated with the formulas

\[ =\text{PERCENTILE(Salary,0.05)} \]

and

\[ =\text{PERCENTILE(Salary,0.95)} \]

They say that 5% of all salaries are below $23,690 (so that 95% are above $23,690), and 95% of all salaries are below $35,810 (so that 5% are above $35,810). These two percentiles (and sometimes others) are frequently quoted.

Similarly, Figure 3.2 lists the lower and upper quartiles in cells B8 and B9. These are the 25th and 75th percentiles, so they can be calculated with the PERCENTILE function. They can also be calculated with the formulas

\[ =\text{QUARTILE(Salary,1)} \]

and

\[ =\text{QUARTILE(Salary,3)} \]

That is, they are the first and third quartiles. (The median is the second quartile.) In short, 25% of the salaries are below $27,325, 25% are above $32,300, and the other 50% are in between.

The difference between the first and third quartiles is called the interquartile range (IQR). It measures the spread between the largest and smallest of the middle half of the data. In the salary example the IQR is $4,975 (= $32,300 − $27,325). We come back to the IQR when we discuss boxplots later in this chapter.

**Excel Tip** Excel uses a rather obscure interpolation formula (which is not spelled out in its online help) for determining percentiles and quartiles. This can be particularly disconcerting—and wrong—for small data sets. For example, Excel reports that the 25th percentile (and the first quartile) of the 10 numbers 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 is 3.25. This does not appear to be consistent with our definition of the 25th percentile, which states that 25% of the numbers are below it and 75% are above it. As another example, it always reports the 25th percentile of 13 numbers as the fourth smallest of these numbers, which seems strange. Therefore, we recommend not putting too much reliance on Excel’s PERCENTILE and QUARTILE functions for small data sets.

3.4 MINIMUM, MAXIMUM, AND RANGE

Three other descriptive measures of a variable are its minimum, maximum, and range. The minimum is the smallest value, the maximum is the largest value, and the range is the difference between the maximum and minimum. These are listed in Figure 3.2 for the
salary data. The minimum and maximum are calculated in cells B4 and B5 with the formulas

\[ =\text{MIN}(\text{Salary}) \]

and

\[ =\text{MAX}(\text{Salary}) \]

The range is then calculated in cell B12 with the formula

\[ =B5-B4 \]

We see that no salary is below $17,100, no salary is above $38,200, and all salaries are contained within an interval of length $21,100.

The minimum and maximum provide bounds on the data set, and the range provides a crude measure of the variability of the data. These measures are often worth reporting, but they can obviously be affected by one or two extreme values. The range, in particular, is usually not as good a measure of variability as the measures discussed next.

### 3.5 MEASURES OF VARIABILITY: VARIANCE AND STANDARD DEVIATION

To really understand a data set, we need to know more than measures of central location; we also need measures of variability. To see this, consider the following example.

**Example 3.3 Variability of Elevator Rail Diameters at Otis Elevator**

Suppose Otis Elevator is going to stop manufacturing elevator rails. Instead, it is going to buy them from an outside supplier. Otis would like each rail to have a diameter of 1 inch. The company has obtained samples of 10 elevator rails from each supplier. These are listed in columns A and B of Figure 3.4. (See the file Otis4.xls.) Which supplier should Otis prefer?

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Diameters from two suppliers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Data</td>
<td>Summary measures</td>
<td>Supplier1</td>
<td>Supplier2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Supplier1</td>
<td>Supplier2</td>
<td>Mean</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1.00</td>
<td>0.96</td>
<td>Median</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.98</td>
<td>1.05</td>
<td>Mode</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1.02</td>
<td>1.00</td>
<td>Variance</td>
<td>0.000133</td>
<td>0.001200</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1.01</td>
<td>0.97</td>
<td>Standard deviation</td>
<td>0.0115</td>
<td>0.0346</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.99</td>
<td>1.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.99</td>
<td>0.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1.00</td>
<td>1.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1.01</td>
<td>0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1.00</td>
<td>1.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.4
Two Samples with Different Amounts of Variability

86 Chapter 3 Describing Data: Summary Measures
**Objective**  
To calculate the variability for two suppliers and choose the one with the least variability.

**Solution**

Observe that the mean, median, and mode are all exactly 1 inch for each supplier. Based on these measures, the two suppliers are equally good and both are right on the mark. It is clear from a glance at the data, however, that supplier 1 is somewhat better than supplier 2. The reason is that supplier 2’s rails exhibit more variability about the mean than supplier 1’s rails. If we want rails to have a diameter of 1 inch, then variability around the mean is bad! ■

The most commonly used measures of variability are the variance and standard deviation. The variance is essentially the average of the squared deviations from the mean. We say “essentially” because there are two versions of variance: the population variance, usually denoted by $\sigma^2$, and the sample variance, usually denoted by $s^2$. The formulas for them are given by equations (3.2) and (3.3). These formulas differ basically in their denominators. Also, their numerical values are practically the same when $n$, the number of observations, is large.

**Formula for Population Variance**

$$\sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n} \quad (3.2)$$

**Formula for Sample Variance**

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n - 1} \quad (3.3)$$

We say that each term in these sums is a squared deviation from the mean, where deviation means difference. Therefore, you can remember variance as the average of the squared deviations from the mean. This is not quite true for the sample variance (because of dividing by $n - 1$, not $n$), but it is nearly true. If we use the supplier 2 data in Figure 3.4 for illustration, the deviations from the mean are $-0.04, 0.05, 0.00, -0.03, 0.00, 0.03, -0.02, 0.02, -0.05, \text{ and } 0.04$; the squares of these deviations are $0.0016, 0.0025, 0.0000, 0.0009, 0.00000, 0.0004, 0.0004, 0.0025, \text{ and } 0.0016$; and the sum of these squares is $0.0108$. Therefore, the population variance is $0.0108/10 = 0.0011$, and the sample variance (the one quoted in the figure) is $0.0108/9 = 0.0012$.

As their names imply, $\sigma^2$ is relevant if the data set includes the entire population, whereas $s^2$ is relevant for a sample from a population. Excel has a built-in function for each. To obtain $\sigma^2$ we use the VARP function; to obtain $s^2$ we use the VAR function. In this chapter we illustrate only the sample variance $s^2$.

The important part about either variance formula is that the variance tends to increase when there is more variability around the mean. Indeed, large deviations from the mean contribute heavily to the variance because they are squared. One consequence of this is that the variance is expressed in squared units (squared dollars, for example) rather than original units. Therefore, a more intuitive measure is the standard deviation, defined as the square root of the variance (3.4). The standard deviation is measured in original units, such as dollars, and, as we discuss shortly, it is much easier to interpret.

**Relationship Between Standard Deviation and Variance**

$$\text{Standard deviation} = \sqrt{\text{Variance}} \quad (3.4)$$
Of course, depending on which variance measure we use, we obtain the corresponding standard deviation, either \( \sigma \) (population) or \( s \) (sample). Excel has built-in functions for each of these: we use STDEV.P for \( \sigma \) and STDEV for \( s \). In this chapter we illustrate only \( s \), but again, the difference is negligible when \( n \) is large.

The variances and standard deviations of the diameters from the two suppliers in Example 3.3 appear in Figure 3.4. To obtain them, enter the formulas

\[
= \text{VAR(A5:A14)}
\]

and

\[
= \text{STDEV(A5:A14)}
\]

in cells E8 and E9, and enter similar formulas for supplier 2 in cells F8 and F9. Because of the relationship between variance and standard deviation, we could also have used the formula

\[
= \text{SQRT(E8)}
\]

in cell E9, but we instead took advantage of the STDEV function.

As we mentioned previously, it is difficult to interpret these variances numerically because they are expressed in squared inches, not inches. All we can say is that the variance from supplier 2 is considerably larger than the variance from supplier 1. The standard deviations, on the other hand, are expressed in inches. The standard deviation for supplier 1 is approximately 0.012 inch, and supplier 2’s standard deviation is approximately three times as large. This is a considerable disparity, as we explain next.

### Interpretation of the Standard Deviation: Empirical Rules

Many data sets follow certain “empirical rules.” In particular, suppose that a histogram of the data is approximately symmetric and “bell shaped.” That is, heights of the bars rise to some peak and then decline. Such behavior is quite common, and it allows us to interpret the standard deviation intuitively. Specifically, we can state that

- approximately 68% of the observations are within 1 standard deviation of the mean, that is, within the interval \( X \pm s \);
- approximately 95% of the observations are within 2 standard deviations of the mean, that is, within the interval \( X \pm 2s \); and
- approximately 99.7%—almost all—of the observations are within 3 standard deviations of the mean, that is, within the interval \( X \pm 3s \).

We illustrate these empirical rules with the following example.

---

**EXAMPLE 3.4 DISTRIBUTION OF MONTHLY DOW RETURNS**

The file Dow.xlsx contains monthly closing prices for the Dow Jones index from January 1947 through January 1993. The monthly returns from the index are also shown, starting with the February 1947 value. Each return is the monthly percentage change (expressed as a decimal) in the index. How well do the empirical rules work for these data?

**Objective**

To illustrate the empirical rules for Dow Jones monthly returns.
In contrast, the measures we have been discussing are relevant for the series of returns, which fluctuates around a stable mean. In Figure 3.7 we first calculate the mean and standard deviation of returns with the AVERAGE and STDEV functions in cells B4 and B5. These indicate an average return of about 0.59% and a standard deviation of about 3.37%. Therefore, the empirical rules (if they apply) imply, for example, that about 2/3 of all returns are within the interval 0.59% ± 3.37%, that is, from −2.78% to 3.95%.
We can use a frequency table to check whether the empirical rules apply to these returns. We first enter the upper limits of suitable categories in the range A8:A15. Although any categories could be chosen, it is convenient to choose values of the form $\overline{X} \pm ks$ as
breakpoints for the categories, where the open-ended categories on either end are “more than 3 standard deviations from the mean.” Here, $k$ is a typical multiple of the standard deviation $s$. We let $k$ have the values 1, 2, and 3. Then the upper limits of the categories can be calculated in column B by entering the formulas

\[ =B4-3*5 \]

and

\[ =B8+B5 \]

in cells B8 and B9, and then copying this latter formula to the range B10:B14. In words, each breakpoint is 1 standard deviation higher than the previous one.

Next, we use the FREQUENCY function to fill in column C. Specifically, we highlight the range C8:C15, type the formula

\[ =\text{FREQUENCY}(\text{Return}, \text{Bins}) \]

and press Ctrl-Shift-Enter. (Here, Return is the range name for the Returns variable, and Bins is the range name of the range B8:B14.)

Finally, we use the frequencies in column C to calculate the actual percentages of returns within $k$ standard deviations of the mean for $k = 1$, $k = 2$, and $k = 3$, and we compare these with percentages from the empirical rules. (See rows 19 and 20.) The agreement between these percentages is not perfect—there are a few more observations within 1 standard deviation of the mean than the empirical rules predict—but in general the empirical rules work quite well.

3.6 OBTAINING SUMMARY MEASURES WITH ADD-INS

In the past few sections we used Excel’s built-in functions (AVERAGE, STDEV, and so on) to calculate a number of summary measures. A quicker way is to use the StatTools add-in. We illustrate the StatTools add-in in the following example.

EXAMPLE 3.5 SUMMARIZING MOVIE STAR SALARIES

Refer again to the Salary.xls data set used in Example 3.1, and find a set of useful summary measures for the salaries.

Objective To use StatTools’s One-Variable Summary Statistics procedure to calculate summary measures of salaries.

Solution

This is easy with StatTools’s One-Variable Summary Statistics procedure. As usual, begin by creating a StatTools data set with the StatTools/Data Set Manager menu item. Then select the StatTools/Summary Measures/One-Variable Summary menu item, select all variables you want to summarize, and select the summary measures you want from the dialog box shown in Figure 3.8. (Here is a handy tip. When you first see the dialog box shown in Figure 3.8, it will probably have all of the options checked, which is probably more than you want. In this case, uncheck the options you don’t want, and then click on the middle “Save” button on the lower left. This allows you to save your setup—the options you want for future analyses.)
A typical output appears in Figure 3.9. It includes many of the summary measures we have discussed, plus a few more. The **mean absolute deviation** is similar to the variance, except that it is an average of the **absolute** (not squared) deviations from the mean. The **kurtosis** and **skewness** indicate the relative peakedness of the distribution and its skewness. These are relatively technical measures that we will not discuss here.
By clicking on any of the cells in column B of Figure 3.9, you see that StatTools provides formulas for the outputs, using its own built-in statistical functions. The effect is that if any of the original data change, the summary measures we just produced change automatically. Finally, note that all outputs are formatted with a certain number of decimal places by default. You can reformat them in a more appropriate manner if you like.

### PROBLEMS

#### Level A

1. A human resources manager at Beta Technologies, Inc., is interested in compiling some statistics on full-time Beta employees. In particular, she is interested in summarizing age, number of years of relevant full-time work experience prior to coming to Beta, number of years of full-time work experience at Beta, number of years of post-secondary education, and salary. Find appropriate summary measures for these factors using data in the file P02_01.xls.

2. A production manager is interested in determining the typical proportion of defective items in a shipment of one of the computer components that her company manufactures. The spreadsheet provided in the file P02_02.xls contains the proportion of defective components for each of 500 randomly selected shipments collected during a 1-month period. Is the mean, median, or mode the most appropriate measure of central location for these data?

3. Business Week’s Guide to the Best Business Schools provides enrollment data on graduate business programs that it rates as the best in the United States. Specifically, this guide reports the percentages of women, minority, and international students enrolled in each of the top programs, as well as the total number of full-time students enrolled in each program. Use the most appropriate measure(s) of central location to find the typical percentage of women, minority, and international students enrolled in these elite programs. These data are contained in the file P02_03.xls.

4. The manager of a local fast-food restaurant is interested in improving the service provided to customers who use the restaurant’s drive-up window. As a first step in this process, the manager asks his assistant to record the time (in minutes) it takes to serve 200 customers at the final window in the facility’s drive-up system. The 200 customer service times given in the file P02_04.xls are all observed during the busiest hour of the day for this fast-food operation.
   a. Compute the mean, median, and mode of this sample of customer service times.
   b. Which of these measures do you believe is the most appropriate one in describing this distribution? Explain the reasoning behind your choice.

5. A finance professor has just given a midterm examination in her corporate finance course. In particular, she is interested in learning how her class of 100 students performed on this exam. The 100 exam scores are given in the file P02_05.xls.
   a. What are the mean and median scores (out of 100 possible points) on this exam?
   b. Explain why the mean and median values are different for these data.

6. Compute the mean, median, and mode of the given set of average annual household income levels of citizens from selected U.S. metropolitan areas in the file P03_06.xls. What can you infer about the shape of this particular income distribution from the computed measures of central location?

7. The operations manager of a toll booth, located at a major exit of a state turnpike, is trying to estimate the typical number of vehicles that arrive at the toll booth during a 1-minute period during the peak of rush-hour traffic. In an effort to estimate this typical throughput value, he records the number of vehicles that arrive at the toll booth over a 1-minute interval commencing at the same time for each of 365 normal weekdays. These data are contained in the file P02_09.xls.
   a. Find the most appropriate measure of the given distribution’s central location.
   b. Is this distribution of arrivals skewed somewhat? Explain.

8. The proportions of high school graduates annually taking the SAT test in each of the 50 states and the District of Columbia are provided in the file P02_10.xls.
   a. Compute the mean, median, and mode of this set of proportions.
   b. Which of these measures do you believe is the most appropriate one in describing this distribution? Explain the reasoning behind your choice.

9. Consider the average time (in minutes) it takes a citizen of each metropolitan area to travel to work and back home each day. Refer to the data given in the file P02_11.xls.
   a. Find the most representative average commute time across this distribution.
10. Five hundred households in a middle-class neighborhood were recently surveyed as a part of an economic development study conducted by the local government. Specifically, for each of the 500 randomly selected households, the survey requested information on several variables, including the household’s level of indebtedness (excluding the value of any home mortgage). These data are provided in the file P02_06.xls.

- a. Find the maximum and minimum debt levels for the households in this sample.
- b. Find the indebtedness levels at each of the 25th, 50th, and 75th percentiles.
- c. Compute and interpret the interquartile range for these data.

11. A real estate agent has gathered data on 150 houses that were recently sold in a suburban community. This data set includes observations for each of the following variables: the appraised value of each house (in thousands of dollars), the selling price of each house (in thousands of dollars), the size of each house (in hundreds of square feet), and the number of bedrooms in each house. Refer to the file P02_07.xls in answering the following questions.

- a. Find the house(s) at the 80th percentile of all sample houses with respect to appraised value.
- b. Find the house(s) at the 80th percentile of all sample houses with respect to selling price.
- c. Find the maximum and minimum sizes (measured in square footage) of all sample houses.
- d. What is the typical number of bedrooms in a recently sold house in this suburban community?

12. The U.S. Department of Transportation regularly publishes the Air Travel Consumer Report, which provides a variety of performance measures of major U.S. commercial airlines. One dimension of performance reported is each airline’s percentage of domestic flights arriving within 15 minutes of the scheduled arrival time at major reporting airports throughout the country. Use these data, given in the file P02_12.xls, to answer the following questions:

- a. Which major U.S. airline has the highest third quartile on-time arrival percentage?
- b. Which major U.S. airline has the lowest first quartile on-time arrival percentage?
- c. Which major U.S. airline has the largest range of on-time arrival percentages?
- d. Which major U.S. airline has the smallest range of on-time arrival percentages?

13. Having computed measures of central location for various numerical attributes of full-time employees, a human resources manager at Beta Technologies, Inc., is now interested in compiling some statistics on the variability of sample data values about their respective means. Using the data provided in the file P02_01.xls, assist this manager by computing sample standard deviations for each of the following numerical variables: age, number of years of relevant full-time work experience prior to coming to Beta, number of years of full-time work experience at Beta, number of years of post-secondary education, and salary. For each variable, explain why the empirical rules apply or do not apply.

14. A production manager is interested in determining the variability of the proportion of defective items in a shipment of one of the computer components that her company manufactures. The spreadsheet provided in the file P02_02.xls contains the proportion of defective components for each of 500 randomly selected shipments collected during a 1-month period. Compute the sample standard deviation of these data. Discuss whether the empirical rules apply.

15. In an effort to provide more consistent customer service, the manager of a local fast-food restaurant would like to know the dispersion of customer service times about their average value for the facility’s drive-up window. The file P02_04.xls contains 200 customer service times, all of which were observed during the busiest hour of the day for this fast-food operation. Use these data to answer the following questions:

- a. Find and interpret the variance and standard deviation of these sample values.
- b. Are the empirical rules applicable for these data? If so, apply these rules and interpret your results. If not, explain why the empirical rules are not applicable here.

16. Compute the standard deviation of the average annual household income levels of citizens from selected U.S. metropolitan areas in the file P03_06.xls. Is it appropriate to apply the empirical rules for these data? Explain.

17. The file P02_26.xls contains annual percentage changes in consumer prices. Are the empirical rules applicable in this case? If so, apply these rules and interpret your results.

18. The diameters of 100 rods produced by Rodco are listed in the file P03_18.xls. Based on these data, you can be approximately 99.7% sure that the diameter of a typical rod will be between what two numbers?

19. The file P03_19.xls contains the thickness (in centimeters) of some mica pieces. A piece meets specifications if it is between 7 and 15 centimeters in thickness.

- a. What fraction of mica pieces meet specifications?
- b. Do the empirical rules appear to be valid for these data?
Level B

20. A finance professor has just given a midterm examination in her corporate finance course. She is now interested in assigning letter grades to the scores earned by her 100 students who took this exam. The top 10% of all scores should receive a grade of A. The next 10% of all scores should be assigned a grade of B. The third 10% of all scores should receive a grade of C. The next 10% of all scores should be assigned a grade of D. All remaining scores should receive a grade of F. The 100 exam scores are given in the file P02_05.xls. Assist this instructor in assigning letter grades to these finance exam scores.

21. The operations manager of a toll booth, located at a major exit of a state turnpike, is trying to estimate the variability of the number of vehicles that arrive at the toll booth during a 1-minute period during the peak of rush-hour traffic. In an effort to estimate this measure of dispersion, he records the number of vehicles that arrive at the toll booth over a 1-minute interval commencing at the same time for each of 365 normal weekdays. These data are contained in the file P02_09.xls.
   a. Is the sample variance (or sample standard deviation) a reliable measure of dispersion for these data? Explain why or why not.
   b. If the sample variance (or sample standard deviation) is not a reliable measure of dispersion for these data, how can this operations manager most appropriately measure the variability of the values in this data set?

22. The file P02_10.xls contains the proportions of high school graduates annually taking the SAT test in each of the 50 states and the District of Columbia. Approximately 95% of these proportions fall between what two fractions?

23. The file P03_23.xls contains the salaries of all Indiana University business school professors.
   a. If you increased every professor’s salary by $1000, what would happen to the mean and median salary?
   b. If you increased every professor’s salary by $1000, what would happen to the sample standard deviation of the salaries?
   c. If you increased everybody’s salary by 5%, what would happen to the sample standard deviation of the salaries?

24. The file P03_24.xls contains a sample of family incomes (in thousands of 1980 dollars) for a set of families sampled in 1980 and 1990. Assume that these families are representative of the whole United States. The Republicans claim that the country was better off in 1990 than 1980, because average income increased. Do you agree?

25. According to the Educational Testing Service, the scores of people taking the SAT were as listed in Table 3.1.
   a. Estimate the average and standard deviation of SAT scores for students whose families made at least $18,000 and for those whose families made no more than $6000. (Hint: Assume all scores in a group are concentrated at the group’s midpoint.)
   b. Do these results have any implications for college admissions?

26. The file P03_26.xls lists the fraction of U.S. men and women of various heights. Use these data to estimate the mean and standard deviation of the height of American men and women. (Hint: Assume all heights in a group are concentrated at the group’s midpoint.)

27. The file P03_27.xls lists the fraction of U.S. men and women of various weights. Use these data to estimate the mean and standard deviation of the weights of U.S. men and women. (Hint: Assume all weights in a group are concentrated at the group’s midpoint.)

3.7 MEASURES OF ASSOCIATION: COVARIANCE AND CORRELATION

All of the summary measures to this point involve a single variable. It is also useful to summarize the relationship between two variables. Specifically, we would like to summarize the type of behavior often observed in a scatterplot. Two such measures are covariance and correlation. We discuss them briefly here and in more depth in later chapters. Each
measures the strength (and direction) of a linear relationship between two numerical variables. Intuitively, the relationship is “strong” if the points in a scatterplot cluster tightly around some straight line. If this straight line rises from left to right, then the relationship is positive and the measures are positive numbers. If it falls from left to right, then the relationship is negative and the measures are negative numbers.

If we want to measure the covariance or correlation between two variables $X$ and $Y$—indeed, even if we just want to form a scatterplot of $X$ versus $Y$—then $X$ and $Y$ must be “paired” variables. That is, they must have the same number of observations, and the $X$ and $Y$ values for any observation should be naturally paired. For example, each observation could be the height and weight for a particular person, the time in a store and the amount purchased for a particular customer, and so on.

With this in mind, let $X_i$ and $Y_i$ be the paired values for observation $i$, and let $n$ be the number of observations. Then the covariance between $X$ and $Y$, denoted by Cov($X$, $Y$), is given by equation (3.5):

\[
\text{Cov}(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}
\]

Scatterplots that rise from lower left to upper right will tend to have positive covariance and correlation. Those that fall from upper left to lower right will tend to have negative covariance and correlation.

Statisticians use the symbol $r$ for the correlation based on sample data. If they want to denote the correlation based on the entire population, they use the symbol $\rho$ (rho).

You probably will never have to use this formula directly—Excel has a built-in COVAR function that does it for you—but the formula does indicate what covariance is all about. It is essentially an average of products of deviations from means. If $X$ and $Y$ vary in the same direction, then when $X$ is above (or below) its mean, $Y$ will also tend to be above (or below) its mean. In either case, the product of deviations will be positive—a positive times a positive or a negative times a negative—so the covariance will be positive. The opposite is true when $X$ and $Y$ vary in opposite directions. Then the covariance will be negative.

The limitation of covariance as a descriptive measure is that it is affected by the units in which $X$ and $Y$ are measured. For example, we can inflate the covariance by a factor of 1000 simply by measuring $X$ in dollars rather than in thousands of dollars. The correlation, denoted by Corr($X$, $Y$), remedies this problem. It is a unitless quantity defined by equation (3.6), where Stdev($X$) and Stdev($Y$) denote the standard deviations of $X$ and $Y$. Again, you’ll probably never have to use this formula for calculations—Excel does it for you with the built-in CORREL function—but it does show that to produce a unitless quantity, we need to divide the covariance by the product of the standard deviations.

The correlation is unaffected by the units of measurement of the two variables, and it is always between $-1$ and $+1$. The closer it is to either of these two extremes, the closer the points in a scatterplot are to some straight line, either in the negative or positive direction. On the other hand, if the correlation is close to 0, then the scatterplot is typically a “cloud” of points with no apparent relationship. However, it is also possible that the points are close to a curve and have a correlation close to 0. This is because correlation is relevant only for measuring linear relationships.

When there are more than two variables in a data set, it is often useful to create a table of covariances and/or correlations. Each value in the table then corresponds to a particular pair of variables. StatTools allows you to do this easily, as illustrated in the following example.
The only relationships that stand out are the positive relationships between salary and cultural expenses and between salary and dining expenses, and the negative relationship between cultural and sports-related expenses. In contrast, there is very little linear relationship between salary and sports expenses or between dining expenses and either culture or sports expenses. To confirm these graphically, we show scatterplots of Salary versus Culture and Culture versus Sports in Figures 3.11 and 3.12. These indicate more intuitively what a correlation of approximately \( \pm 0.5 \) really means.

In general, we point out the following properties that are evident from Figure 3.10:

- The correlation between a variable and itself is always 1.
- The correlation between \( X \) and \( Y \) is the same as the correlation between \( Y \) and \( X \). Therefore, it is sufficient to list the correlations below (or above) the diagonal in the table. (The same is true for covariances.) StatTools provides these options.
- The covariance between a variable and itself is the variance of that variable.

### Example 3.6 Household Spending for Various Categories of Items

A survey questions members of 100 households about their spending habits. The data in the file Expenses.xls represent the salary, expenses for cultural activities, expenses for sports-related activities, and expenses for dining out for each household over the past year. Do these variables appear to be related linearly?

**Objective** To use StatTools’s Correlation and Covariance procedure to measure the relationship between expenses in various categories.

**Solution**

Scatterplots of each variable versus each other variable answer the question quite nicely, but six scatterplots are required, one for each pair. To get a quick indication of possible linear relationships, we can use StatTools to obtain a table of correlations and/or covariances. To do so, make sure a StatTools data set has been created in the usual way, select the StatTools/Summary Measures/Correlation and Covariance menu item, select all four variables of interest, and choose to obtain correlations and covariances. Otherwise, accept all of StatTools's default settings. The tables of correlations and covariances appear in Figure 3.10.

- **Correlation Table**
  - Salary: 1.000
  - Culture: 0.506
  - Sports: -0.081
  - Dining: 0.558

- **Covariance Table**
  - Salary: 91130278.79
  - Culture: 1105845.25
  - Sports: -221238.79
  - Dining: 2590600.81

The only relationships that stand out are the positive relationships between salary and cultural expenses and between salary and dining expenses, and the negative relationship between cultural and sports-related expenses. In contrast, there is very little linear relationship between salary and sports expenses or between dining expenses and either culture or sports expenses. To confirm these graphically, we show scatterplots of Salary versus Culture and Culture versus Sports in Figures 3.11 and 3.12. These indicate more intuitively what a correlation of approximately \( \pm 0.5 \) really means.

In general, we point out the following properties that are evident from Figure 3.10:

- The correlation between a variable and itself is always 1.
- The correlation between \( X \) and \( Y \) is the same as the correlation between \( Y \) and \( X \). Therefore, it is sufficient to list the correlations below (or above) the diagonal in the table. (The same is true for covariances.) StatTools provides these options.
- The covariance between a variable and itself is the variance of that variable.
It is difficult to interpret the magnitudes of the covariances. These depend on the fact that the data are measured in dollars rather than, say, thousands of dollars. It is much easier to interpret the magnitudes of the correlations because they are scaled to be between $-1$ and $+1$.

**Figure 3.11**
Scatterplot
Indicating a Positive Relationship

**Figure 3.12**
Scatterplot
Indicating a Negative Relationship

---

**PROBLEMS**

**Level A**

28. Explore the relationship between the selling prices and the appraised values of the 150 homes in the file P02_07.xls by computing a correlation.
   a. Is there evidence of a linear relationship between the selling price and appraised value for these data? If so, characterize the relationship (i.e., indicate whether the relationship is a positive or negative one).
   b. For which of the two remaining variables, the size of the home and the number of bedrooms in the home, is the relationship with the home’s appraised value stronger? Justify your choice.
29. A human resources manager at Beta Technologies, Inc., is trying to determine the variable that best explains the variation of employee salaries using the sample of 52 full-time employees in the file P02_01.xls. Create a table of correlations to help this manager identify whether the employee’s (a) gender, (b) age, (c) number of years of relevant work experience prior to employment at Beta, (d) the number of years of employment at Beta, or (e) the number of years of postsecondary education has the strongest linear relationship with annual salary.

30. Consider the enrollment data for Business Week’s top U.S. graduate business programs in the file P02_03.xls. Specifically, compute correlations to assess whether there is a linear relationship between the total number of full-time students and each of the following: (a) the proportion of female students, (b) the proportion of minority students, and (c) the proportion of international students enrolled at these distinguished business schools.

31. What is the relationship between the number of short-term general hospitals and the number of medical specialists in metropolitan areas? Explore this question by the correlation between these two variables using the data in the file P02_17.xls. Interpret your result.

32. Consider the economic development data in the file P02_06.xls.
   a. Calculate the correlation between each variable and the household’s average monthly expenditure on utilities.
   b. Which of the variables have a positive linear relationship with the household’s average monthly expenditure on utilities?
   c. Which of the variables have a negative linear relationship with the household’s average monthly expenditure on utilities?
   d. Which of the variables have essentially no linear relationship with the household’s average monthly expenditure on utilities?

33. Motorco produces electric motors for use in home appliances. One of the company’s production managers is interested in examining the relationship between the dollars spent per month in inspecting finished motor products and the number of motors produced during that month that were returned by dissatisfied customers. He has collected the data in the file P02_18.xls to explore this relationship for the past 36 months. Calculate the correlation between these two variables, and interpret it for this production manager.

34. A large number of metropolitan areas in the United States have been ranked with consideration of the following aspects of life in each area: cost of living, transportation, jobs, education, climate, crime, arts, health, and recreation. The data are in the file P02_55.xls.
   a. Create a table of correlations to discern the relationship between the metropolitan area’s overall score and each of these numerical factors.
   b. Which variables are most strongly associated with the overall score? Are you surprised by any of the results here?

35. The ACCRA Cost of Living Index provides a useful and reasonably accurate measure of cost-of-living differences among many urban areas. Items on which the index is based have been carefully chosen to reflect the different categories of consumer expenditures. The data are in the file P02_19.xls. Calculate a table of correlations to explore the relationship between the composite index and each of the various expenditure components.
   a. Which expenditure component has the strongest linear relationship with the composite index?
   b. Which expenditure component has the weakest linear relationship with the composite index?

36. Based on the annual data in the file P02_25.xls from the U.S. Department of Agriculture, determine whether a linear relationship exists between the number of farms and the average size of a farm in the United States during these years. Specifically, calculate an appropriate correlation and interpret it.

### 3.8 Describing Data Sets with Boxplots

We now introduce the boxplot, a very useful graphical method for summarizing data. We saved boxplots for this chapter because they are based on a number of summary measures we discussed in Section 3.2. Boxplots can be used in two ways: either to describe a single variable in a data set or to compare two (or more) variables. We illustrate these uses in the following examples.
Example 3.7 Distribution of Monthly Dow Returns

Recall that the Dow.xls file lists the monthly returns on the Dow from February 1947 through January 1993. Use a boxplot to summarize the distribution of these returns.

Objective To use StatTools’s Boxplot procedure to describe the distribution of monthly Dow returns.

Solution Excel has no boxplot option, but this option is included in the StatTools add-in. To create a boxplot for Dow returns, make sure a StatTools data set has been created, select the StatTools/Summary Graphs/Box-Whisker Plot menu item, click on the Format button and make sure the Unstacked option is selected, select the Return variable, and check the Include Key Describing Plot Elements option. (We discuss the difference between the Stacked and Unstacked options in later sections.) The resulting boxplot appears in Figure 3.13. Also, because you checked the Include Key option, the description of the boxplot elements in Figure 3.14 is placed below the boxplot. These elements are described next.

Figure 3.13 Boxplot of Dow Returns

Boxplot of Return

1Boxplots are also called box-whisker plots to indicate the whiskers (lines) on either side of the box.
3.8 Describing Data Sets with Boxplots

UNDERSTANDING A BOXPLOT

- The right and left of the box are at the third and first quartiles. Therefore, the length of the box equals the interquartile range (IQR), and the box itself represents the middle 50% of the observations. The height of the box has no significance. (Our boxplot conventions are not the only ones, as we see in the chapter opener about movie revenues. However, the variations are primarily cosmetic.)
- The vertical line inside the box indicates the location of the median. The point inside the box indicates the location of the mean.
- Horizontal lines are drawn from each side of the box. They extend to the most extreme observations that are no farther than 1.5 IQRs from the box. They are useful for indicating variability and skewness.²
- Observations farther than 1.5 IQRs from the box are shown as individual points. If they are between 1.5 IQRs and 3 IQRs from the box, they are called mild outliers and are hollow. Otherwise, they are called extreme outliers and are solid.³

The boxplot implies that the Dow returns are approximately symmetric on each side of the median, although the mean is slightly below the median. In addition, there are a few mild outliers but no extreme outliers.

Boxplots are probably most useful for comparing two populations graphically, as we illustrate in the following example.

²Some statistical software packages create vertical boxplots rather than the horizontal type StatTools creates. However, the information is identical.
³These conventions are due to the statistician John Tukey.
Example 3.8 Graphical Comparison of Male and Female Movie Stars’ Salaries

Recall that the salaries of famous actors and actresses are listed in the file Actors.xls. Use side-by-side boxplots to compare the salaries of male and female actors and actresses.

Objective To use StatTools’s Boxplot procedure to compare actors’ salaries across gender.

Solution

The data setup for this type of “comparison” problem can be in one of two forms: stacked or unstacked. The data are **stacked** if there is a categorical variable such as Gender that designates which gender each observation is in, and there is a numeric variable such as Salary that lists the salaries for both genders. (In this case, StatTools dialog boxes have a “Cat” column for the categorical variable and a “Val” column for the numeric variable.) The data are **unstacked** if there is a separate Salary column for each gender (one for males and one for females). As Figure 3.15 indicates, the data in the Actors.xls file are in stacked form. Therefore, to obtain side-by-side boxplots of male and female salaries, make sure a StatTools data set has been created, and select the StatTools/Summary Graphs/Box-Whisker Plot menu item. In the resulting dialog box, click on the Format button and make sure the Stacked option is checked. This gives you two columns of checkboxes. Select Gender as the “Cat” variable and Salary as the “Val” variable. (This time you can leave the Include Key option unchecked. It always provides exactly the same output and is included only as a learning tool.)

```
<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Famous actors and actresses</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Note: All monetary values are in $ millions.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Name</td>
<td>Gender</td>
<td>DomesticGross</td>
<td>ForeignGross</td>
</tr>
<tr>
<td>6</td>
<td>Angela Bassett</td>
<td>F</td>
<td>32</td>
<td>17</td>
</tr>
<tr>
<td>7</td>
<td>Jessica Lange</td>
<td>F</td>
<td>21</td>
<td>27</td>
</tr>
<tr>
<td>8</td>
<td>Winona Ryder</td>
<td>F</td>
<td>36</td>
<td>30</td>
</tr>
<tr>
<td>9</td>
<td>Michelle Pfeiffer</td>
<td>F</td>
<td>66</td>
<td>31</td>
</tr>
<tr>
<td>10</td>
<td>Whoopi Goldberg</td>
<td>F</td>
<td>32</td>
<td>33</td>
</tr>
<tr>
<td>11</td>
<td>Emma Thompson</td>
<td>F</td>
<td>26</td>
<td>44</td>
</tr>
<tr>
<td>12</td>
<td>Julia Roberts</td>
<td>F</td>
<td>57</td>
<td>47</td>
</tr>
<tr>
<td>13</td>
<td>Sharon Stone</td>
<td>F</td>
<td>32</td>
<td>47</td>
</tr>
<tr>
<td>14</td>
<td>Meryl Streep</td>
<td>F</td>
<td>34</td>
<td>47</td>
</tr>
<tr>
<td>15</td>
<td>Susan Sarandon</td>
<td>F</td>
<td>38</td>
<td>49</td>
</tr>
<tr>
<td>16</td>
<td>Nicole Kidman</td>
<td>F</td>
<td>55</td>
<td>51</td>
</tr>
<tr>
<td>17</td>
<td>Holly Hunter</td>
<td>F</td>
<td>51</td>
<td>53</td>
</tr>
<tr>
<td>18</td>
<td>Meg Ryan</td>
<td>F</td>
<td>43</td>
<td>55</td>
</tr>
<tr>
<td>19</td>
<td>Andie Macdowell</td>
<td>F</td>
<td>26</td>
<td>75</td>
</tr>
<tr>
<td>20</td>
<td>Jodie Foster</td>
<td>F</td>
<td>62</td>
<td>85</td>
</tr>
<tr>
<td>21</td>
<td>Rene Russo</td>
<td>F</td>
<td>69</td>
<td>85</td>
</tr>
<tr>
<td>22</td>
<td>Sandra Bullock</td>
<td>F</td>
<td>64</td>
<td>104</td>
</tr>
<tr>
<td>23</td>
<td>Demi Moore</td>
<td>F</td>
<td>65</td>
<td>125</td>
</tr>
<tr>
<td>24</td>
<td>Danny Glover</td>
<td>M</td>
<td>42</td>
<td>4</td>
</tr>
<tr>
<td>25</td>
<td>Billy Crystal</td>
<td>M</td>
<td>52</td>
<td>14</td>
</tr>
</tbody>
</table>
```

Excel Tip Many statistical software packages require data to be in stacked form. StatTools allows data to be in either stacked or unstacked form. In addition, StatTools can convert from stacked to unstacked form, or vice versa.
The resulting side-by-side boxplots appear in Figure 3.16. It is clear that the female salary box is considerably to the left of the male salary box, although both have about the same IQR. Each boxplot has three indications that the salary distributions are skewed to the right: (1) the means are larger than the medians, (2) the medians are closer to the left sides of the boxes than to the right sides, and (3) the horizontal lines (“whiskers”) extend farther to the right than to the left of the boxes. However, there are no outliers—not even the big stars like Harrison Ford or Sylvester Stallone!

**Figure 3.16**
Side-by-side Boxplots of Female and Male Salaries

---

### PROBLEMS

#### Level A

37. Consider the data in the file P03_37.xls on various performance measures for the largest U.S. airlines. Create boxplots to summarize each of the given measures of airline performance. Are these four performance measures linearly associated with one another? Explain your findings.

38. In a recent ranking of top graduate business schools in the United States published by *U.S. News & World Report*, the average starting base salaries for recent graduates from recognized graduate programs were provided. These data are given in the file P02_08.xls. Create a boxplot to characterize this distribution of average starting MBA salaries. Is this distribution essentially symmetric or skewed?

39. Consider the average annual rates for various forms of violent crime (including murder, rape, robbery, and aggravated assault) and the average annual rates for various forms of property-related crime (including burglary, larceny-theft, and motor vehicle theft) in selected U.S. metropolitan areas. The data are provided in the file P02_50.xls. Create side-by-side boxplots to compare the rates of these two general classes of crimes. Summarize your findings.

40. The annual average values of the Consumer Confidence Index are given in the file P02_28.xls. Create a boxplot to find the middle 50% of this distribution of values. Characterize the nature and amount of the variability about the interquartile range for these data.

41. Consider the given set of average annual household income levels of citizens from selected U.S. metropolitan areas in the file P03_06.xls. What can you infer about the shape of this particular income distribution from a boxplot of the given data? If applicable, note the presence of any outliers in this data set.

42. Using cost-of-living data from the *ACCRA Cost of Living Index* in the file P02_19.xls, create a boxplot to summarize the composite cost-of-living index values.
Level B

43. In 1970 a lottery was held to determine who would be drafted and sent to Vietnam. For each date of the year, a ball was put into an urn. For instance, January 1 was number 305 and February 14 was number 4. Thus a person born on February 14 would be drafted before a person born on January 1. The file P03_43.xls contains the “draft number” for each date for the 1970 and 1971 lotteries. Do you notice anything unusual about the results of either lottery? What do you think might have caused this result? (Hint: Create a boxplot for each month’s numbers.)

3.9 APPLYING THE TOOLS

Now that you are equipped with a collection of tools for describing data, it’s time to apply these tools to some serious data analysis. We examine three data sets in this section. Each of these is rather small by comparison with the data sets real companies often face, but they are large enough to make the analysis far from trivial. In each example we illustrate some of the output that might be obtained by the company involved, but you should realize that we are never really finished. With data sets as rich as these, there are always more numbers that could be calculated, more tables that could be formed, and more charts that could be created. We encourage you to take each analysis a few steps beyond what we present there.

Each example has a decision problem lurking behind it. If these data belonged to real companies, the companies would not only want to describe the data but they would want to use the information from their data analysis as a basis for decision making. We are not yet in a position to perform this decision making, but you should appreciate that the data analysis we perform here is really just the first step in an overall business analysis.

EXAMPLE 3.9 ACCOUNTS RECEIVABLE AT SPRING MILLS

The Spring Mills Company produces and distributes a wide variety of manufactured goods. Due to this variety, it has a large number of customers. The company classifies these customers as small, medium, and large, depending on the volume of business each does with Spring Mills. Recently, Spring Mills has noticed a problem with its accounts receivable. It is not getting paid back by its customers in as timely a manner as it would like. This obviously costs Spring Mills money. If a customer delays a payment of $300 for 20 days, say, then the company loses potential interest on this amount. The company has gathered data on 280 customer accounts. For each of these accounts, the data set lists three variables: Size, the size of the customer (coded 1 for small, 2 for medium, 3 for large); Days, the number of days since the customer was billed; and Amount, the amount the customer owes. (See the file Receive.xls.) What information can we obtain from these data?

Objective To use charts, summary measures, and pivot tables to understand data on accounts receivable at Spring Mills.

Solution

It is always a good idea to get a rough sense of the data first. We do this by calculating several summary measures for Days and Amount, a histogram of Amount, and a scatterplot of Amount versus Days in Figures 3.17, 3.18, and 3.19. Figure 3.17 indicates positive skewness in the Amount variable—the mean is considerably larger than the median, probably because of some large amounts due. Also, the standard deviation of Amount is quite large. This positive skewness is confirmed by the histogram. The scatterplot suggests some suspicious behavior, with two distinct groups of points. (You can check that the upper group
of points in the scatterplot corresponds to the large customers, whose amounts owed are uniformly greater than those in the other two groups.)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Summary measures for selected variables</td>
<td>Days</td>
<td>Amount</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Count</td>
<td>280,000</td>
<td>280,000</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Sum</td>
<td>4102,000</td>
<td>1300000,000</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Mean</td>
<td>14,650</td>
<td>464,286</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Median</td>
<td>13,000</td>
<td>320,000</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Standard deviation</td>
<td>7,221</td>
<td>378,055</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Minimum</td>
<td>2,000</td>
<td>140,000</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Maximum</td>
<td>39,000</td>
<td>2220,000</td>
<td></td>
</tr>
</tbody>
</table>

The next logical step is to see whether the different customer sizes have any effect on either Days, Amount, or the relationship between Days and Amount. To do this, it is useful to “unstack” the Days and Amount variables—that is, to create a new Days and Amount variable for each group of customer sizes. For example, the Days and Amount variables for customers of size 1 are labeled Days(1) and Amount(1). (Note that StatTools can work with the stacked variables directly. Just make sure that when you run any StatTools procedure, you click on the Format button and check the Stacked option.) Summary measures and a variety of charts based on these variables appear in Figures 3.20–3.28.
**Figure 3.19** Scatterplot of Amount versus Days for All Customers

![Scatterplot of Amount vs Days](image)

**Figure 3.20** Summary Measures Broken Down by Size

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td><strong>One Variable Summary</strong></td>
<td>Days (1)</td>
<td>Days (2)</td>
<td>Days (3)</td>
</tr>
<tr>
<td>9</td>
<td>Mean</td>
<td>9.800</td>
<td>20.550</td>
<td>19.233</td>
</tr>
<tr>
<td>10</td>
<td>Std. Dev.</td>
<td>3.128</td>
<td>6.622</td>
<td>6.191</td>
</tr>
<tr>
<td>11</td>
<td>Median</td>
<td>10.000</td>
<td>20.000</td>
<td>19.000</td>
</tr>
<tr>
<td>12</td>
<td>Minimum</td>
<td>2.000</td>
<td>8.000</td>
<td>3.000</td>
</tr>
<tr>
<td>13</td>
<td>Maximum</td>
<td>17.000</td>
<td>39.000</td>
<td>32.000</td>
</tr>
<tr>
<td>14</td>
<td>Count</td>
<td>150</td>
<td>100</td>
<td>30</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td><strong>One Variable Summary</strong></td>
<td>Amount (1)</td>
<td>Amount (2)</td>
<td>Amount (3)</td>
</tr>
<tr>
<td>18</td>
<td>Mean</td>
<td>$254.53</td>
<td>$481.90</td>
<td>$1454.33</td>
</tr>
<tr>
<td>19</td>
<td>Std. Dev.</td>
<td>$49.28</td>
<td>$99.15</td>
<td>$293.89</td>
</tr>
<tr>
<td>20</td>
<td>Median</td>
<td>$250.00</td>
<td>$470.00</td>
<td>$1390.00</td>
</tr>
<tr>
<td>21</td>
<td>Minimum</td>
<td>$140.00</td>
<td>$280.00</td>
<td>$930.00</td>
</tr>
<tr>
<td>22</td>
<td>Maximum</td>
<td>$410.00</td>
<td>$750.00</td>
<td>$2220.00</td>
</tr>
<tr>
<td>23</td>
<td>Count</td>
<td>150</td>
<td>100</td>
<td>30</td>
</tr>
</tbody>
</table>
Figure 3.21  Histogram of Amount for Small Customers

Figure 3.22  Histogram of Amount for Medium Customers
Figure 3.23 Histogram of Amount for Large Customers

Figure 3.24 Boxplots of Days Owed by Different Size Customers
Figure 3.25  Boxplots of Amounts Owed by Different Size Customers

Boxplot of Comparison of Amount

Size = 3

Size = 2

Size = 1

Figure 3.26  Scatterplot of Amount versus Days for Small Customers

Scatterplot of Amount(1) vs Days(1)
Figure 3.27  Scatterplot of Amount versus Days for Medium Customers

Figure 3.28  Scatterplot of Amount versus Days for Large Customers
There is obviously a lot going on here, and most of it is clear from the charts. We point out the following: (1) there are far fewer large customers than small or medium customers; (2) the large customers tend to owe considerably more than small or medium customers; (3) the small customers do not tend to be as long overdue as the medium or large customers; and (4) there is no relationship between Days and Amount for the small customers, but there is a definite positive relationship between these variables for the medium and large customers.

We have now done the “obvious” analysis. There is still much more we can do, however. For example, suppose Spring Mills wants a breakdown of customers who owe at least $500. We first create a new variable called “Large?” next to the original variables that equals 1 for all amounts greater than $500 and equals 0 otherwise.\(^4\) (See Figure 3.29 for some of the data.) We do this by entering the formula

\[
\text{=IF(C6>}$B$3,1,0)
\]

in cell D6 and copying down. We then use a pivot table to create a count of the number of 1’s in this new variable for each value of the Size variable.

Figure 3.29 shows the results. Actually, we created this pivot table twice, once (on top) showing counts as percentages of each column, and once showing them as percentages of each row. The top table shows, for example, that about 73% of all customers with amounts less than $500 are small customers. The bottom table shows, for example, that 45% of all medium-size customers owe at least $500. When you hear the expression “slicing and dicing the data,” this is what it means. These two pivot tables are based on the same counts, but they portray them in slightly different ways. Neither is better than the other; each provides useful information.

Finally, we investigate the amount of interest Spring Mills is losing by the delays in its customers’ payments. We assume that the company can make 12% annual interest on excess cash. Then we create a Lost variable for each customer size that indicates the amount of interest Spring Mills loses on each customer group. (See Figure 3.31.) The typical formula for lost interest in cell C10 is

\[
\text{=B10*A10*$C$7/365}
\]

\(^4\)We could just as well enter the labels “yes” and “no” in the Large? column. However, it is common to use 0 to 1 values for such a variable.
**Figure 3.30**

Pivot Tables for Counts of Customers Who Owe More Than $500

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Count of Large?</td>
<td>Large?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Size</td>
<td></td>
<td>1</td>
<td>Grand Total</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>73.17%</td>
<td>0.00%</td>
<td>53.67%</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>26.83%</td>
<td>50.00%</td>
<td>35.71%</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>0.00%</td>
<td>40.00%</td>
<td>10.71%</td>
</tr>
<tr>
<td>7</td>
<td>Grand Total</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

**Figure 3.31**

Summary Measures of Lost Interest

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Interest Lost</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Summary measures for selected variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Lost1</td>
<td>Lost2</td>
<td>Lost3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Sum</td>
<td>$122.68</td>
<td>$338.65</td>
<td>$287.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Annual interest rate</td>
<td>12%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Days1</td>
<td>Amount1</td>
<td>Lost1</td>
<td>Days2</td>
<td>Amount2</td>
<td>Lost2</td>
<td>Days3</td>
<td>Amount3</td>
<td>Lost3</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>$180.00</td>
<td>$0.41</td>
<td>17</td>
<td>$470.00</td>
<td>$2.63</td>
<td>19</td>
<td>$1,330.00</td>
<td>$8.31</td>
</tr>
<tr>
<td>11</td>
<td>8</td>
<td>$210.00</td>
<td>$0.55</td>
<td>22</td>
<td>$540.00</td>
<td>$3.91</td>
<td>20</td>
<td>$1,400.00</td>
<td>$9.21</td>
</tr>
<tr>
<td>12</td>
<td>10</td>
<td>$210.00</td>
<td>$0.69</td>
<td>28</td>
<td>$560.00</td>
<td>$5.16</td>
<td>14</td>
<td>$1,550.00</td>
<td>$7.13</td>
</tr>
<tr>
<td>13</td>
<td>8</td>
<td>$150.00</td>
<td>$0.39</td>
<td>24</td>
<td>$470.00</td>
<td>$3.71</td>
<td>15</td>
<td>$1,460.00</td>
<td>$7.20</td>
</tr>
<tr>
<td>14</td>
<td>9</td>
<td>$300.00</td>
<td>$0.89</td>
<td>26</td>
<td>$650.00</td>
<td>$5.56</td>
<td>23</td>
<td>$2,030.00</td>
<td>$15.35</td>
</tr>
<tr>
<td>15</td>
<td>5</td>
<td>$240.00</td>
<td>$0.39</td>
<td>29</td>
<td>$530.00</td>
<td>$5.05</td>
<td>19</td>
<td>$1,520.00</td>
<td>$9.49</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>$330.00</td>
<td>$0.43</td>
<td>21</td>
<td>$550.00</td>
<td>$3.80</td>
<td>15</td>
<td>$1,330.00</td>
<td>$6.56</td>
</tr>
<tr>
<td>17</td>
<td>10</td>
<td>$290.00</td>
<td>$0.95</td>
<td>33</td>
<td>$620.00</td>
<td>$6.73</td>
<td>17</td>
<td>$1,520.00</td>
<td>$8.50</td>
</tr>
<tr>
<td>18</td>
<td>5</td>
<td>$240.00</td>
<td>$0.39</td>
<td>16</td>
<td>$430.00</td>
<td>$2.26</td>
<td>21</td>
<td>$1,390.00</td>
<td>$9.60</td>
</tr>
<tr>
<td>19</td>
<td>13</td>
<td>$270.00</td>
<td>$1.15</td>
<td>27</td>
<td>$460.00</td>
<td>$4.08</td>
<td>24</td>
<td>$1,590.00</td>
<td>$12.55</td>
</tr>
</tbody>
</table>

This is the amount owed multiplied by the number of days owed multiplied by the interest rate, divided by the number of days in a year. Then we calculate sums of these amounts in row 5. Although Spring Mills is losing more per customer from the large customers, it is losing more in total from the medium-size customers—because there are more of them. This is shown graphically in Figure 3.32 with a pie chart of the sums in row 5. This pie chart shows, for example, that 46% of the lost interest is due to the medium-size customers.

**Figure 3.32**

Pie Chart of Lost Interest by Customer Size
If Spring Mills really wants to decrease its receivables, it might want to target the medium-size customer group, from which it is losing the most interest. Or it could target the large customers because they owe the most on average. The most appropriate action depends on the cost and effectiveness of targeting any particular customer group. However, the analysis presented here gives the company a much better picture of the current state of affairs.

### Example 3.10 Customer Arrival and Waiting Patterns at R&P Supermarket

The R&P Supermarket is open 24 hours a day, 7 days a week. Lately, it has been receiving a lot of complaints from its customers about excessive waiting in line for checking out. R&P has decided to investigate this situation by gathering data on arrivals, departures, and line lengths at the checkout stations. It has collected data in half-hour increments for an entire week—336 observations—starting at 8 A.M. on Monday morning and ending at 8 A.M. the following Monday.

Specifically, it has collected data on the following variables: InitialWaiting, the number waiting or being checked out at the beginning of a half-hour period; Arrivals, the number of arrivals to the checkout stations during a period; Departures, the number finishing the checkout process during a period; and Checkers, the number of checkout stations open during a period. (See the file Checkout.xls.)

The data set also includes time variables: Day, day of week; StartTime, clock time at the beginning of each half-hour period; and TimeInterval, a descriptive term for the time of day, such as Lunch rush for 11:30 A.M. to 1:30 P.M. (The comment in cell C3 of the data sheet spells these out.) Finally, the data set includes the calculated variable EndWaiting, the number waiting or being checked out at the end of a half-hour period. For any time period, it equals InitialWaiting plus Arrivals minus Departures; it also equals InitialWaiting for the next period. A partial listing of the data appears in Figure 3.33.

![Figure 3.33](image)

A Partial Listing of the Supermarket Checkout Data

The manager of R&P wants to analyze these data to discover any trends, particularly in the pattern of arrivals throughout a day or across the entire week. Also, the store currently uses a “by the seat of your pants” approach to opening and closing checkout stations each half hour. The manager would like to see how well the current approach is working. Of course, she would love to know the “best” strategy for opening and closing checkout stations—but this is beyond her (and our) capabilities at this point.
Objective To use charts, summary measures, and pivot tables to understand time patterns of arrivals and congestion at R&P Supermarket.

Solution

Obviously, time plays a crucial role in this example, so a good place to start is to create one or more time series graphs. The graph in Figure 3.34 shows the time series behavior of InitialWaiting (the lower line) and Arrivals during the entire week. (This looks much better on a PC monitor, where the two lines are in different colors.) There is almost too much clutter in this graph to see exactly what’s happening, but it is clear that (1) Fridays and Saturdays are the busiest days; (2) the time pattern of arrivals is somewhat different—more spread out—during the weekends than during the weekdays; (3) there are fairly regular peak arrival periods during the weekdays; and (4) the number waiting is sometimes as large as 10 or 20, and the largest of these tend to be around the peak arrival times. You can decide whether it might be better (with less clutter) to separate this plot into two time series graphs, one for InitialWaiting and one for Arrivals. The advantage of the current plot is that we can match up time periods for the two variables.

Figure 3.34 Time Series Graph of InitialWaiting and Arrivals Variables

A similar time series graph appears in Figure 3.35. This shows Arrivals and Departures, although it is difficult to separate the two time series—they are practically on top of one another. Perhaps this is not so bad. It means that for the most part, the store is checking out customers approximately as quickly as they are arriving.
A somewhat more efficient way to obtain this time series behavior is with pivot tables. Figure 3.36 shows one possibility. To create this pivot table, we drag the InitialWaiting variable to the Data area, express it as an average, drag the StartTime variable to the Row area, and drag the Day variable to the Page area. (We also condensed the information from half-hour periods to hour-long periods by using the grouping option on the Pivot Table toolbar.) Finally, we create a time series graph from the data in the pivot table. Note how the variable in the Page area works. For example, we obtain the graph in Figure 3.36 if we choose Monday in the Page area (the drop-down list in cell B1). However, if we choose another day in cell B1, the data in the pivot table and the graph change automatically. So we can make comparisons across the days of the week with a couple of clicks of the mouse!

Similarly, the pivot table and corresponding column chart in Figure 3.37 indicate the average number of arrivals per half-hour period for each interval in the day. To obtain this output, we drag the Arrival variable to the Data area, express it as an average, drag the TimeInterval variable to the Row area, and drag the Day variable to the Page area. You can check that the pattern shown here for Friday is a bit different than for the other days—it has a significant bulge during the afternoon rush period.

**Excel Tip** If you try to create this pivot table on your own, you will no doubt wonder how we got the time intervals in column A in the correct chronological order. The trick is to create a custom sort list. To do so, select the Tools/Options menu item, and click on the Custom Lists tab. Then type a list of items in the List Entries box in the order you want them, and click on the Add button. To sort in this list order in the pivot table, place the cursor on any item in column A, select the Data/Sort menu item, click on Options, and select your new list from the drop-down list. This custom list will then be available in this or any other workbook you develop.
The manager of R&P is ultimately interested in whether the “right” number of checkout stations are available throughout the day. Figures 3.38 and 3.39 provide some evidence. The first of these is a scatterplot of Checkers versus TotalCustomers. (We calculated the TotalCustomers variable as the sum of the InitialWaiting and the Arrivals variables to measure the total amount of work presented to the checkout stations in any half-hour period.) There is an obvious positive relationship between these two variables. Evidently, management is reacting as it should—it is opening more checkout stations when there is more traffic. The second scatterplot shows EndWaiting versus Checkers. There is again a definite upward trend. Periods when more checkout stations are open tend to be associated with periods where more customers still remain in the checkout process. Presumably, management is reacting with more open checkout stations in busy periods, but it is not reacting strongly enough.
Figure 3.38  Scatterplot of Checkers versus TotalCustomers

Figure 3.39  Scatterplot of EndWaiting versus Checkers
If you think you can solve the manager’s problem just by fiddling with the numbers in the Checkers column, think again. There are two problems. First, there is a trade-off between the “cost” of having customers wait in line and the cost of paying extra checkout people. This is a difficult trade-off for any supermarket manager. Second, the number of departures is clearly related to the number of checkout stations open. (The relationship is a complex one that requires mathematical queuing theory to quantify.) Therefore, it doesn’t make sense to change the numbers in the Checkers column without changing the numbers in the Departures (and hence the InitialWaiting and EndWaiting) columns in an appropriate way. This is not an easy problem!

In real situations like this one, company analysts often use a technique called discrete-event simulation to study the effect on waiting times and costs of using various policies for opening and closing checkout counters. We do not cover this technique in this book.

**Example 3.11 Demographic and Catalog Mailing Effects on Sales at HyTex**

The HyTex Company is a direct marketer of stereophonic equipment, personal computers, and other electronic products. HyTex advertises entirely by mailing catalogs to its customers, and all of its orders are taken over the telephone. The company spends a great deal of money on its catalog mailings, and it wants to be sure that this is paying off in sales. Therefore, HyTex has collected data on 1000 customers at the end of the current year. (See the file `Catalogs.xls`.) For each customer the company has data on the following variables:

- **Age:** coded as 1 for 30 years or younger, 2 for 31 to 55 years, and 3 for 56 years or older
- **Gender:** coded as 1 for males, 0 for females
- **OwnHome:** coded as 1 if customer owns a home, 0 otherwise
- **Married:** coded as 1 if customer is currently married, 0 otherwise
- **Close:** coded as 1 if customer lives reasonably close to a shopping area that sells similar merchandise, 0 otherwise
- **Salary:** combined annual salary of customer and spouse (if any)
- **Children:** number of children living with customer
- **History:** coded as “NA” if customer had no dealings with the company before this year, 1 if customer was a low-spending customer last year, 2 if medium-spending, 3 if high-spending
- **Catalogs:** number of catalogs sent to the customer this year
- **AmountSpent:** total amount of purchases made by the customer this year

HyTex wants to analyze these data carefully to understand its customers better. Also, the company wants to see whether it is sending the catalogs to the right customers. Currently, each customer receives either 6, 12, 18, or 24 catalogs through the mail each year. However, the decision as to who receives how many has not really been thought out carefully. Is the current distribution of catalogs effective? Is there room for improvement?

**Objective** To use charts, summary measures, and pivot tables to understand the demographics of HyTex’s customers, and to understand how these demographics, as well as the number of catalogs mailed, affect amounts spent.
Solution

This is the most difficult example we have faced so far, but it pales in comparison to the difficulty real direct marketing companies face. They have all sorts of data on millions of customers. How can they make sense of all these data? Using our relatively small data set, we will get the ball rolling. We challenge you to discover additional patterns in the data. Furthermore, we see only an indication of whether the current distribution of catalog mailings is effective. It is well beyond our abilities at this point to find a more effective catalog distribution policy.

HyTex is primarily interested in the AmountSpent variable, so it makes sense to create scatterplots of AmountSpent versus selected “explanatory” variables. We do this in Figures 3.40 to 3.42. Figure 3.40 shows AmountSpent versus Salary. It is clear that customers with higher salaries tend to spend more, although the variability in amounts spent increases significantly as salary increases. Figure 3.41 shows that there is some tendency toward higher spending among customers who receive more catalogs. But do the catalogs cause more spending, or are more catalogs sent to customers who would tend to spend more anyway? There is no way to answer this “cause-and-effect” question with the data the company has collected. Figure 3.42 shows the interesting tendency of customers with more children to spend less. Perhaps customers with more children are already spending so much on $100-plus athletic shoes that they have little left to spend on electronic equipment!
Pivot tables and accompanying charts are very useful in this type of situation. We show several. First, Figures 3.43 and 3.44 can be used to better understand the demographics of the customers. Each row of Figure 3.43 shows the percentages of an age group who own homes. By changing the page variables Gender and Married in cells B5 and B6, we can see how these percentages change for married women, unmarried men, and so on. You can check that these percentages remain relatively stable for the various groups. Specifically, a small percentage of the younger people own their own home, regardless of marital status or gender.

Figure 3.43
Percent Home Owners versus Age, Married, and Gender

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Married</td>
<td>(All)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>OwnHome</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Age</td>
<td></td>
<td>Count</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0.2124%</td>
<td>0.52%</td>
<td>100.00%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>0.3183%</td>
<td>0.52%</td>
<td>100.00%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>0.2124%</td>
<td>0.52%</td>
<td>100.00%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Grand Total</td>
<td>0.52%</td>
<td>0.52%</td>
<td>100.00%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.44 is similar. It shows the percentages of each age group who are married, for any combination of the Gender and OwnHome variables. Here the percentages change considerably for different settings of the page variables. For example, you can check that the married/unmarried split is quite different for women who don’t own a home than for the male home owners shown in the figure.

Figure 3.44
Percent Married versus Age, OwnHome, and Gender

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>OwnHome</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>Gender</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>Count</td>
<td>Married</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Age</td>
<td></td>
<td>0.52</td>
<td>0.52%</td>
<td>100.00%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0.52%</td>
<td>0.52%</td>
<td>100.00%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>0.52%</td>
<td>0.52%</td>
<td>100.00%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>0.52%</td>
<td>0.52%</td>
<td>100.00%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Grand Total</td>
<td>0.52%</td>
<td>0.52%</td>
<td>100.00%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Excel Tip  We do not have to create a new pivot table and pivot chart to get those shown in Figure 3.44. Starting with the pivot table shown in Figure 3.43, we can simply drag the gray OwnHome button to just above the gray Married button, and then drag the Married button to where the OwnHome button was. Everything updates automatically, including the chart! (This assumes that we requested a pivot chart in the first place.) The pivot table/pivot chart combinations for the rest of this example were created from the original pivot table/pivot chart combination in Figure 3.43 in the same way—that is, by dragging field buttons to the appropriate locations in the pivot table. What a time-saver!
Figure 3.45 provides more demographic information. Now we show the average Salary broken down by Age and Gender, with page variables for OwnHome and Married. You can check that the shape of the resulting chart is practically the same for any combination of the page variables. However, the heights of the bars change appreciably. For example, the average salaries are considerably larger for the married home owners shown in the figure than for unmarried customers who are not home owners.

The pivot table and pivot chart in Figure 3.46 break the data down in another way. Each column in the pivot table shows the percentages in the various History categories for a particular number of children. Each of these columns corresponds to one of the bars in the “stacked” bar chart. Also, we have used Close as a page variable. Two interesting points emerge. First, customers with more children tend to be more heavily represented in the low-spending History category (and less heavily represented in the high-spending category). Also, as you can check by changing the setting of the Close variable from 1 to 0, the percentage of high-spenders among customers who live far from electronics stores is much higher than for those who live close to such stores.

Although we are not told exactly how HyTex determined its catalog mailing distribution, Figure 3.47 provides an indication. Each row of the pivot table shows the percentages of a particular History category that were sent 6, 12, 18, or 24 catalogs. The company’s distribution policy is still somewhat unclear—and there is probably hope for improvement—but it did evidently send more catalogs to high-spending customers and fewer to low-spending customers.
Finally, Figure 3.48 shows the average AmountSpent versus History and Catalogs, with a variety of demographic variables in the page area. There are so many possible combinations that it is difficult to discover all the existing patterns. However, one thing stands out loud and clear from the graph: the more catalogs customers receive, the more they tend to spend. In addition, if they were large spenders last year, they tend to be large spenders this year.

In a pivot table with this many combinations, there will almost certainly be some combinations with no observations. For example, it turns out that there are no young married males who were low-spenders last year and received 18 catalogs. In this case you see a blank in the corresponding pivot table cell. In addition, there are no young married male home owners who received 12 catalogs. If you try this combination, the whole “12” column of the pivot table will disappear—and the corresponding bars will disappear from the chart.
3.10 CONCLUSION

This chapter and the previous chapter have illustrated the tremendous variety of descriptive measures you can obtain with Excel's built-in tools and add-ins such as StatTools. The concepts in these chapters are relatively simple. The key, therefore, is to have simple-to-use tools available to produce tables, graphs, and numerical summary measures in a matter of minutes (or even seconds). This is now possible, not only with statistical software packages but with spreadsheet packages, particularly Excel. These tools allow you to concentrate on presenting the data in the most appropriate way, so that interesting information hidden in the data is brought to the surface.

Summary of Key Terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Explanation</th>
<th>Excel</th>
<th>Pages</th>
<th>Equation Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-variable summary statistics</td>
<td>General term for measures that summarize distribution of a variable</td>
<td>StatTools/ Summary Statistics/ One-Variable Summary</td>
<td>82–93</td>
<td></td>
</tr>
<tr>
<td>Mean X (for sample) ( \mu ) (for population)</td>
<td>Average of observations</td>
<td>=AVERAGE(range)</td>
<td>82</td>
<td>3.1</td>
</tr>
<tr>
<td>Median</td>
<td>Middle observation after observations are sorted from low to high</td>
<td>=MEDIAN(range)</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>Mode</td>
<td>Most frequently occurring observation</td>
<td>=MODE(range)</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>Percentile</td>
<td>Value such that a specified percentage of observations are below it</td>
<td>=PERCENTILE(range,pct), where pct is decimal between 0 and 1</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>Quartile</td>
<td>First quartile has 25% of observations below it. Third quartile has 25% of observations above it. Second quartile is the median.</td>
<td>=QUARTILE(range,n), where n is 1, 2, or 3</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>Interquartile range</td>
<td>Difference between third and first quartiles</td>
<td>=QUARTILE(range,3) − QUARTILE(range,1)</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>Minimum</td>
<td>Smallest observation</td>
<td>=MIN(range)</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>Maximum</td>
<td>Largest observation</td>
<td>=MAX(range)</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>Difference between largest and smallest observations</td>
<td>=MAX(range) − MIN(range)</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>Variance ( s^2 ) (for sample) ( \sigma^2 ) (for population)</td>
<td>Measure of variability Basically, average of squared deviations from mean</td>
<td>=VAR(range) (for sample) =VARP(range) (for population)</td>
<td>87</td>
<td>3.2, 3.3</td>
</tr>
<tr>
<td>Standard deviation ( s ) (for sample) ( \sigma ) (for population)</td>
<td>Measure of variability in same units as observations Square root of variance</td>
<td>=STDEV(range) (for sample) =STDEVVP(range) (for population)</td>
<td>87</td>
<td>3.4</td>
</tr>
</tbody>
</table>
3.10 Conclusion

### Empirical rules
They specify approximate percentages of observations within 1, 2, or 3 standard deviations of mean for bell-shaped distributions.

### Measures of association
- **Covariance**: Measure of association, affected by units of measurement
  - $=\text{COVAR}(\text{range1}, \text{range2})$
  - Pages: 96
  - Equation Number: 3.5

- **Correlation**
  - **$r$** (for sample): Measure of association, unaffected by units of measurement, always between $-1$ and $+1$
  - $=\text{CORREL}(\text{range1}, \text{range2})$
  - Pages: 96
  - Equation Number: 3.6

- **Boxplot**: Chart that indicates the distribution of one or more variables; box captures middle $50\%$ of data, lines and points indicate possible skewness and outliers
  - StatTools/Summary Graphs/Box-Whisker Plot
  - Pages: 99

### Conceptual Exercises

**C.1.** Suppose that the histogram of a given income distribution is positively skewed. What does this fact imply about the relationship between the mean and median of this distribution?

**C.2.** “The midpoint of the line segment joining the first quartile and third quartile of any distribution is the median.” Is this statement true or false? Explain your answer.

**C.3.** Explain why the standard deviation would likely not be a reliable measure of variability for a distribution of data that includes at least one extreme outlier.

**C.4.** “Two numerical variables are unrelated if the correlation between them is close to 0.” Indicate whether this statement is true or false. Explain your choice.

**C.5.** Explain how a boxplot could be used to determine whether the associated distribution of values is essentially symmetric.

### Level A

**44.** The annual base salaries for 200 students graduating from a reputable MBA program this year (see the file P02_74.xls) are of interest to those in the admissions office who are responsible for marketing the program to prospective students. What salary level is most indicative of those earned by students graduating from this MBA program this year?

**45.** In its annual ranking of top graduate business schools in the United States, *U.S. News & World Report* provides data on a number of attributes of recognized graduate programs (refer to the file P02_08.xls). One variable of interest is the annual out-of-state tuition paid by affected full-time students in each program.

  **a.** Find the annual out-of-state tuition levels at each of the 25th, 50th, and 75th percentiles.

  **b.** Identify the schools with the largest and smallest annual out-of-state tuitions. Does there appear to be a relationship between the program’s overall ranking and its out-of-state tuition level?

**46.** Consider the Consumer Price Index, which provides the annual percentage change in consumer prices, for the period in the file P02_26.xls. Find and interpret the interquartile range of these annual percentage changes.

**47.** The Consumer Confidence Index (CCI) attempts to measure people’s feelings about general business conditions, employment opportunities, and their own income prospects. The annual average values of the CCI are given in the file P02_28.xls.

  **a.** Fifteen percent of all years in this sample have annual CCI values that exceed what value?
b. Forty percent of all years in this sample have annual CCI values that are less than or equal to what value?

c. In which year of the given sample period were U.S. consumers most confident, as measured by the CCI?

d. In which year of the given sample period were U.S. consumers least confident, as measured by the CCI?

48. Consider the proportion of Americans under the age of 49. The annual averages of the discount rate, federal funds rate, and the prime rate are given in the file P02_29.xls.
   a. In which years of the sample has the poverty rate for American children exceeded the rate that defines the third quartile of these data?
   b. In which years of the sample has the poverty rate for American children fallen below the rate that defines the first quartile of these data?
   c. What is the typical poverty rate for American children during this period?

49. The annual averages of the discount rate, federal funds rate, and the prime rate are given in the file P02_30.xls. For each of these three key interest rates, determine the following:
   a. Thirty percent of all years in the sample period have annual average rates that exceed what value?
   b. Twenty-five percent of all years in the sample period have annual average rates that are less than or equal to what value?
   c. What is the most typical annual average rate over the given sample period?

50. Given data in the file P02_13.xls from a recent survey of chief executive officers from 350 of the largest U.S. public companies, respond to the following questions:
   a. Find the annual salary below which 75% of all given CEO salaries fall.
   b. Find the annual bonus above which 55% of all given CEO bonuses fall.
   c. Determine the range of the middle 50% of all given total annual compensation figures (i.e., of the amounts found in column Sum03).

51. The file P02_44.xls contains the number of classroom teachers and the average salary of classroom teachers for each of the 50 states and the District of Columbia. Which of the states paid their teachers average salaries that exceeded approximately 90% of all average salaries? Which of the states paid their teachers average salaries that exceeded only about 10% of all average salaries?

52. The annual base salaries for 200 students graduating from a reputable MBA program this year are given in the file P02_74.xls.
   a. Do the empirical rules apply to these data? Explain.
   b. If the empirical rules apply here, between what two numbers can we be about 68% sure that the salary of any one of these 200 students will fall?

53. Refer to the data given in the file P02_11.xls. Consider the average time (in minutes) it takes a citizen of each metropolitan area to travel to work and back home each day.
   a. Find a reliable measure of the dispersion of these average commute times around the overall sample mean.
   b. Between what two numbers can we be approximately 99.7% sure that any one of these average travel times will fall?

54. Is there a strong relationship between a chief executive officer’s annual compensation and her or his organization’s recent profitability? Explore this question by generating correlations for the survey data in the file P02_13.xls. In particular, compute and interpret correlation measures for the change in the company’s net income from 2002 to 2003 (see Comp_NetInc03P column) and the CEO’s 2003 base salary, as well as for the change in the company’s net income from 2002 to 2003 and the CEO’s 2003 bonus. Summarize your findings.

55. Create boxplots to compare recent job growth rates with forecasted job growth rates for selected towns in the United States. These growth rates are all provided in the file P03_55.xls. Do you detect the presence of any outliers in either of these two distributions? Also, calculate the correlation between these two sets of job growth rates and interpret it.

56. The percentage of private-industry jobs that are managerial has steadily declined in recent years as companies have found middle management a ripe area for cutting costs. How have women and various minority groups fared in gaining management positions during this period of corporate downsizing of the management ranks? Relevant data are given in the file P02_75.xls. Create boxplots using these data to make general comparisons across the various groups included in the set.

57. The U.S. Bureau of Labor Statistics provides data on the year-to-year percentage changes in the wages and salaries of workers in private industries, including both white-collar and blue-collar occupations. Consider the data in the file P02_56.xls. Create side-by-side boxplots to summarize these distributions of annual percentage changes. Interpret the interquartile range for each of three distributions.

58. Explore the given distribution of the numbers of beds in short-term general hospitals in selected U.S. metropolitan areas by creating a boxplot (refer to the data in the file P02_17.xls). Use your boxplot to answer the following questions:
   a. Is it more likely for these metropolitan areas to have larger or smaller numbers of hospital beds?
   b. Is the mean or median the more accurate measure of central location in this case? Explain.
60. Consider various characteristics of the U.S. civilian labor force provided in the file P03_60.xls. In particular, examine the given unemployment rates across the United States to answer the following questions.

a. Characterize the distribution of total unemployment rates. What is the most typical value? How are the other total unemployment rates distributed about the typical rate?

b. Compare the distribution of total unemployment rate to those of the male unemployment rate and the female unemployment rate. How are these distributions similar? How are they different?

c. Which is more strongly associated with the total unemployment rate in the United States: the male unemployment rate or the female unemployment rate?

61. The file P03_61.xls contains the sale price of gasoline in each of the 50 states.

a. Compare the distributions of gasoline sale price data (one for each year). Specifically, do you find the mean and standard deviation of these distributions to be changing over time? If so, how do you explain the trends?

b. In which regions of the country have gasoline prices changed the most?

c. In which regions of the country have gasoline prices remained relatively stable?

62. Examine life expectations (in years) at birth for various countries across the world. These data are provided in the file P03_62.xls.

a. Generate an estimate of the typical human’s life span at birth using the 2002 data. What are the limitations of the method you have employed in estimating this world population parameter?

b. Characterize the variability of the life spans at birth using the 2002 data. Is this distribution fairly symmetric or skewed? How do you know?

c. How strongly are the 2002 life expectations associated with projections for births in 2025 and 2050? Explain why the degree of linear association between the 2002 data and each set of projections diminishes somewhat over time.

63. This problem focuses on the per capita circulation of daily newspapers in the United States. The file P03_63.xls contains these data.

a. Compare the yearly distributions of daily newspaper per capita circulation over the period.

b. Note any clear trends, both nationally and regionally, in the average value of and variability of per capita newspaper circulation during the given years.

64. Have the proportions of Americans receiving public aid changed in recent years? Explore this question through a careful examination of the data provided in the file P03_64.xls. In particular, generate numerical summary measures to respond to each of the following.

a. Report any observed changes in the overall mean or median rates during the given time period.

b. Can you find evidence of regional changes in the proportions of Americans receiving public aid? If so, summarize your specific findings.

65. The file P03_65.xls contains the measured weight (in ounces) of a particular brand of ready-to-eat breakfast cereal placed in each of 500 randomly selected boxes by one of five different filling machine operators. Quality assurance personnel at this company are interested in determining how well these five operators are performing their assigned task of consistently placing 15 ounces of cereal in each box.

a. Employ descriptive graphs and summary measures to ascertain whether some or all of these operators are consistently missing the target weight of 15 ounces per box.

b. If you were charged with selecting the “Outstanding Employee of the Month” from this set of filling machine operators, which operator would you select based on the given data? Defend your choice.

66. Electro produces voltage-regulating equipment in New York and ships the equipment to Chicago. The voltage held is measured in New York before each unit is shipped to Chicago. The voltage held by each unit is also measured when the unit arrives in Chicago. The file P03_66.xls contains a sample of voltage measurements at each city. A voltage regulator is considered acceptable if it can hold a voltage of between 25 and 75 volts.

a. Using boxplots and descriptive statistics, what can you learn about the voltage held by units before shipment and after shipment?

b. What percentage of units are acceptable before and after shipping?

c. Do you have any suggestions about how to improve the quality of Electro’s regulators?

d. Ten percent of all New York regulators have a voltage exceeding what value?

e. Five percent of all New York regulators have a voltage less than or equal to what value?

67. The file P03_67.xls contains the individual scores of students in two different accounting sections who took the same exam. Comment on the differences between exam scores in the two sections.
68. The file P03_68.xls contains the monthly interest rates on 3-month government T-bills. For example, in January 1985, 3-month T-bills yielded 7.76% annual interest. To succeed in investments, it is important to understand the characteristics of the monthly changes in T-bill rates.
   a. Create a histogram of the monthly changes in interest rates. Choose categories so that you get an “interesting” histogram.
   b. Do the empirical rules hold for changes in monthly interest rates?
   c. Based on the given data, there is a 5% chance that during a given month T-bill rates will increase by less than what value? (A negative number is allowed here.)
   d. Based on the given data, there is a 10% chance that during a given month T-bill rates will increase by at least what value?
   e. Based on the given data, estimate the chances that T-bill rates during a given month will increase by more than 0.5%.

Level B

69. Data on the numbers of insured commercial banks in the United States are given in the file P03_69.xls.
   a. Compare these distributions of the numbers of U.S. commercial banks (one for each year). Are the mean and standard deviation of these numbers changing over time? If so, how do you explain the trends?
   b. What trends do you notice in the numbers of commercial banks by region? For example, how do the numbers of commercial banks appear to be changing in the northeastern United States over the given period? Summarize your findings for each region of the country.

70. Educational attainment in the United States is the focus of this problem. Employ descriptive methods with the data provided in the file P03_70.xls to characterize the educational achievements of Americans in the given year. Do your findings surprise you in some way?

71. Data on U.S. homeownership rates are given in the file P03_71.xls.
   a. Employ numerical summary measures to characterize the changes in homeownership rates across the country during this period.
   b. Do the trends appear to be uniform across the United States or are they unique to certain regions of the country? Explain.

72. Data on community hospital average daily cost are provided in the file P03_72.xls. Do the yearly distributions of average cost figures tend to become more or less variable over the given time period? Justify your answer with descriptive graphs and/or relevant summary measures.

73. The median sales price of existing one-family homes in selected metropolitan areas is the variable of interest in this exercise. Using the data contained in the file P03_73.xls, characterize the distribution of median sales prices of existing single-family homes in 1996. How is this distribution different from that for the median sales prices of such homes in 1992? Summarize the essential differences.

74. Are U.S. traffic fatalities related to the speed limit and/or road type? Consider the data found in the file P03_74.xls.
   a. Do the average number and/or variability in the number of traffic fatalities occurring on interstate highways tend to increase as the speed limit is raised above 55 miles per hour? Explain your answer.
   b. Do the average number and/or variability in the number of traffic fatalities occurring on noninterstate roads tend to increase as the speed limit rises above 35 miles per hour? Explain your answer.
   c. Do the average number and/or variability in the number of traffic fatalities occurring on a road with a posted speed limit of 55 miles per hour tend to change with the road type (i.e., interstate versus noninterstate highway)? Explain your answer.
   d. Based on these data, which combination of speed limit and road type appears to be most lethal for U.S. drivers?
   e. Based on these data, which combination of speed limit and road type appears to be safest for U.S. drivers?

75. Have greater or lesser proportions of Americans joined labor unions during the past two decades? Respond to this question by applying descriptive summary measures and graphical tools to the data provided in the file P03_75.xls. Interpret your output. What conclusions can you draw from an analysis of these data?

76. Consider the percentage of the U.S. population without health insurance coverage. The file P03_76.xls contains such percentages by state for 1998 through 2003.
   a. Describe the distribution of state percentages of Americans without health insurance coverage in the period 2002–2003. Be sure to employ both measures of central location and dispersion in developing your characterization of these data.
   b. Compare the 2002–2003 distribution with the corresponding set of percentages taken in 2001–2002. How are these two sets of figures similar? In what ways are they different?
   d. Based on your answers in parts b and c above, what would you expect to find upon analyzing similar data for 2003–2004?
77. Given data in the file P02_13.xls from a recent survey of chief executive officers from the largest U.S. public companies, apply your knowledge of numerical summary measures to determine whether the typical levels and variances of the 2003 annual salaries and bonuses earned by CEOs depend in part on the types of companies in which they serve.

78. Consider survey data collected from 1000 randomly selected Internet users, given in the file P02_43.xls.
   a. Use these data to formulate a profile of the typical female Internet user. Consider such attributes as age, education level, marital status, annual income, and family size in formulating your profile.
   b. Use these data to formulate a profile of the typical married Internet user. Consider such attributes as gender, age, education level, annual income, and family size in formulating your profile.
   c. Use these data to formulate a profile of the typical high-income (say, with an annual income in excess of $80,000) Internet user. Consider such attributes as gender, age, education level, marital status, and family size in formulating your profile.

79. Consider the economic development data in the file P02_06.xls.
   a. Use these data to formulate a profile of the typical household residing within each of the four neighborhood locations. Consider such attributes as family size, home ownership status, gross annual income(s) of household wage earner(s), monthly home mortgage or rent payment, average monthly expenditure on utilities, and the total indebtedness (excluding the value of a home mortgage) of the household in formulating your profile.
   b. Do differences arise in the mean or median income levels of those wage earners from households located in different quadrants of this neighborhood? If so, summarize these differences.
   c. Do differences arise in the mean or median monthly home mortgage or rent payment paid by households located in different quadrants of this neighborhood? If so, summarize these differences.
   d. Do differences arise in the mean or median debt levels of households located in different quadrants of this neighborhood? If so, summarize these differences.

80. A human resources manager at Beta Technologies, Inc., is interested in developing a profile of the highest paid full-time Beta employees based on the given representative sample of 52 of the company’s full-time workers in the file P02_01.xls. In particular, she is interested in determining the typical age, number of years of relevant full-time work experience prior to coming to Beta, number of years of full-time work experience at Beta, and number of years of post-secondary education for those employees in the highest quartile with respect to annual salary. Employ appropriate descriptive methods to help the human resources manager develop this desired profile.

81. Using cost-of-living data from the ACCRA Cost of Living Index (see the file P02_19.xls), examine the relationship between the geographical location of an urban area within the United States (e.g., northeast, southeast, midwest, northwest, or southwest) and its composite cost-of-living index. In other words, is the overall cost of living higher or lower and more or less variable for urban areas in particular geographical regions of the country? You must first assign the given urban areas systematically to one of several geographical regions before you can apply appropriate summary measures in responding to this question. Summarize your findings in detail.

82. The file P03_82.xls contains monthly interest rates on bonds that pay money a year after the day they are bought. It is often suggested that interest rates are more volatile (tend to change more) when interest rates are high. Do these data support this statement?

83. The file P03_83.xls contains data on 1000 of Marvak’s best customers. Marvak is a direct-marketing firm that sells electronic items. It has collected these data to learn more about its customers. The variables are self-explanatory, although a few cell notes have been added in row 3. Your boss at Marvak would like you to do the following.
   a. She wants a breakdown of gender by age group. That is, she wants a pivot table that lists, for each age group, the percentages of females and males. She wants to be able to access this information easily (with a couple of clicks) for any values of Home and Married.
   b. She guesses that customers with larger salaries tend to spend more at Marvak. To check this, she wants you to append a new variable called SalaryCat to the data set that contains the four category labels in column J. A customer with salary below $30,000 is categorized as “LowSal”; between $30,000 and $70,000 as “MedSal”; between $70,000 and $120,000 as “HighSal”; and over $120,000 as “HugeSal.” (By the way, no incomes are exactly equal to $30,000, $70,000, or $120,000.) You can use the lookup table in columns I and J to form this new SalaryCat column. Then create a pivot table that shows the average amount spent for each of the four salary categories, and comment briefly (on that sheet) whether your boss’s conjecture appears to be correct.

84. The file P03_84.txt contains the largest 100 public companies in the world, as listed in The Wall Street Journal on September 24, 1992, ranked by market value. The rankings are shown for 1992, as well as for 1991. The following variables are included:
   - Company: name of company
   - Location: 1 for United States, 2 for Japan/Australia, 3 for Europe
Bank: 1 if bank (or savings institution), 0 otherwise
Rank92: rank according to market value in 1992
Rank91: rank according to market value in 1991
MarketVal: market value in millions of U.S. dollars (12/31/91 exchange rates used)
Sales91: sales in 1991 in millions of U.S. dollars
PctChSales: percent change in sales from 1990, based on home currency
Profit91: profit in 1991 in millions of U.S. dollars
PctChProfit: percent change in profit from 1990, based on home currency

a. This file is in ASCII (text) form, with blanks between the items and names (text data) in double quotes. Open this file in Excel. (There is no “import” command; you simply open the file.) Excel will recognize that this is a text file, and a “wizard” will lead you through the steps to open it properly into Excel format. The key is that it is “delimited” with blanks. Once the file is opened, look at it to make sure everything is lined up correctly. Then use the Save As command to save the file as an .XLS file. (Once you do this, the .TXT version is no longer needed.)

b. Note that percentage changes from 1990 to 1991 are given for sales and profits. Use formulas to create variables Sales90 and Profit90, the sales and profits for 1990. [Hint: For example, the formula for Sales90 is 100*Sales91/(100+PctChSales).]

c. Create a scatterplot of Profit91 (vertical axis) versus Profit90. Most of the points lie close to a line. For this problem, consider an outlier to be any point that is obviously not very close to this line. Which companies are the worst outliers in this sense? In business terms, what makes these companies outliers?

d. The companies designated “banks” are clearly different from the other companies in that their sales figures are much larger. For this question, consider only the subset of nonbanks. Define an outlier with respect to any variable as an observation that is at least 1.5 IQRs above the third quartile or below the first quartile. (This is the boxplot definition of outliers.) How many outliers are there with respect to Sales91?

e. There is a variable that codes the location of the company: 1 for United States, 2 for Japan/Australia, 3 for Europe. With regard to Profit91, are there any obvious differences among these? (Use summary statistics and/or charts.)

f. The data in this file are somewhat dated. See if you can find similar, but up-to-date, data (possibly on the Web). Then do the same analysis on the new data.

The file P03_86.xls contains data on close to 1800 professional NFL football players. (In case you are not a pro football fan, there are two conferences in the NFL: the NFC and the AFC. Also, each player plays either on offense or defense.) The data for each player include his base salary, any bonus, and the total of the base salary and the bonus. For each part, do the requested analysis for the total salary.

a. Create histograms of the salaries for (1) all of the players, (2) all of the NFC players, and (3) all of the AFC players. Also, create tables of summary statistics for these three groups.

b. Proceed as in part a, but now make the distinction between offensive and defensive players instead of which conference they are in.

c. Repeat parts a and b, but now eliminate the quarterbacks (QB), who are often the highest paid players. Does it make a difference?

d. Use one or more pivot tables to break the data down in other interesting ways, such as by position. Your goal is to understand the salary structure in the NFL. Write up your results in a brief report.

The file P03_87.xls contains questionnaire data from a random sample of 200 TV viewers. (The variable name headings actually begin in row 25.) The questionnaire was taken by the local station XYZ, an affiliate of one of the three main networks. Like the local affiliates of the other two networks, XYZ’s local newsmagazine program follows the national news program. The purpose of the questionnaire was to discover characteristics of the viewing public, presumably with the intention of doing something to increase XYZ’s ratings. Your assignment is very open-ended—purposely so. Summarize any aspects of the data that you think are relevant—find means, proportions, scatterplots, pivot tables, whatever—to help XYZ management understand these viewers. (All 200 people watch both national and local news.)
The Dow Jones Industrial Average (DJIA) is a composite index of 30 of the largest “blue-chip” companies in the United States. It is probably the most quoted index from Wall Street, partly because it is old enough that many generations of investors have become accustomed to quoting it, and partly because the U.S. stock market is the world’s largest. Besides longevity, two other factors play a role in the Dow’s widespread popularity: It is understandable to most people, and it reliably indicates the market’s basic trend. As this edition was going to press, the Dow was hovering around 10,000, after having been above 11,000. It is difficult to predict where it might be when you read this.

Unlike most other market indexes that are weighted indexes (usually by market capitalization, that is, price times shares outstanding), the DJI is an unweighted index. It was originally an average, namely, the sum of the stock prices divided by the number of stocks. In fact, the very first average price of industrial stocks, on May 26, 1896, was 40.94. However, because of stock splits, the DJI is now calculated somewhat differently to preserve historical continuity. To calculate the DJI, the prices of the 30 stocks in the index are summed, and this sum is divided by a “divisor,” which is currently slightly greater than 0.33 varies over time.

The Dow originally consisted of 12 stocks in 1896 and increased to 20 in 1916. The 30-stock average made its debut in 1928, and the number has remained constant ever since. However, the 30 stocks comprising the Dow do not remain the same. For example, after the first edition of this book, SBC Communications, Home Depot, Honeywell International, Intel, Microsoft, and Citigroup have replaced Union Carbide, Sears Roebuck, Aluminum Company of America, Allied Signal, Chevron, and Traveler’s Group in the select 30. Since then, more changes have been made. The editors of The Wall Street Journal select the components of the DJI. They take a broad view of what “industrial” means. In essence, it is almost any company that isn’t in the transportation business and isn’t a utility. In choosing a new company for the DJI, they look among substantial industrial companies with a history of successful growth and wide interest among investors. The components of the DJI are not changed often. It isn’t a “hot stock” index, and the Journal editors believe that stability of composition enhances the trust that many people have in the averages. The most frequent reason for changing a stock is that something is happening to one of the components (for example, a company is being acquired).

Some people make predictions about where the stock market is headed in part on their interpretation of DJI movements, as well as movements of the transportation and utilities averages. But indexes don’t predict anything. They are doing their job if they accurately reflect where the market has been. However, there is a great deal of common ground between the economy and the market. Stock investors try to anticipate future profits, and corporate profits are a prime fuel for the U.S. economy. So, not surprisingly, the market frequently rises ahead of economic expansion and falls prior to economic slowdown or contraction. The trouble is that this relationship isn’t perfectly correlated; there are other factors that move markets and still others that affect the economy. Moreover, many people make the mistake of calibrating their economic expectations to the DJI’s movements. The result is, as Nobel-laureate economist Paul A. Samuelson put it, “The market has predicted nine of the last five recessions.”

As indicated previously, the DJI is not the only market index. There are two other Dow Jones indexes, one for transportation (DJT), consisting of 20 stocks, and one for utilities (DJU), consisting of 15 stocks. An elaborate analytical system dubbed Dow Theory holds that the DJT must “confirm” the movement of the industrial average for a market trend to have staying power. If the industrials reach a new high, the transportations would need to reach a new high to “confirm” the broad trend. The trend reverses when both averages experience sharp downturns at around the same time. If they diverge—for example, if the industrial average keeps climbing while the transportations decline—watch out! The underlying fundamentals of the Dow Theory
hold that the industrials make and the transportations take. If the transportations aren’t taking what the industrials are making, it portends economic weakness and market problems. Similarly, according to analysts who study the averages, a rise in utility stock prices indicates that investors anticipate falling interest rates because utilities are big borrowers and their profits are enhanced by lower interest costs. But the utility average tends to decline when investors expect rising interest rates. Because of this interest-rate sensitivity, the utility average is regarded by some as a leading indicator for the stock market as a whole.

This information and other interesting facts about the Dow Jones averages are available http://www.dowjones.com. Other Web sites have data on the averages themselves. We wrote an Excel macro to allow you to download the historical data (daily, weekly, or monthly) from Yahoo’s Web site for any of these indexes and the stock they contain. The resulting data are in the files DJI_StockQuery.xls, DJT_StockQuery.xls, and DJU_StockQuery.xls. Each file contains a sheet for the Dow Jones average and a sheet for each stock in the average. (The sheet names for the latter indicate the ticker symbols for the stocks, such as HP for Hewlett-Packard Co.) All of the data are summarized in a Consolidate Sheet.

Use the tools you’ve learned in the past two chapters to analyze these data sets (or more recent Dow Jones data sets if you can download them). Here are some suggested directions for analysis.

1. The return for any period is the percentage change in the price over that period. That is, the return is

   \[
   \frac{p_{\text{end}} - p_{\text{beg}}}{p_{\text{beg}}}
   \]

   where \( p_{\text{end}} \) is the ending price and \( p_{\text{beg}} \) is the beginning price. (The return also includes dividends, but you can ignore these here.) Are the weekly returns for the 30 stocks in the DJI highly correlated with each other? Are the weekly returns correlated with the DJI itself? What about daily returns? What about monthly returns? Answer the same questions for the DJT; for the DJU.

2. How would you evaluate portfolios of any of these stocks over some period of time? How much better (or worse) are some portfolios?

3. As stated previously, some analysts believe that the DJU is a leading indicator of the market. Do the data bear this out, assuming we identify the DJI as “the market”? One way to answer this is with “cross-correlations,” such as the correlation between the DJI today and the DJU a week ago. Alternatively, we could compare a time series graph of the DJI with a “shifted” version of the DJU.

4. Similarly, is there any relationship between the DJI and the DJT?
Following up on Case Study 3.1, there are many market indexes other than the Dow Jones averages. These include broad U.S. indexes such as the Nasdaq Composite Index (mainly technology stocks), the NYSE Composite Index (an index of many stocks on the New York Stock Exchange), the S&P 500 index (an index of 500 of the largest U.S. companies), and others. There are also U.S. indexes for particular industries, such as the AMEX Biotechnology Index, and foreign indexes, such as the AMEX Japan Index. Daily data (from the same source as in the previous case) for several of these indexes are listed in the file OtherIndexes_StockQuery.xls. Formulate and answer any interesting questions relating to these data and the Dow Jones data. In particular, do all of these indexes tend to move together, or do they tend to move in opposite directions?
A mean, as defined in this chapter, is a pretty simple concept—it is the average of a set of numbers. But even this simple concept can cause confusion if we aren’t careful. The data in Figure 3.49 are typical of data presented by marketing researchers for a type of product, in this case, beer. Each value is an average of the number of six-packs of beer purchased per customer during a month. For the individual brands, the value is the average only for the customers who purchased at least one six-pack of that brand. For example, the value for Miller is the average number of six-packs purchased of all of these brands for customers who purchased at least one six-pack of Miller. In contrast, the “Any” average is the average number of six-packs purchased of these brands for all customers in the population.

Is there a paradox in these averages? On first glance, it might appear unusual, or even impossible, that the “Any” average is less than each brand average. Make up your own (small) data set, where you list a number of customers, along with the numbers of six-packs of each brand of beer each customer purchased, and calculate the averages for your data that correspond to those in Figure 3.49. Do you get the same result (that the “Any” average is lower than all of the others)? Are you guaranteed to get this result? Does it depend on the amount of brand loyalty in your population, where brand loyalty is greater when customers tend to stick to the same brand, rather than buying multiple brands? Write up your results in a concise report.

Figure 3.49  Average beer purchases

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Miller</td>
<td>Budweiser</td>
<td>Coors</td>
<td>Michelob</td>
<td>Heineken</td>
<td>Old Milwaukee</td>
<td>Rolling Rock</td>
<td>Any</td>
</tr>
<tr>
<td>2</td>
<td>6.77</td>
<td>6.62</td>
<td>6.64</td>
<td>7.11</td>
<td>7.29</td>
<td>7.30</td>
<td>7.17</td>
<td>4.71</td>
</tr>
</tbody>
</table>