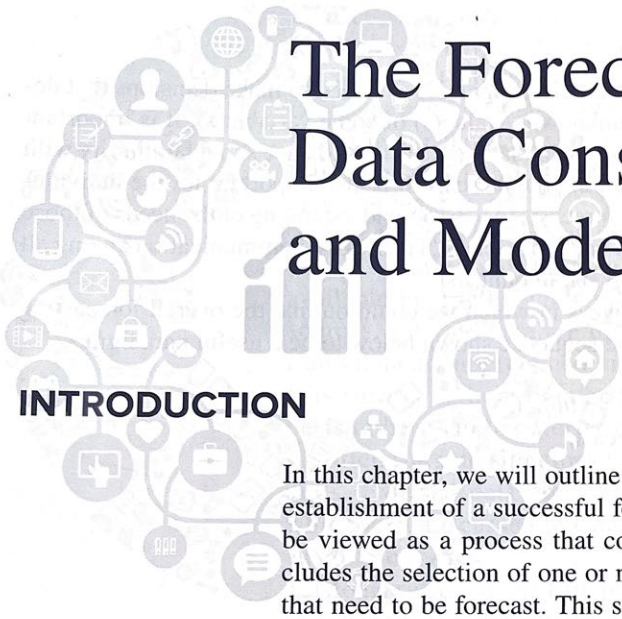
Chapter Two

The Forecast Process, Data Considerations, and Model Selection

INTRODUCTION



©VLADGRIN/Getty Images

In this chapter, we will outline a forecasting process that is a useful guide to the establishment of a successful forecasting system. It is important that forecasting be viewed as a process that contains certain key components. This process includes the selection of one or more forecasting techniques applicable to the data that need to be forecast. This selection, in turn, depends on the type of data that are available. In selecting a forecasting model, one should first evaluate the data for trend, seasonal, and cyclical components.

In evaluating a data series for its trend, seasonal, and cyclical components, it is useful to look at the data in graphic form. In this chapter, we evaluate data for the U.S. population, total new houses sold, disposable personal income, and The Gap sales to see which time-series components exist in each. This chapter also includes a review of statistics and an introduction to the use of autocorrelation coefficients, which can provide useful information about the underlying components in a time series.

LEARNING OBJECTIVES

After studying this chapter, you should be able to:

1. Explain a process for developing forecasts.
2. Distinguish between trend, seasonal, and cyclical data patterns.
3. Identify forecasting methods that would be good candidates for a given series to be forecast.
4. Explain the differences between the mean, median, and mode for a set of data.
5. Explain the most common measures of dispersion in data.
6. Discuss the normal and Student's t distributions.
7. Describe three common forms of statistical hypotheses.

8. Explain what a statistical correlation measures.
9. Explain how an autocorrelation function (ACF) can be useful in forecasting.

THE FORECAST PROCESS

The forecast process begins with recognizing the need to make decisions that depend on the future—and unknown—value(s) of some variable(s). It is important for managers who use forecasts in making decisions to have some familiarity with the methods used in developing the forecast. It is also important for the individuals involved in developing forecasts to have an understanding of the needs of those who make decisions based on the forecasts. Thus, good communication among all involved with forecasting is paramount.

There are a variety of ways in which we could outline the overall forecasting process. We have found the sequence shown below to be a useful paradigm.

1. Specify objectives.
2. Determine what to forecast.
3. Identify time dimensions.
4. Data considerations.
5. Model selection.
6. Model evaluation.
7. Forecast preparation.
8. Forecast presentation.
9. Tracking results.

This flow of relationships in the forecasting process will be discussed in more detail in Chapter 12, after a base of understanding of quantitative forecasting methods has been established.

It may seem obvious that the forecasting process should begin with a clear statement of objectives that includes how the forecast will be used in a decision context. Objectives and applications of the forecast should be discussed between the individual(s) involved in preparing the forecast and those who will utilize the results. Good communication at this phase will help ensure that the effort that goes into developing the forecast results in improved decision outcomes.

The second step of the process involves specifying explicitly what to forecast. For a traditional sales forecast, you must decide whether to forecast unit sales or dollar sales. Should the forecast be for total sales, or sales by product line, or sales by region? Should it include domestic sales, export sales, or both? A hospital may want to forecast patient load, which could be defined as admissions, discharges, patient-days, or acuity-days. In every forecasting situation, care must be taken to carefully determine exactly what variable(s) should be forecast.

Next, two different issues that relate to the time dimensions of the forecast need to be considered. One of these dimensions involves the length and periodicity of the forecast. Is the forecast needed on an annual, quarterly, monthly, weekly, or daily basis? In some situations, an even shorter time period may be necessary, such as in forecasting electricity demand for a generating facility. The second time dimension to be considered is related to the urgency of the forecast. If there is little time available before the forecast is needed, the choice of methods that can be used will be limited. When forecasting involves hundreds, or thousands, or tens of thousands SKUs (stock keeping units), a forecaster will be limited to methods that can be automated and done in an efficient manner.

The fourth element of the forecasting process involves a consideration of the quantity and the type of data that are available. Some data may be available internally, while other data may have to be obtained from external sources. Internal data are often the easiest to obtain, but not always. Sometimes data are not retained in a form useful for the development of a forecast. It is surprising how frequently we find that data are kept only on an annual basis rather than for shorter periods, such as quarterly or monthly. Similarly, we often run into situations where only dollar values are available rather than units. External data are available from a wide variety of sources, some of which were discussed in Chapter 1. Most external sources provide data in an electronic form.

Model selection, the fifth phase of our forecasting process, depends on a number of criteria, including:

1. The pattern exhibited by the data
2. The quantity of historic data available
3. The length of the forecast horizon

Table 2.1 summarizes how these criteria relate to the quantitative forecasting methods that are included in this text. While all of these criteria are important, the first is the most important. We will discuss the evaluation of patterns in data and model selection in greater detail after completing a review of the forecasting process.

Fit refers to how well the model works retrospectively. *Accuracy* relates to how well the model works in the forecast horizon

The sixth phase of the forecasting process involves testing models on the specific series to be forecast. This is often done by evaluating how each model works in a retrospective sense. That is, we see how well the results fit the historic data that were used in developing the models. A measures such as the mean absolute percentage error (MAPE) is typically used for this evaluation. We often make a distinction between *fit* and *accuracy* in evaluating a forecast model. *Fit* refers to how well the model works retrospectively. *Accuracy* relates to how well the model works in the forecast horizon (i.e., outside the period used to develop the model). When we have sufficient data, we often use a "holdout" period to evaluate forecast accuracy. For example, suppose that you have 10 years of historic quarterly sales data and want to make a two-year

TABLE 2.1 A Guide to Selecting a Traditional Forecasting Method* **

Forecasting Method	Data Pattern	Quantity of Historical Data (Number of Observations)	Forecast Horizon
Naive	Stationary	1 or 2	Very short
Moving averages	Stationary	Number equal to the periods in the moving average	Very short
Exponential smoothing			
Simple	Stationary	5 to 10	Short
Adaptive response	Stationary	10 to 15	Short
Holt's	Linear trend	10 to 15	Short to medium
Winters'	Trend and seasonality	At least 4 or 5 per season	Short to medium
Bass model	S-curve	Small, 3 to 10	Short to Medium
Regression-based			
Trend	Linear and nonlinear trend with or without seasonality	Minimum of 10 with 4 or 5 per season if seasonality is included	Short to medium
Causal	Can handle nearly all data patterns	Recommend a minimum of 10 per independent variable	Short, medium, and long
Time-series decomposition	Can handle trend, seasonal, and cyclical patterns	Enough to see two peaks and two troughs in the cycle	Short, medium, and long
ARIMA	Stationary or transformed to stationary	Minimum of 50	Short, medium, and long

*The methods presented in this table are the most commonly used techniques. There are many other methods available, most of which are included in the ForecastX™ software that accompanies this text.

** Data and text mining methods that are discussed in Chapters 8 through 11 are based on different criteria.

forecast. In developing and evaluating potential models, you might use just the first eight years of data to forecast the last two years of the historical series. MAPEs could then be calculated for the two holdout years to determine which model or models provide the most accurate forecasts. These models would then be respecified using all 10 years of historic data, and a forecast would be developed for the true forecast horizon. If the models selected in phase 6 did not yield an acceptable level of accuracy, you would return to step 5 and select an alternative model.

Phase 7, forecast preparation, is the natural result of having found models that are believed to produce acceptably accurate results. We recommend that more than one technique be used whenever possible. When two, or more, methods that have different information bases are used, their combination will frequently provide better forecasts than would either method alone. The process of combining forecasts is sufficiently important that the appendix to Chapter 5 is devoted to this topic.

The eighth phase of the forecasting process involves the presentation of forecast results to those who rely on them to make decisions. Here, clear communication is critical. Sometimes analysts who develop forecasts become so enamored with the sophistication of their models that they focus on technical issues rather than on the substance of the forecast. In both written and oral presentations, the use of objective visual representations of the results is very important.[1]

Finally, the forecasting process should include continuous tracking of how well forecasts compare with the actual values observed during the forecast horizon. Over time, even the best of models are likely to deteriorate in terms of accuracy and need to be respecified, or replaced with an alternative method. Forecasters can learn from their mistakes. A careful review of forecast errors may be helpful in leading to a better understanding of what causes deviations between the actual and forecast series.

TREND, SEASONAL, AND CYCLICAL DATA PATTERNS

The data that historically have been used most often in forecasting are time series. For example, you might have sales data by month from January 2010 through December 2017, or you might have the number of visitors to a national park every year for a 30-year period, or you might have stock prices on a daily basis for several years. These would all be examples of time-series data.

Such time series can display a wide variety of patterns when plotted over time. Displaying data in a time-series plot is an important first step in identifying various component parts of the time series. A time series is likely to contain some, or all, of the following components:

Trend
Seasonal
Cyclical
Irregular (often called random)

We will first define and discuss each of these in general terms, and then we will look at several specific data series to see which components we can visualize through graphic analyses.

Data are considered *stationary* when there is neither a positive nor a negative trend.

The *trend* in a time series is the long-term change in the level of the data. If, over an extended period of time, the series moves upward, we say that the data show a positive trend. If the level of the data diminishes over time, there is a negative trend. Data are considered ***stationary*** when there is neither a positive nor a negative trend (i.e., the series is essentially flat in the long term).

[1] An excellent discussion of how to present information in graphic form can be found in Edward R. Tufte, *The Visual Display of Quantitative Information* (Cheshire, CT: Graphics Press, 1983).

A *seasonal* pattern occurs in a time series when there is a regular variation in the level of the data that repeats itself at the same time each year.

A *seasonal* pattern occurs in a time series when there is a regular variation in the level of the data that repeats itself at the same time each year. For example, ski lodges in Killington, Vermont, have very regular high occupancy rates during December, January, and February (as well as regular low occupancy rates in the spring of the year). Housing starts are always stronger in the spring and summer than during the fall and winter. Retail sales for many products tend to peak in November and December because of holiday sales. Most university enrollments are higher in the fall than in the winter or spring and are typically the lowest in the summer. All of these patterns recur with reasonable regularity year after year. No doubt you can think of many other examples of time-series data for which you would expect similar seasonal patterns.

A *cyclical* pattern is represented by wavelike upward and downward movements of the data around the long-term trend. Cyclical fluctuations are of longer duration and are less regular than are seasonal fluctuations. The causes of cyclical fluctuations are less readily apparent as well. They are usually attributed to the ups and downs in the general level of business activity that are frequently referred to as *business cycles*.

The *irregular* component of a time series contains the fluctuations that are not part of the other three components. These are often called *random* fluctuations. As such, they are the most difficult to capture in a forecasting model. There is always some noise in the data that would be a part of this irregular component.

To illustrate these components, let us analyze three specific sets of data. One of these is a quarterly series for the population in the United States (POP), which is an important driver for many types of business activities. POP tends to increase at a fairly constant linear rate. The second series is monthly data for total new houses sold (TNHS), which is also important for many businesses to forecast since it drives so many other types of sales (such as drapes, furniture, appliances, etc.). TNHS has a lot of seasonality, some upward trend since 2010, and a cyclical component. The third series is disposable personal income (DPI, in billions of dollars), which also has a positive trend. The trend is slightly nonlinear with DPI increasing at an increasing rate. DPI is also sometimes referred to as a prime mover because income is the revenue source for personal consumption.

The top panel of Figure 2.1 shows a times-series plot of population on a quarterly basis starting with the first quarter of 2000 and ending with the last quarter of 2016. From a visual inspection of this graph, it is fairly easy to see that there has been a positive trend to POP over the period shown. The long-term linear trend is shown by the dotted red line in the lower panel of Figure 2.1. (In later chapters, you will learn how to determine an equation for this long-term trend line.) You see that population is nonstationary. Because POP is nonstationary, some models would not be appropriate in forecasting POP (see Table 2.1).

Total new houses sold (TNHS) is plotted in Figure 2.2 for the period from January 1978 through July 2007. Probably the most striking feature of this visualization of the TNHS data is the regular and sharp upward and downward movements

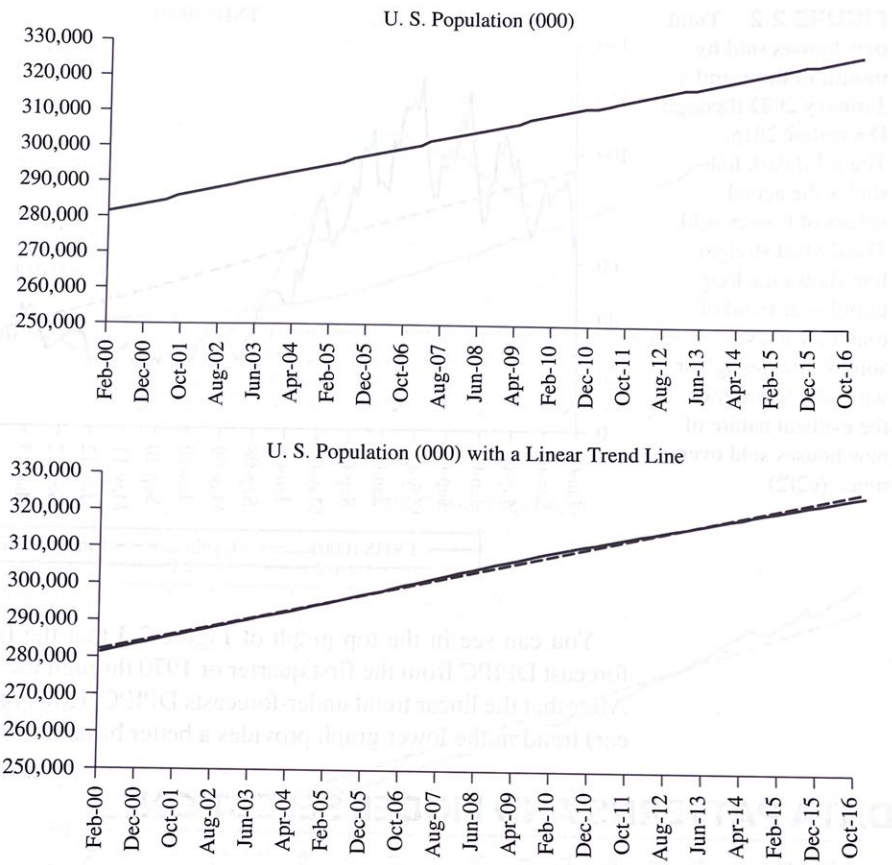


FIGURE 2.1 Total population of the United States in thousands. The upper graph shows the U.S. population growth from the first quarter of 2000 through the last quarter of 2016. The bottom graph illustrates how closely a linear trend approximates the actual population growth. (c2f1)

that repeat year after year. This indicates a seasonal pattern, with housing sales reaching a peak in the spring of each year. Overall, there also appears to be some upward trend to the data and some cyclical movement as well.

The straight solid line in Figure 2.2 shows the long-term trend in the TNHS series. The dotted line, which moves above and below the long-term trend but is smoother than the plot of TNHS, is what the TNHS series looks like after the seasonality has been removed. Such a series is said to be "deseasonalized," or "seasonally adjusted" (SA). By comparing the deseasonalized series with the trend, the cyclical nature of houses sold becomes clearer. You will learn how to deseasonalize data in Chapter 6.

Now let us turn to a visual analysis of disposable personal income per capita (DPIPC). Figure 2.3 shows DPIPC from the first quarter of 1959 through the last quarter of 2016. Clearly, there is an upward trend in the data, and it is a trend that appears to be accelerating slightly (i.e., becoming increasingly steep). You will learn to forecast such nonlinear trends later in this text. There does not appear to be a cyclical component to the series, and there is no seasonality.

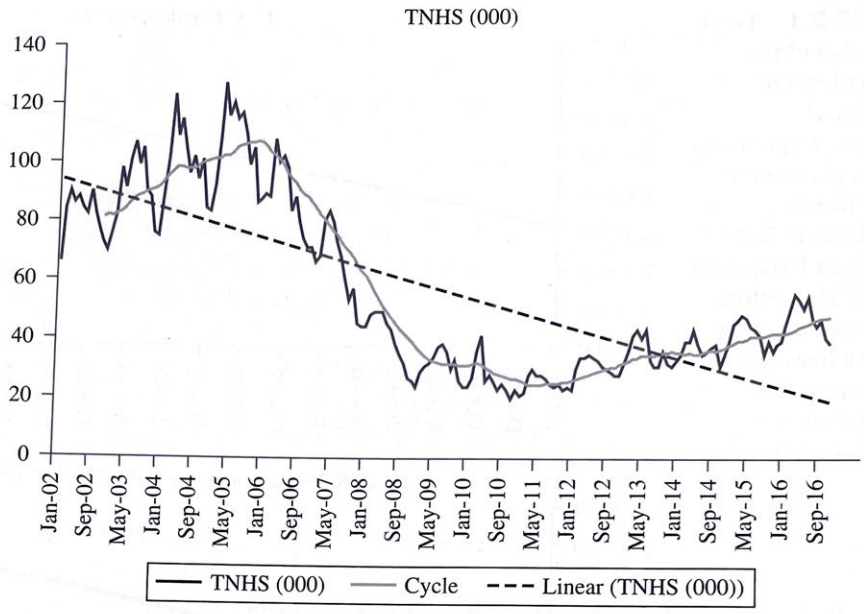


FIGURE 2.2 Total new houses sold by month in thousands: January 2002 through December 2016. The solid dark line shows the actual values of houses sold. The dashed straight line shows the long-term linear trend of total new houses sold, while the lighter wavelike line shows the cyclical nature of new houses sold over time. (c2f2)

You can see in the top graph of Figure 2.3 that the linear trend would over-forecast DPIPC from the first quarter or 1970 through the second quarter of 2003. After that the linear trend under-forecasts DPIPC. However, the quadratic (nonlinear) trend in the lower graph provides a better basis for forecasting.

DATA PATTERNS AND MODEL SELECTION

The pattern that exists in the data is an important consideration in determining which forecasting techniques are appropriate.

As discussed earlier in this chapter, the pattern that exists in the data is an important consideration in determining which forecasting techniques are appropriate. On the basis only of the pattern of data, let us apply the information in Table 2.1 to determine which methods might be good candidates for forecasting each of the three specific series just discussed and plotted in Figures 2.1 through 2.3.

For POP, which has a trend but no cycle and no seasonality, the following might be most appropriate:

Holt's exponential smoothing
Linear regression trend

Total new houses sold (TNHS) has a trend, seasonality, and a cycle. Therefore, some likely candidate models for forecasting TNHS would include:

Winters' exponential smoothing
Linear regression trend with seasonal adjustment
Causal regression
Time-series decomposition

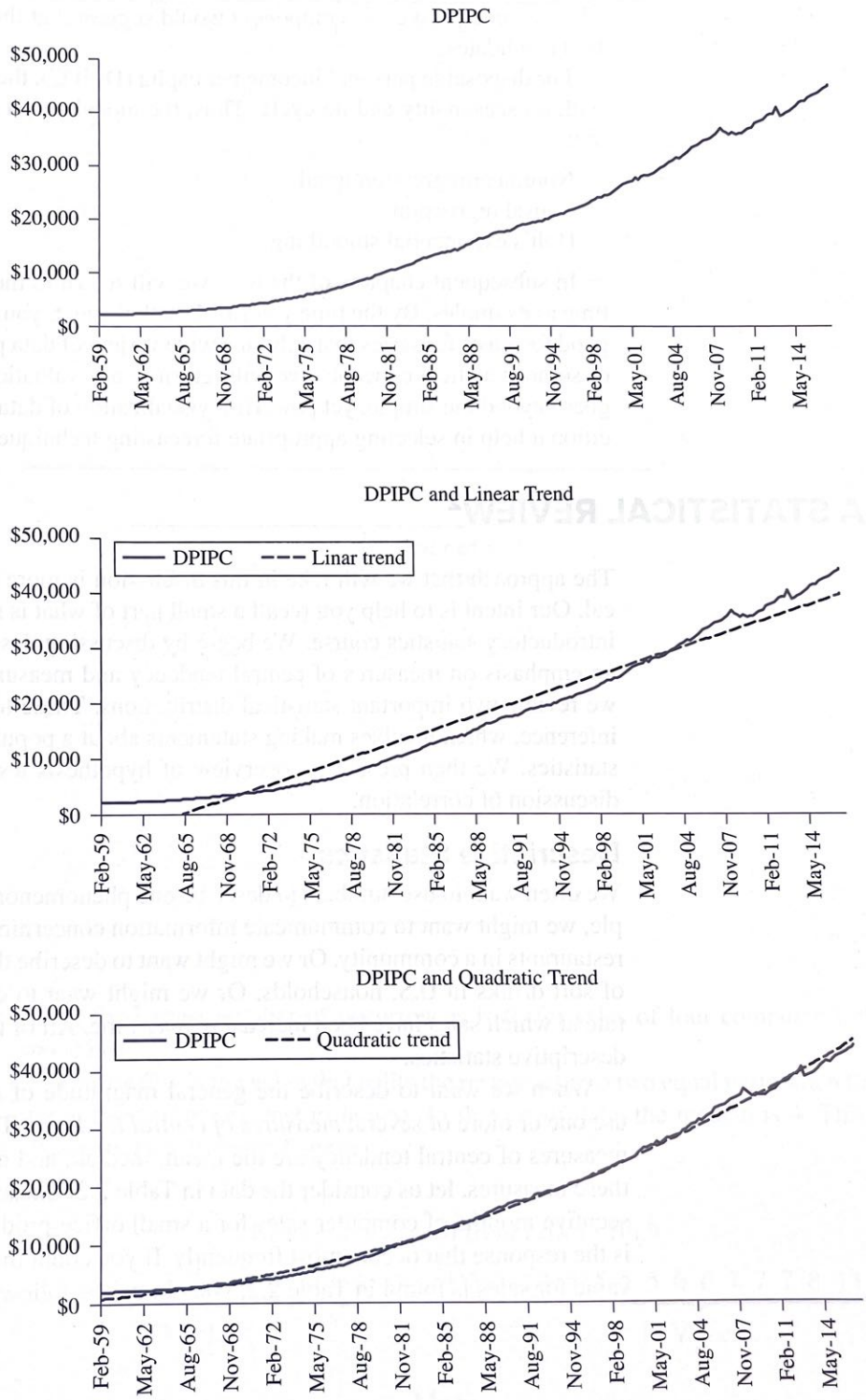


FIGURE 2.3
Disposable personal income per capita in the United States. The solid line in the upper graph shows the actual DPIPC from the first quarter of 1959 through the last quarter of 2016, The dashed line in the lower graph shows a linear trend for DPIPI. The bottom graph illustrates how a quadratic trend (shown by the dashed line) approximates the actual DPIPC more accurately than the linear trend. (c2f3)

The existence of a cycle component would suggest that the latter two may be the best candidates.

For disposable personal income per capita (DPIPC), there is a nonlinear trend, with no seasonality and no cycle. Thus, the models most likely to be successful are:

Nonlinear regression trend
Causal regression
Holt's exponential smoothing

In subsequent chapters of the text, we will return to these series from time to time as examples. By the time you finish with the text, you will be able to develop good forecasts for series that exhibit a wide variety of data patterns. After a review of some statistical concepts, we will return to an evaluation of data patterns that goes beyond the simple, yet powerful, visualization of data and that will be of additional help in selecting appropriate forecasting techniques.

A STATISTICAL REVIEW[2]

The approach that we will take in this discussion is more intuitive than theoretical. Our intent is to help you recall a small part of what is normally covered in an introductory statistics course. We begin by discussing descriptive statistics, with an emphasis on measures of central tendency and measures of dispersion. Next we review two important statistical distributions. These topics lead to statistical inference, which involves making statements about a population based on sample statistics. We then present an overview of hypothesis testing and finish with a discussion of correlation.

Descriptive Statistics

We often want to use numbers to describe one phenomenon or another. For example, we might want to communicate information concerning the sales of fast-food restaurants in a community. Or we might want to describe the typical consumption of soft drinks in U.S. households. Or we might want to convey to someone the rate at which sales have been increasing over time. All of these call for the use of descriptive statistics.

When we want to describe the general magnitude of some variable, we can use one or more of several *measures of central tendency*. The three most common measures of central tendency are the mean, median, and mode. To grasp each of these measures, let us consider the data in Table 2.2. These data represent 25 consecutive months of computer sales for a small office-products retailer. The *mode* is the response that occurs most frequently. If you count the number of times each value for sales is found in Table 2.2, you obtain the following results:

[2] Students with a good statistical background may be able to skip this section.

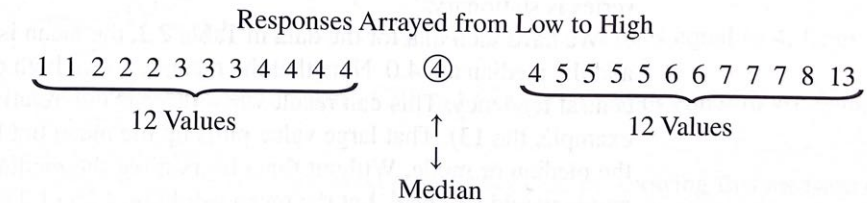
TABLE 2.2
Twenty-Five Consecutive Months of Total Sales (c2t2)

Month	Sales	Month	Sales
1	3	14	4
2	4	15	7
3	5	16	3
4	1	17	4
5	5	18	2
6	3	19	5
7	6	20	7
8	2	21	4
9	7	22	5
10	8	23	2
11	1	24	6
12	13	25	4
13	4		

Sales	Number of Occurrences
1	2
2	3
3	3
4	6
5	4
6	2
7	3
8	1
13	1
Total	25

Since the largest number of occurrences is 6 (for sales of four computers), the mode is 4.

The *median* is the value that splits the responses into two equal parts when they are arrayed from smallest to largest. In this set of data, the median is 4. This is shown in the following diagram:



There are 12 numbers to the left of the circled 4, and 12 numbers to the right. When there are an even number of observations, the median is the midpoint of the two center values. For example, in the series 1, 4, 6, 10, the median is 5. Note that the median may be a number that is not actually in the data array.

The *mean* is the arithmetic average of all the numbers in the data set. To find the mean, add up all the values and divide by the number of observations. If the set of numbers is a population, rather than a sample, the mean is designated by the Greek mu (μ). It is calculated as:

$$\mu = \sum_{i=1}^{N} X_i/N$$

where the subscript i is used to identify each X value and

$$\sum_{i=1}^{N} X_i$$

means the sum of all the values of X_i, in which i ranges from 1 to N. X is simply a shorthand way of representing a variable. For the data in Table 2.2, $X_3 = 5$ and $X_{15} = 7$. N represents the total number of elements, or observations, in the population. In this case $N = 25$. Adding up all 25 values, we get:

$$\Sigma X = 115$$

Note that we have dropped the subscript here. This will often be done to simplify the notation. The population mean is then:

$$\mu = \Sigma X/N = 115/25 = 4.6$$

If the data represent a sample (i.e., a portion of the entire population), the mean is designated $\overline{X}$ and the number of elements in the sample is designated n. Thus, a sample mean is:

$$\overline{X} = \sum_{i=1}^{n} X_i/n$$

If the data in Table 2.2 represented a sample of months, the mean would be calculated as:

$$\overline{X} = \Sigma X/n = 115/25 = 4.6$$

All three of these measures of central tendency provide some feel for what we might think of as a "typical case." For example, knowing that the median and mode for sales are both 4 and the mean is 4.6 gives you an idea about what is a typical month's sales.

These sales data are plotted over time in Figure 2.4, along with the trend line. You see in this plot that sales fluctuate around a nearly flat trend. Thus, this sales series is stationary.

We have seen that for the data in Table 2.2, the mean is 4.6, and both the mode and the median are 4.0. Note that the mean is above both of the other measures of central tendency. This can result when there is one relatively large value (in this example, the 13). That large value pulls up the mean but has little or no effect on the median or mode. Without that observation, the median and mode for this example would still be 4, but the mean would be 4.25 (4.25 = 102/24).

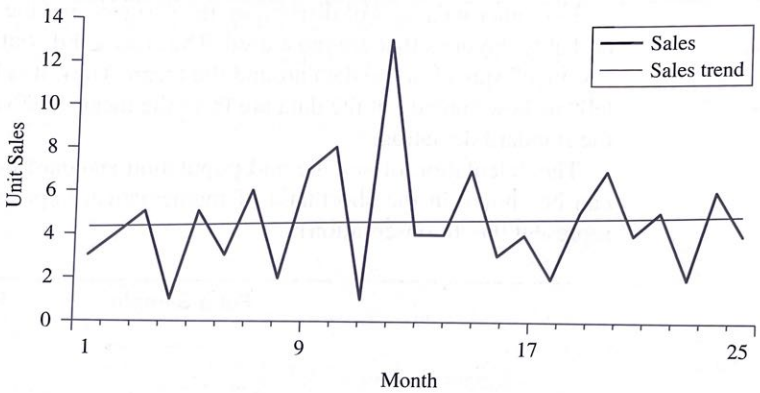


FIGURE 2.4 Sales and sales trend (c2f4). For this sales series, the trend is almost perfectly flat, so that the data are stationary. Note that the level of the trend line is fairly close to the sample mean of 4.6.

Let us now consider dispersion in data. A measure of dispersion tells us something about how spread out (or dispersed) the data are. Such information helps us to gain a clearer picture of the phenomenon being investigated than we get by looking just at a measure of central tendency. Look, for example, at the following two data sets marked *A* and *B*:

A:	18	19	20	21	22
B:	0	10	20	30	40

In both cases, the mean and median are 20. (Since no value occurs more frequently than the others, there is no mode.) However, the two data sets are really very different. Measures of dispersion can be helpful in conveying such a difference.

The simplest measure of dispersion is the *range,* which is the difference between the smallest value and the greatest value. In Table 2.2, the smallest value is 1 (observations 4 and 11); the greatest is 13 (observation 12). Thus,

$$\begin{aligned}\text{Range} &= \text{Greatest value} - \text{Smallest value}\\ &= 13 - 1\\ &= 12\end{aligned}$$

For the two data sets *A* and *B* just given, the range for *A* is 4 and the range for *B* is 40.

Think for a moment about the different perception you get from the following two statements:

"The data set *A* has a mean of 20 and a range of values equal to 4, from 18 to 22."

"The data set *B* has a mean of 20 and a range of values equal to 40, from 0 to 40."

You can see how much your perception is affected by knowing this measure of dispersion in addition to the mean.

Two other measures of dispersion, the variance and the standard deviation, are probably the ones that are most used. The standard deviation is a measure of the "average" spread of the data around the mean. Thus, it is based on the mean and tells us how spread out the data are from the mean. The variance is the square of the standard deviation.

The calculation of sample and population standard deviations and variances can be shown in the shorthand of mathematical expressions as follows (let X_i represent the ith observation):

	For a Sample	For a Population
Standard deviation	$S = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-1}}$	$\sigma = \sqrt{\frac{\Sigma(X_i - \mu)^2}{N}}$
Variance	$S^2 = \frac{\Sigma(X_i - \bar{X})^2}{n-1}$	$\sigma^2 = \frac{\Sigma(X_i - \mu)^2}{N}$

For the computer sales data in Table 2.2, the calculations of the standard deviation and variance are illustrated in Table 2.3. Note that the sum of the unsquared differences between each observation and the mean is equal to zero. This is always true. Squaring the differences gets around the problem of offsetting positive and negative differences. The standard deviation for the sales data is (assuming the data represent a sample) 2.582 units around a mean of 4.6. That is, the "average" spread around the mean is 2.582. The corresponding variance is 6.667 "units squared." You can see that the interpretation of the variance is a bit awkward. What is a "squared computer"? Because of this squaring of the units of measurement, the variance is less useful in communicating dispersion than is the standard deviation. In statistical analysis, however, the variance is frequently far more important and useful than the standard deviation. Thus, both are important to know and understand.

Look back at the two small data sets A and B referred to earlier. For both sets, the mean was 20. Assuming that these are both samples, the standard deviations are:

For A: $S = 1.58$
For B: $S = 15.8$

You see that knowing both the mean and the standard deviation gives you a much better understanding of the data than you would have if you knew only the mean.

The Normal Distribution

Many statistical distributions are important for various applications. Two of them—the normal distribution and Student's t-distribution—are particularly useful for the applications in forecasting to be discussed in this text. In this section,

TABLE 2.3
Calculation of the Standard Deviation and Variance for the Computer Sales Data (Assuming a Sample) (c2t3)

Observation Number	Computer Sales (X_i)	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
1	3	−1.6	2.56
2	4	−0.6	0.36
3	5	0.4	0.16
4	1	−3.6	12.96
5	5	0.4	0.16
6	3	−1.6	2.56
7	6	1.4	1.96
8	2	−2.6	6.76
9	7	2.4	5.76
10	8	3.4	11.56
11	1	−3.6	12.96
12	13	8.4	70.56
13	4	−0.6	0.36
14	4	−0.6	0.36
15	7	2.4	5.76
16	3	−1.6	2.56
17	4	−0.6	0.36
18	2	−2.6	6.76
19	5	0.4	0.16
20	7	2.4	5.76
21	4	−0.6	0.36
22	5	0.4	0.16
23	2	−2.6	6.76
24	6	1.4	1.96
25	4	−0.6	0.36
Total	115	0.0	160.00

$$\text{Mean} = \bar{X} = \frac{\Sigma X_i}{n} = \frac{115}{25} = 4.6$$

$$\text{Variance} = S^2 = \frac{\Sigma(X_i - \bar{X})^2}{n-1} = \frac{160}{25-1} = 6.667$$

$$\text{Standard deviation} = S = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{160}{24}} = \sqrt{6.667} = 2.582$$

we will describe the normal distribution. We will consider the *t*-distribution in a later section.

The normal distribution for a continuous random variable is fully defined by just two characteristics: the mean and the variance (or standard deviation) of the

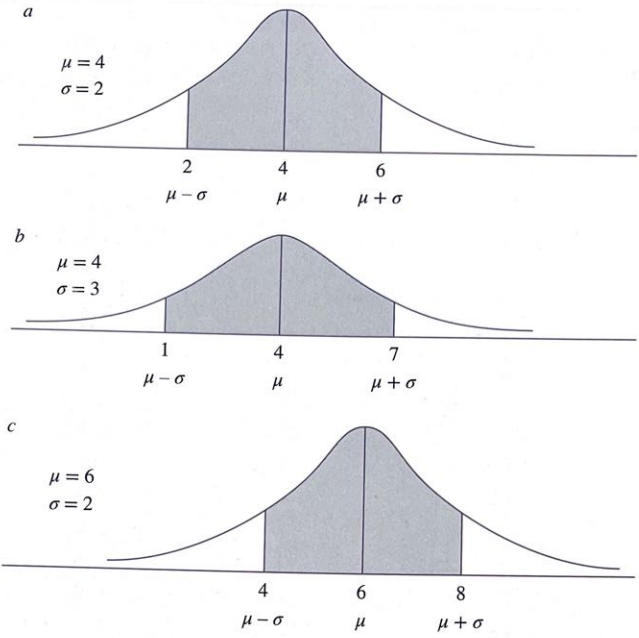


FIGURE 2.5
Three normal distributions. The top and middle distributions have the same mean but different standard deviations. The top and bottom distributions have the same standard deviation but different means.

variable. A graph of the normal distribution has a bell shape such as the three distributions shown in Figure 2.5.[3] All such normal distributions are symmetrical around the mean. Thus, 50 percent of the distribution is above the mean and 50 percent is below the mean. It follows that the median must equal the mean when the distribution is normal.

In Figure 2.5, the top graph represents the normal curve for a variable with a population mean of 4 and a standard deviation of 2. The middle graph is for a variable with the same mean but a standard deviation of 3. The lower graph is for a normal distribution with a mean of 6 and a standard deviation of 2. While each is unique, the three graphs have similar shapes, and they have an important common feature: for each of these graphs the shaded area represents roughly 68 percent of the area under the curve.

This brings us to an important property of all normal curves. The area between one standard deviation above the mean and one standard deviation below

[3] Technically, these are probability density functions, for which the area under the curve between any two points on the horizontal axis represents the probability of observing an occurrence between those two points. For a continuous random variable, the probability of any particular value occurring is considered zero, because there are an infinite number of possible values in any interval. Thus, we discuss only probabilities that values of the variable will lie between specified pairs of points.

the mean includes approximately 68 percent of the area under the curve. Thus, if we were to draw an element at random from a population with a normal distribution, there is a 68 percent chance that it would be in the interval $\mu \pm 1\sigma$. This 68 percent is represented by the shaded areas of the graphs in Figure 2.5.

If you remember that the normal distribution is symmetrical, you will realize that 34 percent must be in the shaded area to the left of the mean and 34 percent in the shaded area to the right of the mean. Since the total area to the right (or left) of the mean is 50 percent, the area in either tail of the distribution must be the remaining 16 percent (these are the unshaded regions in the graphs in Figure 2.5).

If you extend the range to plus or minus two standard deviations from the mean, roughly 95 percent of the area would be in that interval. And if you go out three standard deviations in both directions from the mean, over 99.7 percent of the area would be included. These concepts can be summarized as follows:

$\mu \pm 1\sigma$ includes about 68 percent of the area
$\mu \pm 2\sigma$ includes about 95 percent of the area
$\mu \pm 3\sigma$ includes over 99 percent of the area

These three rules of thumb are helpful to remember.

In Figure 2.5, you saw three similar yet different normal distributions. How many such distributions are there? There may be billions of them. Every variable or measurement you might consider could have a different normal distribution. And yet any statistics text you look in will have just one normal distribution. The reason for this is that every other normal distribution can be transformed easily into a *standard* normal distribution called the Z-distribution. The transformation is simple:

$$Z = \frac{X - \mu}{\sigma}$$

In this way, any observed value (X) can be standardized to a corresponding Z-value. The Z-value measures the number of standard deviations by which X differs from the mean. If the calculated Z-value is positive, then X lies to the right of the mean (X is larger than μ). If the calculated Z-value is negative, then X lies to the left of the mean (X is smaller than μ).

The standard normal distribution is shown in Table 2.4. Note that it is centered on zero. For every value of X, there is a corresponding value for Z, which can be found by using the transformation shown in the preceding equation. For example, let us calculate the Z-values that correspond to $X = 40$ and to $X = 65$ assuming a standard deviation of 10:

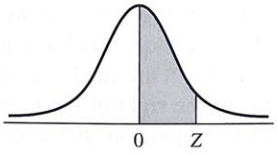
$$Z = \frac{X - \mu}{\sigma}$$

For $X = 40$, $$Z = \frac{40 - 50}{10} = -1$$

For $X = 65$, $$Z = \frac{65 - 50}{10} = 1.5$$

TABLE 2.4 The Standard Normal Distribution*

Source: Hall, Jr., Owen P. and Adelman, Harvey M. *Computerized Business Statistics* (Homewood, Ill.: Richard D. Irwin, 1987), p. 91.



Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2109	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.2051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.49865	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
4.0	.49997									

*Z is the standard normal variable. Other variables can be transformed to Z as follows:

$$Z = \frac{X - \mu}{\sigma}$$

For Z = 1.96, the shaded area in the distribution is 0.4750 (found at the intersection of the 1.9 row and the .06 column).

Through this process every normal variable can be transformed to the standard normal variable Z.

The normal distribution provides a background for many types of data analysis. However, it is not typically appropriate for work with sample data, and in business we almost always have sample data. When working with sample data, we use the t-distribution.

The Student's *t*-Distribution

When the population standard deviation is not known, or when the sample size is small, the Student's t-distribution should be used rather than the normal distribution. The Student's t-distribution resembles the normal distribution but is somewhat more spread out for small sample sizes. As the sample size becomes very large, the two distributions become the same. Like the normal distribution, the t-distribution is centered at zero (i.e., has a mean of zero) and is symmetrical.

Since the t-distribution depends on the number of degrees of freedom (df), there are many t-distributions. The number of degrees of freedom appropriate for a given application depends on the specific characteristics of the analysis. Throughout this text, we will specify the value for df in each application. Table 2.5 has a t-distribution for 29 different degrees of freedom plus infinity. The body of this table contains t-values such that the shaded area in the graph is equal to the subscript on t at the top of each column for the number of degrees of freedom (df) listed along the left.

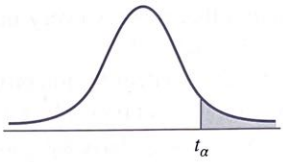
To learn how to read the t-table, let us consider three examples. First, what value of t would correspond to 5 percent of the area in the shaded region if there are 15 degrees of freedom? To answer this, go to the row for 15 degrees of freedom, then to the column that has .050 for the subscript on t. The t-value at the intersection of that row and column is 1.753. Second, if there are 26 degrees of freedom and the t-value is 2.479, how much area would be in the shaded region? Looking across the row for 26 degrees of freedom, we see that 2.479 is in the column for which t is subscripted with .010. Thus, 1 percent of the area would be in that tail.

For our third example, consider the following question: If there are 85 degrees of freedom, what value of t would be associated with finding 97.5 percent of the area in the unshaded portion of the curve? For any number of degrees of freedom greater than 29, we would use the infinity (Inf.) row of the table. If we want 97.5 percent in the clear area, then 2.5 percent must be in the shaded region. Thus, we need the column for which t is subscripted with .025. The t-value at the intersection of this row and column is found to be 1.960. (Note that this is the same as the Z-value for which 2.5 percent would be in the tail, or 0.4750 is in the shaded section of the normal distribution shown in Table 2.4.)

While t-tables are usually limited to four or five areas in the tail of the distribution and perhaps 30 levels for degrees of freedom, most statistical software incorporates the equation for the t-distribution and will give exact areas, given any

TABLE 2.5
Student's *t*-Distribution*

Source: Hall, Jr., Owen P. and Adelman, Harvey M. *Computerized Business Statistics* (Homewood, Ill.: Richard D. Irwin, 1987), p. 91.



df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
Inf.	1.282	1.645	1.960	2.326	2.576

*The t-distribution is used for standardizing when the population standard deviation is unknown and the sample standard deviation is used in its place.

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

t-value and the appropriate number of degrees of freedom. *We will rely on the t-distribution extensively in* Chapters 4 *and* 5 *as part of the evaluation of statistical significance in regression models.*

From Sample to Population: Statistical Inference

We are usually much less interested in a sample than in the population from which the sample is drawn. The reason for looking at a sample is almost always to provide a basis for making some inference about the whole population. For example, suppose we are interested in marketing a new service in Oregon and want to know something about the income per person in the state. Over 3.5 million people live in Oregon. Clearly, trying to contact all of them to determine the mean income per person would be impractical and very costly. Instead we might select a sample and make an inference about the population based on the responses of the people in that sample of Oregon residents.

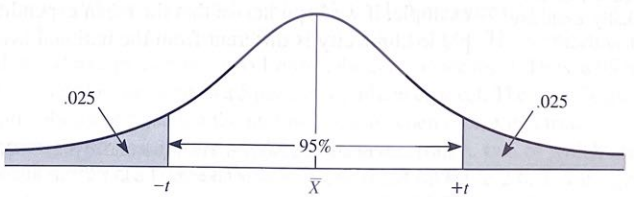
A sample statistic is our best point estimate of the corresponding population parameter. While it is best, it is also likely to be wrong. Thus, in making an inference about a population, it is usually desirable to make an interval estimate.

For example, an interval estimate of the population mean is one that is centered on the sample mean and extends above and below that value by an amount that is determined by how confident we want to be, by how large a sample we have, and by the variability in the data. These elements are captured in the following equation for a confidence interval:

$$\mu = \bar{X} \pm t(s/\sqrt{n})$$

The ratio $s/\sqrt{n}$ is called the standard error of the sample mean and measures dispersion for sample means. The *t*-value is determined from Table 2.5 after choosing the number of degrees of freedom ($n - 1$ in this case) and the level of confidence we desire as reflected by the area in the shaded tail of the distribution.

If we want a 95 percent confidence interval that is symmetrical around the mean, we would want a total of 5 percent in the two extreme tails of the distribution. Thus, 2.5 percent would be in each tail. The following diagram will help you see this:



The t-value that would correspond to 2.5 percent in each tail can be determined from Table 2.5, given the appropriate number of degrees of freedom. Several examples follow:

Number of Degrees of Freedom	t-Value for 95% Confidence Interval
5	2.571
10	2.228
20	2.086
50	1.960
100	1.960

Suppose that a sample of 100 responses gives a mean of \$25,000 and a standard deviation of \$5,000. Our best point estimate for the population mean would be \$25,000, and a 95 percent confidence interval would be:

$$\mu = 25{,}000 \pm 1.96(5{,}000/\sqrt{100})$$
$$= 25{,}000 \pm 980$$

that is,

$$24{,}020 \leq \mu \leq 25{,}980$$

See if you can correctly find the endpoints for a 90 percent confidence interval given this same set of sample results.[4]

Hypothesis Testing

Frequently we have a theory or hypothesis that we would like to evaluate statistically. For example, we might hypothesize that the mean expenditure on entertainment in some city is equal to the national average for all age groups. Or we may theorize that consumption of soft drinks by retired people is less than the national level. Or we may want to evaluate the assumption that women professionals work more than the standard 40-hour work week. All of these can be evaluated by using an appropriate hypothesis testing procedure.

The process begins by setting up two hypotheses, the null hypothesis (designated H_0:) and the alternative hypothesis (designated H_1:). These two hypotheses should be structured so that they are mutually exclusive and exhaustive. For example, if we hypothesize that the mean expenditure on entertainment by people in some city is different from the national average, the null and alternative

The process begins by setting up two hypotheses, the null hypothesis (designated H_0:) and the alternative hypothesis (designated H_1:). These two hypotheses should be structured so that they are mutually exclusive and exhaustive.

[4] The lower bound is \$24,177.5; the upper bound is \$25,822.5. Notice that at this lower confidence level, the value of t is smaller (other things equal), and thus the confidence interval is narrower.

hypotheses would be (let μ_0 = the national average and μ = this city's population mean):

Case I
- H_0: $\mu = \mu_0$
- i.e., H_0: The city mean equals the national mean.
- H_1: $\mu \neq \mu_0$
- i.e., H_1: The city mean is not equal to the national mean.

If we theorize that the consumption of soft drinks by retired people is *less* than the national average, the null and alternative hypotheses would be (let μ_0 = the national average and μ = the mean for retired people):

Case II
- H_0: $\mu \geq \mu_0$
- i.e., H_0: The mean for retired people is greater than or equal to the national average.
- H_1: $\mu < \mu_0$
- i.e., H_1: The mean for retired people is less than the national average.

If we want to evaluate the assumption that women professionals work *more* than the standard 40-hour work week, the null and alternative hypotheses would be (let μ_0 = the standard work week and μ = the mean for professional women):

Case III
- H_0: $\mu \leq \mu_0$
- i.e., H_0: The mean for professional women is less than or equal to the standard.
- H_1: $\mu > \mu_0$
- i.e., H_1: The mean for professional women is greater than the standard.

In each of these cases, the null and alternative hypotheses are mutually exclusive and exhaustive.

In statistical hypothesis testing, the approach is to see whether you find sufficient evidence to reject the null hypothesis. If so, the alternative is found to have support. For questions of the type we are considering, this is done by using a *t*-test. To perform a *t*-test, we must first determine how confident we want to be in our decision regarding whether or not to reject the null hypothesis. In most business applications, a 95 percent confidence level is used. A measure that is closely related to the confidence level is the significance level for the test. The significance level, often denoted α (alpha), is equal to 1 minus the confidence level. Thus, a 95 percent confidence level is the same as a 5 percent significance level. The significance level is the probability of rejecting the null hypothesis when in fact it is true.

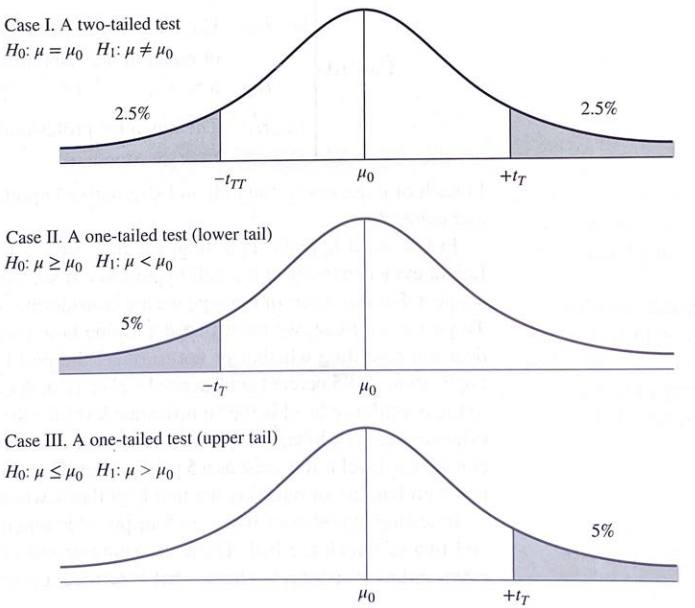
In testing hypotheses, there are four possible outcomes, two of which are good and two of which are bad. These are summarized in Table 2.6. If we reject H_0: when in fact it is true, we have what is termed a *type I error*. The other possible

TABLE 2.6 Type I and Type II Errors

Statistical Decision	The Truth: H_0: Is True	The Truth: H_0: Is Not True
Reject H_0:	Type I error	No error
Fail to Reject H_0:	No error	Type II error

error results when we fail to reject a null hypothesis that is in fact incorrect. This is a *type II error*. These two errors are related in that by reducing the chance of a type I error we increase the chance of a type II error and vice versa. Most of the time, greater attention is given to type I errors. The probability of making a type I error is determined by the significance level (α) we select for the hypothesis test. If the cost of a type I error is large, we would use a low α, perhaps 1 percent or less.

Hypothesis tests may be one- or two-tailed tests. When the sign in the alternative hypothesis is an unequal sign ($\neq$), the test is a two-tailed test. Otherwise, a one-tailed test is appropriate. For a two-tailed test, the significance level (α) is split equally into the two tails of the distribution. For a one-tailed test, the entire significance level (α) goes in the one tail of the distribution that is indicated by the direction of the inequality sign in the alternative hypothesis. Consider the three situations described a few paragraphs back. These are summarized in the following diagrams, which show where the significance level would be (a 5 percent significance level is used in all three cases).



The t_T values are determined from a *t*-distribution, such as that in Table 2.5, at the appropriate number of degrees of freedom ($n - 1$, in the examples used here) and for the tail areas indicated in these diagrams ($\alpha/2$ for two-tailed tests and α for one-tailed tests).

For each hypothesis test, a *t*-value is calculated (t_{calc}) and compared with the critical value from the *t*-distribution (t_T). If the calculated value is further into the tail of the distribution than the table value, we have an observation that is extreme, given the assumption inherent in H_0, and so H_0 is rejected. That is, we have sufficient evidence to reject the null hypothesis (H_0) when the *absolute value* of t_{calc} is greater than t_T. Otherwise we fail to reject the premise in H_0.

The calculated *t*-statistic is found as follows:

$$t_{calc} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

where $\bar{X}$ is our sample mean and our best point estimate of μ. The value we are testing against is μ_0. The sample standard deviation is s and the sample size is n.

Let us now apply these concepts to our three situations. Starting with case I, let us assume that a sample of 49 people resulted in a mean of \$200 per month with a standard deviation of \$84. The national average is \$220 per month. The hypotheses are:

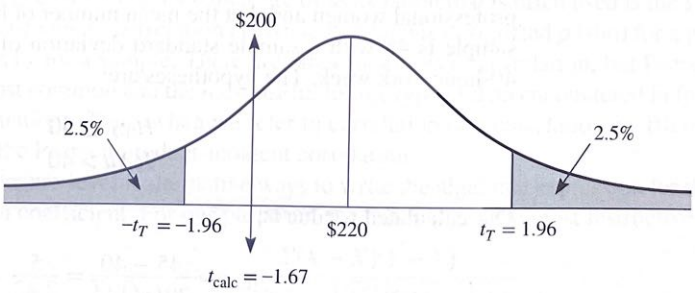
$$H_0: \mu = 220$$
$$H_1: \mu \neq 220$$

The calculated value is:

$$t_{calc} = \frac{200 - 220}{84/\sqrt{49}} = \frac{-20}{12} = -1.67$$

If we want a 95 percent confidence level ($\alpha = 0.05$), the critical or table value of t is $\pm$ 1.96. Notice that the $t_{.025}$ column of Table 2.5 was used. This is because we have a two-tailed test, and the α of 0.05 is split equally between the two tails. Since our calculated *t*-value (t_{calc}) has an absolute value that is less than the critical value from the *t*-table (t_T), we fail to reject the null hypothesis. Thus, we conclude that the evidence from this sample is not sufficient to say that entertainment expenditures by people in this city are any different from the national average.

This result is summarized in the following diagram:



We see here that the observed mean of \$200 or its corresponding t-value (-1.67) is not extreme. That is, it does not fall into either of the shaded areas. These shaded areas taken together are often called the *rejection region*, because t_{calc} values in the shaded areas would call for rejection of H_0.

Let us now look at case II. Assume that for a sample of 25 retired people the mean was 1.2 six-packs per week with a standard deviation of 0.6. The national average (μ_0) is 1.5. The hypotheses are:

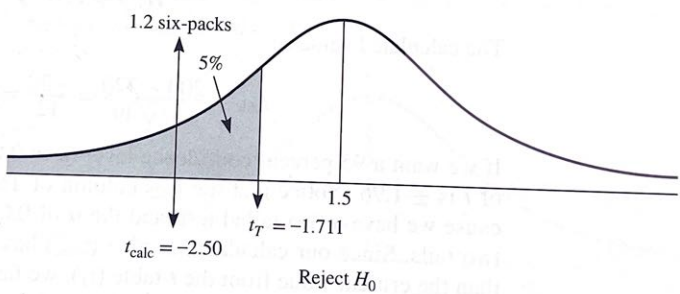
$$H_0: \mu \geq 1.5$$
$$H_1: \mu < 1.5$$

The calculated t-value is:

$$t_{calc} = \frac{1.2 - 1.5}{0.6/\sqrt{25}} = \frac{-0.3}{0.12} = -2.50$$

The critical value from the t-distribution in Table 2.5, assuming a 95 percent confidence level ($\alpha = 0.05$), is $t_T = -1.711$. Note that there are 24 degrees of freedom. Since the absolute value of t_{calc} is greater than the table value of t, we reject H_0. Thus, we conclude that there is sufficient evidence to support the notion that retired people consume fewer soft drinks than the national average.

This result is shown in graphic form as follows:



Here we see that the sample mean of 1.2 is extreme, given $\alpha = 0.05$ and $df = 24$, and so we reject H_0. The calculated value of t falls in the rejection region.

Finally, let us consider case III. We will assume that we have a sample of 144 professional women and that the mean number of hours per week worked for that sample is 45 with a sample standard deviation of 29. The national norm is the 40-hour work week. The hypotheses are:

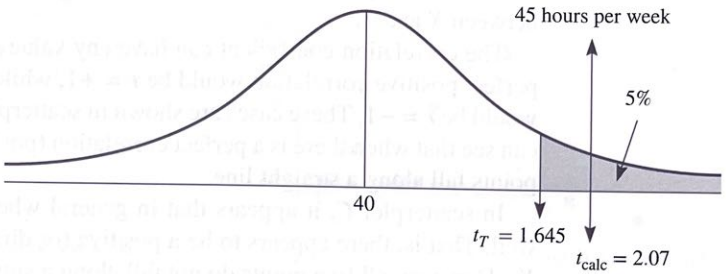
$$H_0: \mu \leq 40$$
$$H_1: \mu > 40$$

Our calculated t-value is:

$$t_{calc} = \frac{45 - 40}{29/\sqrt{144}} = \frac{5}{2.42} = 2.07$$

The relevant table value is 1.645 ($\alpha = 0.05$ and $df = 143$). Since $t_{calc} > t_T$, we reject the null hypothesis and conclude that the mean for professional women is greater than 40 hours per week.

This result is shown graphically as follows:



The calculated t-value lies in the shaded (or rejection) region, and so H_0 is rejected.

The t-tests illustrated in this section involved making judgments about a population mean based on information from a sample. In each t-test, the calculated value of t was determined by dividing some difference $(\overline{X} - \mu_0)$ by a standard error $(s/\sqrt{n})$. All t-statistics are calculated in this general way:

$$t = \frac{\text{the difference being evaluated}}{\text{the corresponding standard error}}$$

We will use this general form later in this chapter as well as in subsequent chapters of the text when t-tests are appropriate.

There are other statistical tests and other distributions that are applicable to forecasting. These include F-tests, Durbin-Watson tests, and chi-square tests, which will be discussed later in the text as they are applied. If you have a basic understanding of the use of t-tests, these other statistical tests will not be difficult to use.

Correlation

It is often useful to have a measure of the degree of association between two variables. For example, if you believe that sales may be affected by expenditures on advertising, you might want to measure the degree of association between sales and advertising. One measure of association that is often used is the Pearson product-moment correlation coefficient, which is designated ρ (rho) for a population and r for a sample. There are other measures of correlation, but Pearson's is the most common and the most useful for the type of data encountered in forecasting situations. Thus, when we refer to correlation or a correlation coefficient, we mean the Pearson product-moment correlation.

There are several alternative ways to write the algebraic expression for the correlation coefficient. For our purposes, the following is the most instructive:

$$r = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{\sqrt{\left[\Sigma\left(X - \overline{X}\right)^2\right]\left[\Sigma\left(Y - \overline{Y}\right)^2\right]}}$$

where X and Y represent the two variables of interest (e.g., advertising and sales). This is the sample correlation coefficient. The calculation of the population correlation coefficient (ρ) is strictly analogous except that the population means for X and Y would be used rather than the sample means. It is important to note that the correlation coefficient defined here measures the degree of linear association between X and Y.

The correlation coefficient can have any value in the range from -1 to $+1$. A perfect positive correlation would be $r = +1$, while a perfect negative correlation would be $r = -1$. These cases are shown in scatterplots A and B of Figure 2.6. You can see that when there is a perfect correlation (positive or negative) all of the data points fall along a straight line.

In scatterplot C, it appears that in general when X increases, Y_C increases as well. That is, there appears to be a positive (or direct) association between X and Y_C. However, all five points do not fall along a single straight line, and so there is not a perfect linear association. In this case, the correlation coefficient is $+0.79$. Scatterplot D shows a negative (or inverse) association between X and Y_D, but one that is not perfectly linear. For scatterplot D, $r = -0.89$.

The remaining two scatterplots in Figure 2.6 illustrate cases for which the correlation coefficient is zero. In both cases, there is no linear association between the variables. However, note that in panel F there is a clear nonlinear association between X and Y_F.

We could perform a hypothesis test to determine whether the value of a sample correlation coefficient (r) gives us reason to believe that the true population correlation coefficient (ρ) is significantly different from zero. If it is not, then there would be no linear association between the two measures. The hypothesis test would be:

$$H_0: \rho = 0$$
$$H_1: \rho \neq 0$$

and t would be calculated as:

$$t = \frac{r - 0}{\sqrt{(1 - r^2)/(n - 2)}}$$

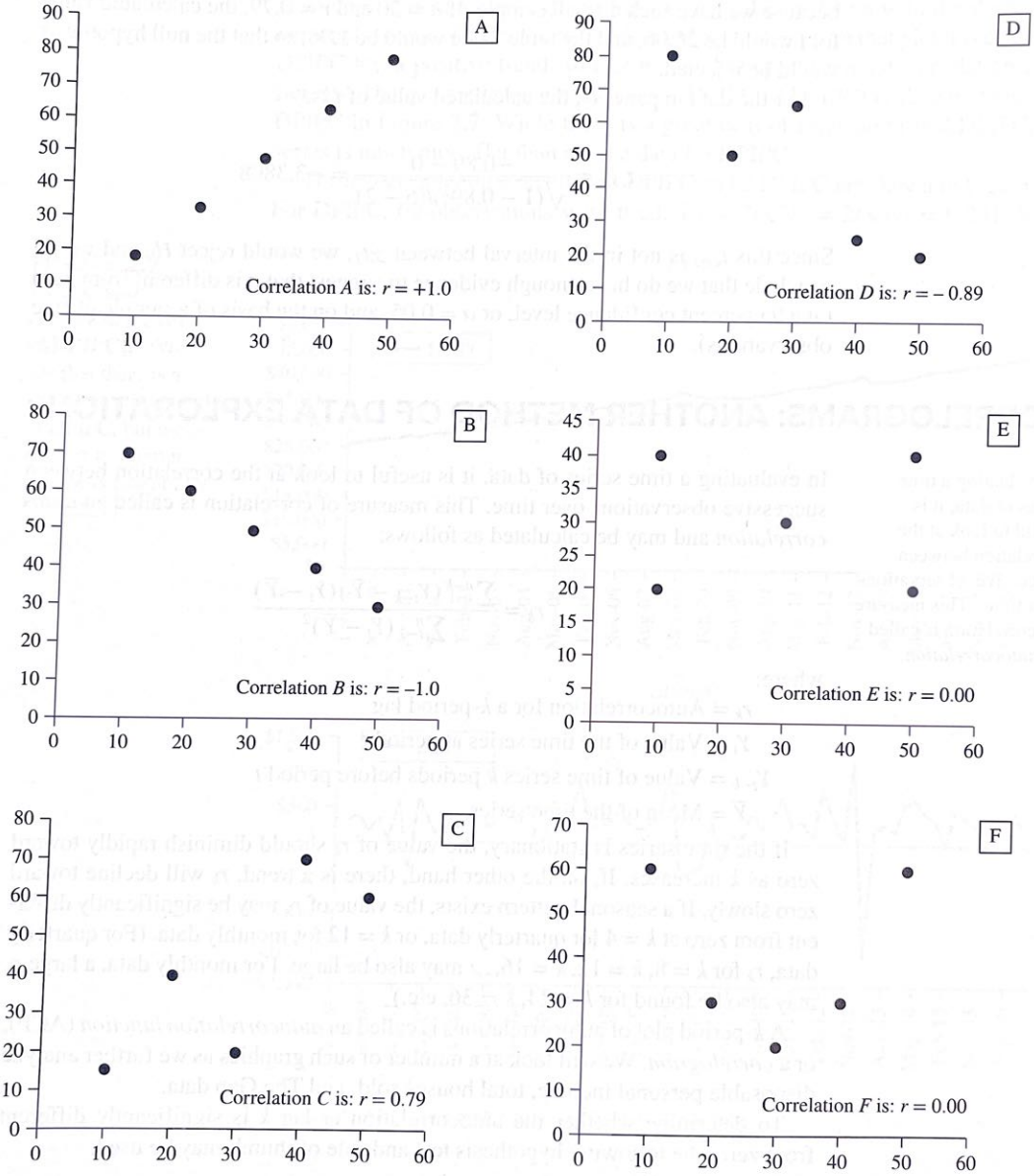
where $\sqrt{(1 - r^2)/(n - 2)}$ is the standard error of r.

Let us apply this to the data in scatterplots C and D of Figure 2.6. In both of these cases, for a two-tailed test, with $\alpha = 0.05$ and $n = 5$, the table value of t_T is 3.182 (there are $n - 2$, or 3 degrees of freedom for this test). For panel C the calculated value of t is:

$$t_{\text{calc}} = \frac{0.79 - 0}{\sqrt{[1 - (0.79)^2]/(5 - 2)}}$$
$$= \frac{0.79}{\sqrt{0.3759/3}} = \frac{0.79}{\sqrt{0.1253}} = -2.2318$$

Since t_{calc} is in the interval between $\pm t_T$ (i.e., ± 3.182), we would fail to reject the null hypothesis on the basis of a sample of five observations at a 95 percent

FIGURE 2.6 Representative scatterplots with the corresponding correlation coefficients.
These scatterplots show correlation coefficients that range from a perfect positive correlation (*A*) and a perfect negative correlation (*B*) to zero correlations (*E* and *F*).



confidence level ($\alpha = 0.05$). Thus, we conclude that there is not enough evidence to say that ρ is different from zero. While the $r = 0.79$ is a fairly strong correlation, we are not able to say it is significantly different from zero in this case, largely because we have such a small sample. If $n = 50$ and $r = 0.79$, the calculated value for t would be 26.06, and the table value would be 1.96, so that the null hypothesis would be rejected.

For the data in panel D, the calculated value of t is:

$$t_{\text{calc}} = \frac{-0.89 - 0}{\sqrt{(1 - 0.89^2)/(5 - 2)}} = -3.3808$$

Since this t_{calc} is not in the interval between $\pm t_T$, we would reject H_0 and would conclude that we do have enough evidence to suggest that ρ is different from zero (at a 95 percent confidence level, or $\alpha = 0.05$, and on the basis of a sample of five observations).

CORRELOGRAMS: ANOTHER METHOD OF DATA EXPLORATION

In evaluating a time series of data, it is useful to look at the correlation between successive observations over time. This measure of correlation is called an *autocorrelation*.

In evaluating a time series of data, it is useful to look at the correlation between successive observations over time. This measure of correlation is called an *autocorrelation* and may be calculated as follows:

$$r_k = \frac{\sum_{t=1}^{n-k} (Y_{t-k} - \overline{Y})(Y_t - \overline{Y})}{\sum_{t-1}^{n} (Y_t - \overline{Y})^2}$$

where:

r_k = Autocorrelation for a k-period lag

Y_t = Value of the time series at period t

Y_{t-k} = Value of time series k periods before period t

$\overline{Y}$ = Mean of the time series

If the time series is stationary, the value of r_k should diminish rapidly toward zero as k increases. If, on the other hand, there is a trend, r_k will decline toward zero slowly. If a seasonal pattern exists, the value of r_k may be significantly different from zero at $k = 4$ for quarterly data, or $k = 12$ for monthly data. (For quarterly data, r_k for $k = 8$, $k = 12$, $k = 16$, ... may also be large. For monthly data, a large r_k may also be found for $k = 24$, $k = 36$, etc.)

A k-period plot of autocorrelations is called an *autocorrelation function* (ACF), or a *correlogram*. We will look at a number of such graphics as we further analyze disposable personal income, total houses sold, and The Gap data.

To determine whether the autocorrelation at lag k is significantly different from zero, the following hypothesis test and rule of thumb may be used:

$$H_0: \rho_k = 0$$
$$H_1: \rho_k \neq 0$$

For any k, reject H_0 if $|r_k|>2/\sqrt{n}$, where n is the number of observations. This rule of thumb is for a 95 percent confidence level.[5]

The use of autocorrelations and correlograms can be illustrated by looking at some of the data used earlier in this chapter. Let us begin with the disposable personal income (DPIPC) data graphed in Figure 2.7. From that plot it is clear that DPIPC has a positive trend, so that we might expect high autocorrelation coefficients. The quarter-to-quarter change in DPIPC (ΔDPIPC) is shown along with DPIPC in Figure 2.7. While there is a great deal of fluctuation in ΔDPIPC, the series is much more flat than are the data for DPIPC.

The autocorrelation structures of DPIPC and ΔDPIPC are shown in Figure 2.8. For DPIPC, 69 observations were used. Thus, $2/\sqrt{n} = 2/\sqrt{69} = 0.241$. Since

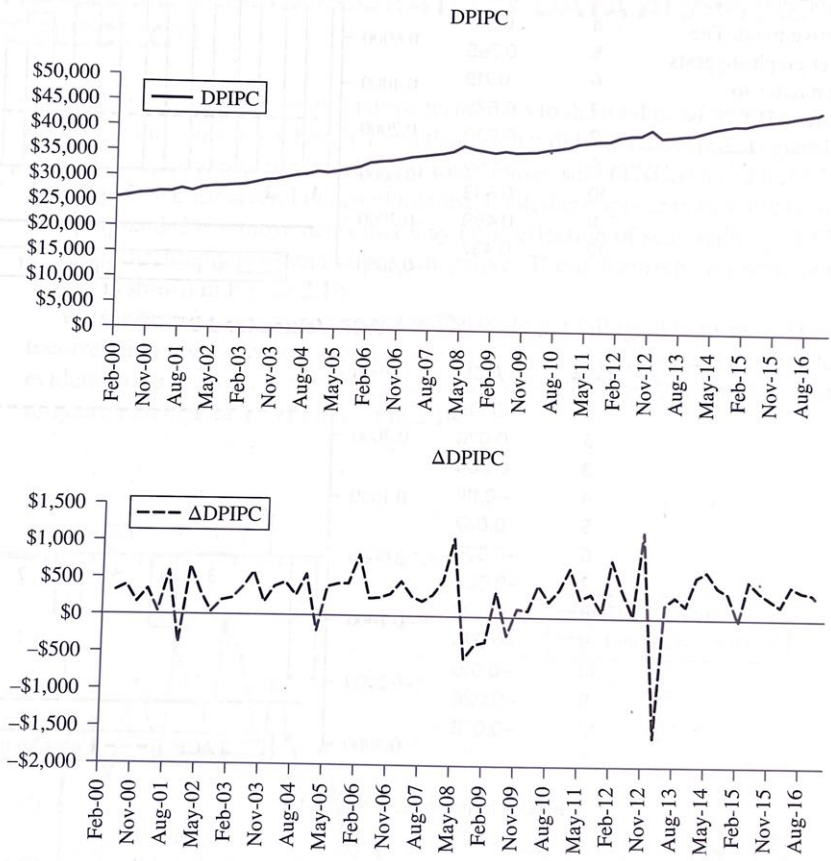


FIGURE 2.7
DPIPC and change in DPIPC (ΔDPIPC). We see that there is a strong positive trend in DPIPC, but the quarter-to-quarter change is essentially flat. (c2f7)

[5] The complete t-test would be to reject H_0 if $|t_{calc}| > t_T$, where:

$$t_{calc} = \frac{(r_k - 0)}{1/\sqrt{(n-k)}}$$

and t_T is from the t-table for $\alpha/2$ and $n - k$ degrees of freedom (n = number of observations, k = period of the lag).

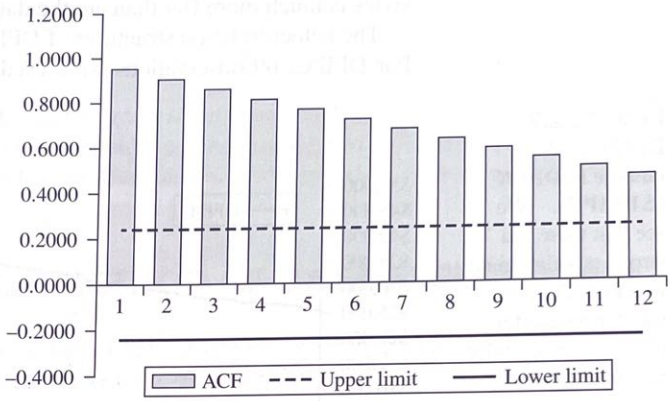
all of the autocorrelation coefficients in Figure 2.8 are greater than 0.241, we can conclude that they are all significantly different from zero. Therefore, we have additional evidence of a trend in the DPIPC data.[6] The actual 95 percent confidence interval is shown by the two horizontal lines labeled "Upper limit" and "Lower limit."

If we want to try a forecasting method for DPIPC that requires stationary data, we must first transform the DPIPC data to a stationary series. Often this

FIGURE 2.8
The ACF graphs for DPIPC and ΔDPIPC. From the upper graph, we see evidence that DPIPC does have a positive trend. The lower graph suggests that quarter-to-quarter ΔDPIPC is stationary. (c2f8)

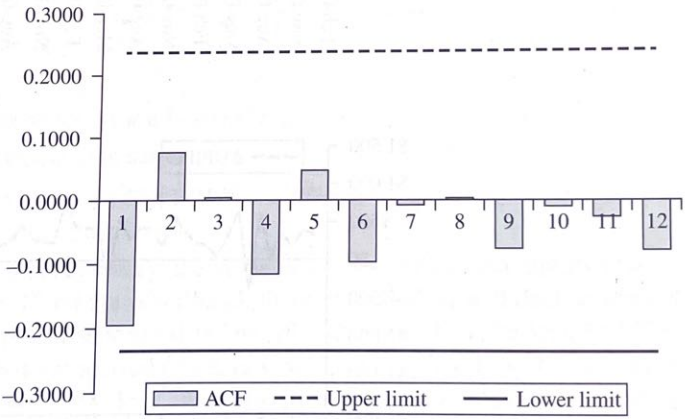
ACF Values for DPIPC

Obs	ACF
1	0.951
2	0.903
3	0.857
4	0.810
5	0.765
6	0.719
7	0.675
8	0.629
9	0.586
10	0.543
11	0.499
12	0.457



ACF Values for ΔDPIPC

Obs	ACF
1	−0.195
2	0.076
3	0.004
4	−0.116
5	0.047
6	−0.097
7	−0.008
8	0.003
9	−0.076
10	−0.010
11	−0.026
12	−0.078



[6] The more formal hypothesis test is:

$$H_0: \rho_k = 0$$
$$H_1: \rho_k \neq 0$$

and the calculated *t*-ratio is:

$$t_{calc} = \frac{r_k - 0}{1/\sqrt{n-k}}$$

For example, for $k = 12$ where $r_k = 0.8124$,

$$t_{calc} = \frac{0.8124 - 0}{1/\sqrt{199-12}} = 11.109$$

which is greater than the table value of 1.96 at $\alpha/2 = 0.025$ (a 95 percent confidence level).

can be done by using first differences. For DPIPC, the first differences can be calculated as:

$$\Delta\text{DPIPC}_t = \text{DPIPC}_t - \text{DPIPC}_{t-1}$$

where ΔDPIPC_t is the first difference (or change) in DPIPC. We can check for stationarity in ΔDPIPC by examining the autocorrelation structure for ΔDPIPC as shown in Figure 2.8. For ΔDPIPC, the autocorrelations are all within the upper and lower bounds, so this series is stationary.

TOTAL NEW HOUSES SOLD: EXPLORATORY DATA ANALYSIS AND MODEL SELECTION

Let us apply exploratory data analysis techniques to the total new houses sold data that were introduced in Chapter 1 and that often will be used as an example in the text. Figure 2.9 shows the raw data for total houses sold (TNHS) and a trend line. In this plot, we see several things of interest. First, there appear to be fairly regular, sharp up-and-down movements that may be a reflection of seasonality in TNHS. Second, the long-term trend appears negative. The autocorrelation structure of TNHS is shown in Figure 2.10.

We see that the autocorrelations for TNHS do not fall quickly to zero. The autocorrelation coefficients are all significantly different from zero. Thus, we have evidence of a significant trend in TNHS. We also show the ACF for the quarter-to-quarter change of TNHS in Figure 2.10.

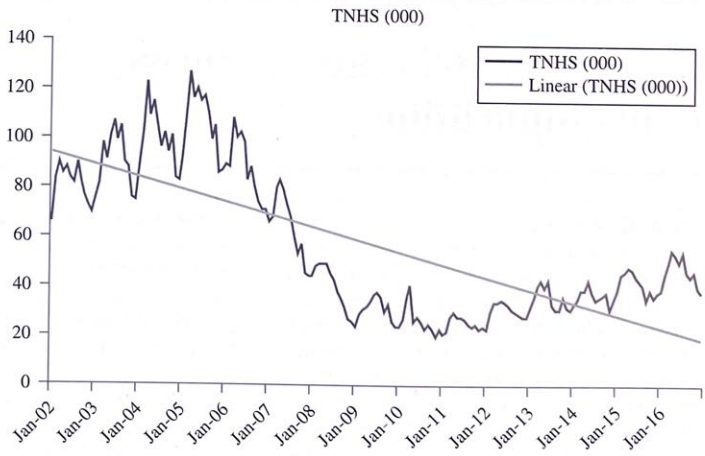
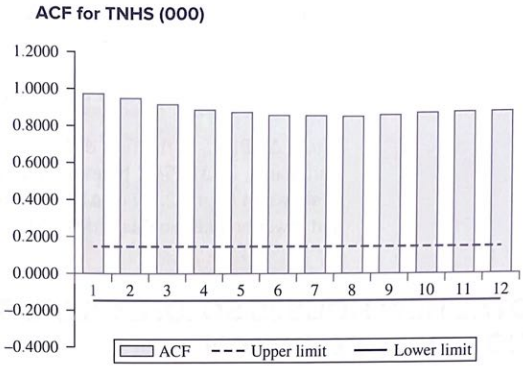


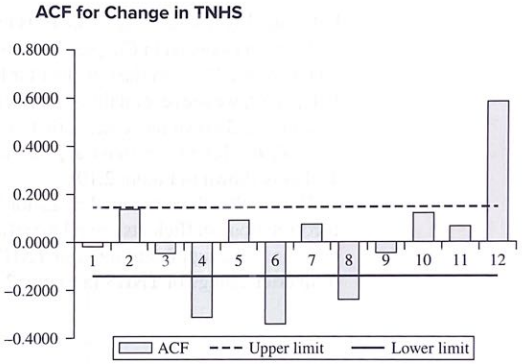
FIGURE 2.9 Total new houses sold. This graph shows total new houses sold (in thousands) by month from January 2002 through December 2016, along with the long-term trend. (c2f9)

FIGURE 2.10 **ACF values for total new houses sold and changes in total new houses sold.** All coefficients are outside the 95 percent confidence band, indicating the positive trend in TNHS. For the change in TNHS, the coefficients fall quickly and are mainly within the 95 percent confidence band, indicating no trend in the month-to-month changes in TNHS. (c2f10)

Obs	ACF
1	.9741
2	.9484
3	.9164
4	.8869
5	.8730
6	.8562
7	.8558
8	.8520
9	.8608
10	.8716
11	.8764
12	.8778



Obs	ACF
1	−.0239
2	.1351
3	−.0506
4	−.3191
5	.0859
6	−.3447
7	.0675
8	−.2432
9	−.0477
10	.1141
11	.0578
12	.5779



Business Forecasting: A Process, Not an Application

1

Charles W. Chase, Jr.

Current literature and experience dictate that the best forecasting system provides easy access, review, and modification of forecast results across all corporate disciplines; provides alternative modeling capabilities (multidimensional); includes the ability to create a knowledge base by which future forecasts can be refined; and provides timely and accurate automated link/feed interfaces with other systems such as I.R.I. (Information Resources Inc.)/Nielsen syndicated databases and the mainframe shipment database. The present industry trend has been redirected away from mainframe systems toward PC-based software applications due to the lack of flexibility associated with mainframe access and reporting. Mainframes are being utilized primarily as storage bins for PC-based systems to extract and store information.

Source: Chase, Charles, Jr. Business Forecasting: A Process, Not an Application, *Journal of Business Forecasting* 11, no. 3 (Fall 1992), pp. 12–13.

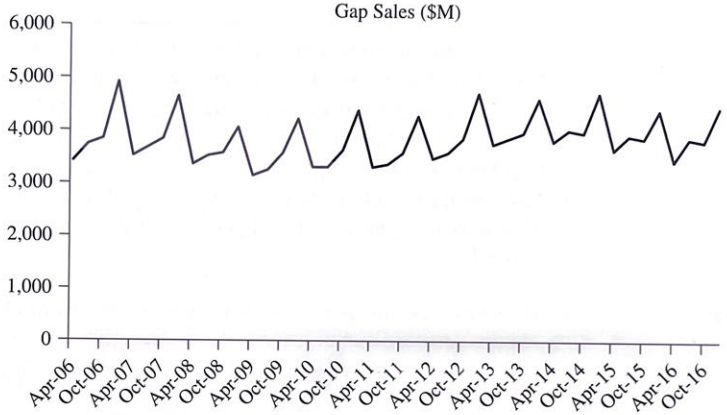
From this exploratory analysis of the total new houses sold, we can conclude that there is trend and seasonality. From Table 2.1, we can, therefore, suggest the following as potential forecasting methods for total houses sold:

Winters' exponential smoothing
Regression trend with seasonality
Causal regression
Time-series decomposition

INTEGRATIVE CASE: THE GAP

The Gap sales year starts in February and ends the following January. This means the first quarter includes February, March, and April. The fourth quarter includes November, December, and January.

Data Analysis of the Gap Sales Data



From this graph, it is clear that The Gap sales are seasonal and essentially stationary. There does not appear to be a cycle. The Gap sales year starts in February and ends the following January. (C2 Gap Sales Data)

Case Questions

1. In 2016, The Gap sales by quarter were as given below:

Quarter	Gap Sales ($M)
2016Q1	3,438
2016Q2	3,851
2016Q3	3,798
2016Q4	4,429

Calculate the mean and standard deviation for this set of quarterly sales. See the following file: C2 Gap Sales Data.

2. The Gap sales on an annual basis are shown in the following table.

Year	Annual Gap Sales ($M)
2006	15,925
2007	15,763
2008	14,526
2009	14,197
2010	14,664
2011	14,549
2012	15,651
2013	16,148
2014	16,435
2015	15,797
2016	15,516

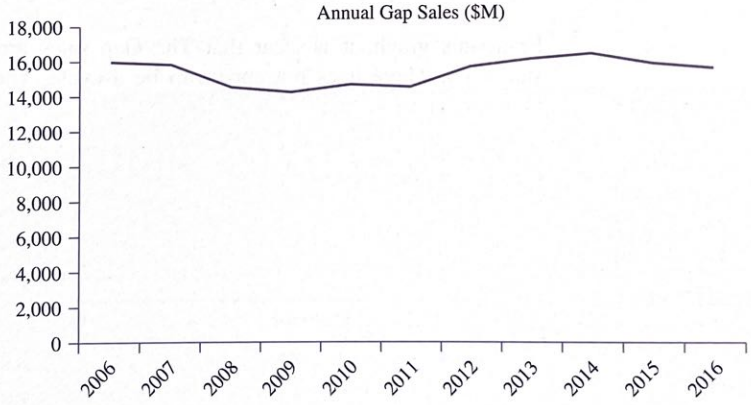
Plot these data in a time-series plot. Based on this graph, what pattern do you see in The Gap's annual sales? See the following file: C2 Gap Sales Data.

3. Using data for 2006Q1 through 2016Q4, construct the correlogram (plot of the autocorrelations) for lags of 1 through 12. (The quarterly data are in the following file: C2 Gap Sales Data.)
4. Based on the plot of The Gap quarterly sales and on what you learned from question 3, what forecasting methods might you suggest if you were to forecast The Gap's quarterly sales?

Solutions to Case Questions

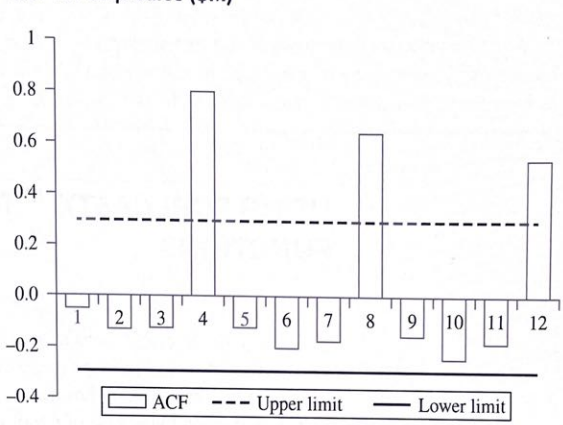
1. The sum of the four quarters is 15,516 which when divided by 4 gives a mean of 3,879. The standard deviation (assuming this is a sample of data) is 410.01. See the following file: C2 Gap Sales Data.

2.



3. As you see from the autocorrelations (ACF) and correlogram below, the autocorrelations do not decline gradually. Thus, we have evidence that there is not a significant trend in The Gap data during the period for which we have data. The higher bars outside the upper 95 percent confidence band suggest seasonality in the fourth quarters.The data are in the following file: C2 Gap Sales Data.

Obs	ACF
1	–.0523
2	–.1305
3	–.1274
4	.7994
5	–.1227
6	–.2048
7	–.1761
8	.6407
9	–.1560
10	–.2462
11	–.1829
12	.5327



4. Based on the plot of The Gap's quarterly sales, as well as the data analysis from question 3, the following forecasting methods might be suggested from the information in Table 2.1:

   Winters' exponential smoothing

   Regression trend with seasonality

   Causal regression

   Time-series decomposition

Comments from the Field 2.1

Anchorage

Economic Development Center Secures Time-Saving Forecasting Accuracy

Anchorage Economic Development Center (AEDC) is a private, nonprofit corporation that has been in operation since 1987 and is seeking to improve the economic conditions in Anchorage by expanding value-added industries, increasing business services, and developing tourism. The AEDC needed to accurately forecast the economic outlook for such industries as mining, government, finance, insurance, real estate, manufacturing, construction, transportation, communications, utilities, trade, and services.

Using historical data from the Alaska Department of Labor, the AEDC had used ratio-to-moving averages classical decomposition formulas in Microsoft Excel to forecast the economic outlook. But this long and fairly complicated process usually took about one month to complete. The results, though complete, were not as accurate as they should be.

The AEDC determined that John Galt Solutions could provide software (ForecastX™ Wizard) that would more accurately—and efficiently—define and forecast the economic conditions in Anchorage. AEDC wanted a solution that would minimize its time formatting and forecasting data and allow more time for analyzing and marketing the results of the forecasts.

The AEDC found ForecastX™ to be an easy-to-integrate tool that required no data preparation. AEDC was also happy to continue using Microsoft Excel and still have the ability to use the advanced forecasting methods. Flawlessly integrated, ForecastX™ Wizard provided the AEDC with Procast (expert selection); the ability to handle unlimited amounts of data; and

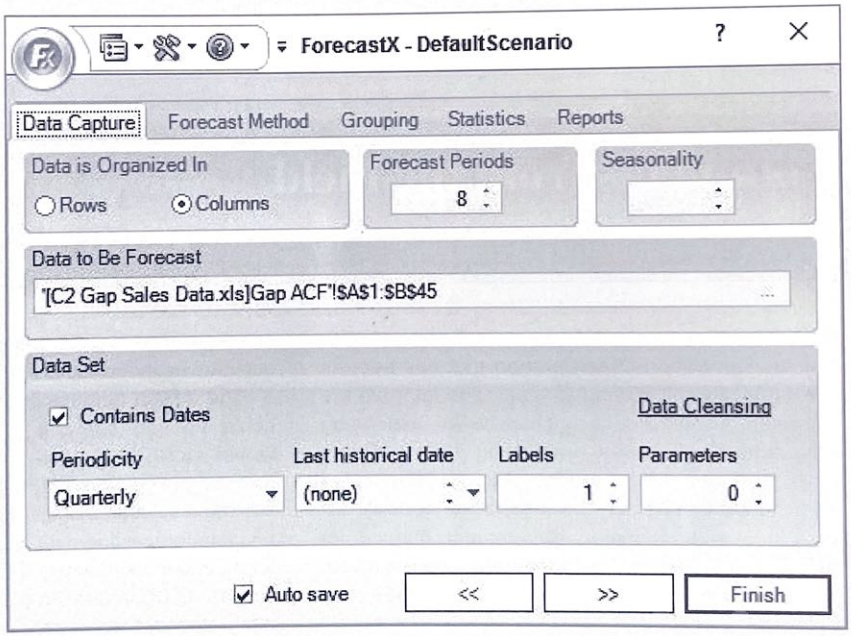
the ability to forecast data on a monthly, quarterly, or yearly basis.

With the advanced features and functionality of ForecastX™ and its ease of use, AEDC was able to cut its forecasting prep time down to one week. More time, therefore, could be spent focusing on evaluating the results of forecasts and bringing more businesses to Anchorage. ForecastX™ Wizard provided AEDC with the tool it needed to more efficiently and accurately complete its forecasts.

USING FORECASTX™ TO FIND AUTOCORRELATION FUNCTIONS

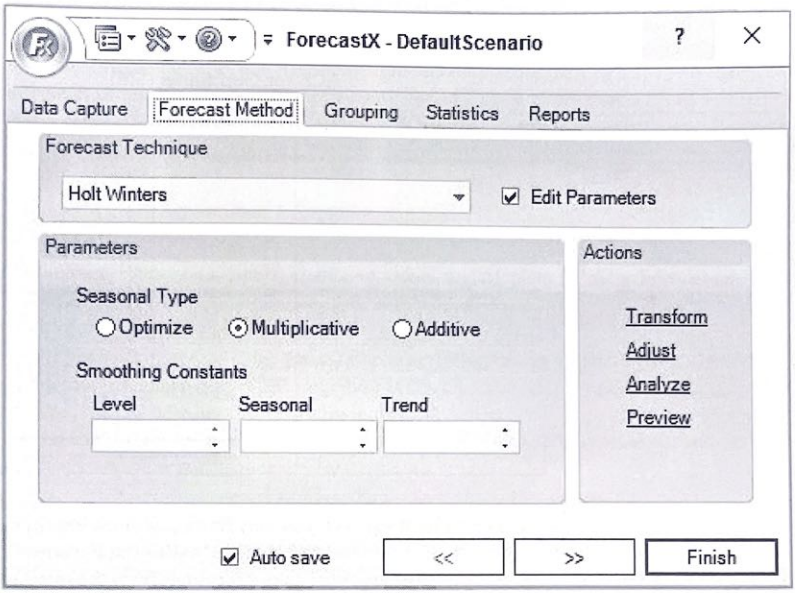
The most difficult calculations in this chapter were the autocorrelation coefficients. These can be calculated easily in the ForecastX™ software that accompanies your text. What follows is a brief discussion of how to use ForecastX™ for this purpose. This also serves as a good introduction to the ease of use of ForecastX™.

First, put your data into an Excel spreadsheet in column format such as the sample of The Gap data shown in C2 Gap Sales Data. Once you have your data in this format, while in Excel highlight the data you want to use and then start ForecastX™. The following dialog box appears:



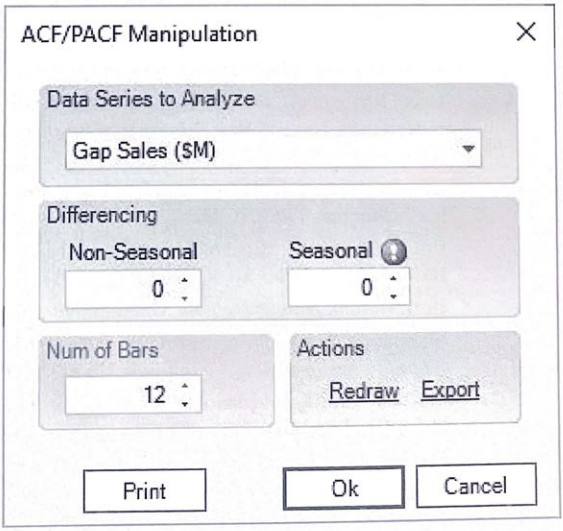
Source: John Galt Solutions

Check the Periodicity box to be sure that it matches the periodicity of your data (**Quarterly** for this example), then click the **Forecast Method** tab at the top and the following screen appears:



Source: John Galt Solutions

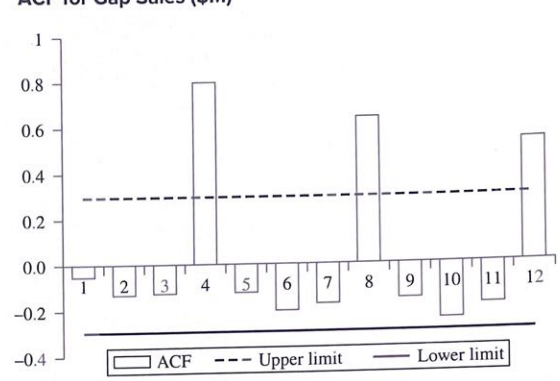
You may get a different forecast method than shown here (Holt Winters) but for now this does not matter. Now click the **Analyze** button in the right side panel and the following screen appears. Click **Export**, and the results will be saved to a new Excel book.



Source: John Galt Solutions

You will have the results shown below (along with some other results) in a new Excel book.

Obs	ACF
1	–.0523
2	–.1305
3	–.1274
4	.7994
5	–.1227
6	–.2048
7	–.1761
8	.6407
9	–.1560
10	–.2462
11	–.1829
12	.5327



Note: **Throughout the text, you may find some situations in which the standard calculations that we show do not match exactly with the ForecastX™ results. This is because they, at times, invoke proprietary alterations from the standard calculations. The results are always very close but sometimes do not match perfectly with "hand" calculations.**

Suggested Readings

Aghazadeh, Seyed-Mahmoud; and Jane B. Romal. "A Directory of 66 Packages for Forecasting and Statistical Analyses." *Journal of Business Forecasting* 11, no. 2 (Summer 1992), pp. 14–20.

"Beyond the Business Cycle?" *The Economist* 353, no. 8142 (October 1999), p. 90.

Chatterjee, Satyajit. "From Cycles to Shocks: Progress in Business-Cycle Theory." *Business Review*, Federal Reserve Bank of Philadelphia (March/April 2000), pp. 27–37.

Chen, Rong, et al. "Forecasting with Stable Seasonal Pattern Models with an Application to Hawaiian Tourism Data." *Journal of Business & Economic Statistics* 17, no. 4 (October 1999), pp. 497–504.

Drumm, William J. "Living with Forecast Error." *Journal of Business Forecasting* 11, no. 2 (Summer 1992), p. 23.

Ermer, Charles M. "Cost of Error Affects the Forecasting Model Selection." *Journal of Business Forecasting* 10, no. 1 (Spring 1991), pp. 10–11.

Huff, Darrell. *How to Lie with Statistics*. New York: W. W. Norton, 1954.

Makridakis, Spyros. "Forecasting: Its Role and Value for Planning and Strategy." *International Journal of Forecasting* 12, no. 4 (December 1996), pp. 513–37.

Mentzer, John T.; and Carol C. Bienstock. *Sales Forecasting Management*. Thousand Oaks, CA: Sage Publications, 1998.

Mentzer, John T.; and Kenneth B. Kahn. "Forecasting Technique Familiarity, Satisfaction, Usage, and Application." *Journal of Forecasting* 14, no. 5 (September 1995), pp. 465–76.

——. "State of Sales Forecasting Systems in Corporate America." *Journal of Business Forecasting* 16, no. 1 (Spring 1997), pp. 6–13.

O'Clock, George; and Priscilla M. O'Clock. "Political Realities of Forecasting." *Journal of Business Forecasting* 8, no. 1 (Spring 1989), pp. 2–6.

Sawhney, Mohanbir S., et al. "A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures." *Marketing Science* 15, no. 2 (1996), pp. 113–31.

Smith, Michael. "Modeling and Short-Term Forecasting of New South Wales Electricity System Load." *Journal of Business & Economic Statistics* 18, no. 4 (October 2000), pp. 465–78.

Tufte, Edward R. *Envisioning Information.* Cheshire, CT: Graphics Press, 1990.

——. *The Visual Display of Quantitative Information.* Cheshire, CT: Graphics Press, 1983.

Winklhofer, Heidi; Adamantios Diamantopoulos; and Stephen F. Witt. "Forecasting Practice: A Review of the Empirical Literature and an Agenda for Future Research." *International Journal of Forecasting* 12, no. 2 (June 1996), pp. 193–221.

xercises

1. The mean volume of sales for a sample of 100 sales representatives is $25,350 per month. The sample standard deviation is $7,490. The vice president for sales would like to know whether this result is significantly different from $24,000 at a 95 percent confidence level. Set up the appropriate null and alternative hypotheses, and perform the appropriate statistical test.
2. Larry Bomser has been asked to evaluate sizes of tire inventories for retail outlets of a major tire manufacturer. From a sample of 120 stores, he has found a mean of 310 tires. The industry average is 325. If the standard deviation for the sample was 72, would you say that the inventory level maintained by this manufacturer is significantly different from the industry norm? Explain why. (Use a 95 percent confidence level.)
3. Twenty graduate students in business were asked how many credit hours they were taking in the current quarter. Their responses are shown as follows (c2p3):

Student Number	Credit Hours	Student Number	Credit Hours	Student Number	Credit Hours
1	2	8	8	15	10
2	7	9	12	16	6
3	9	10	11	17	9
4	9	11	6	18	6
5	8	12	5	19	9
6	11	13	9	20	10
7	6	14	13		

   a. Determine the mean, median, and mode for this sample of data. Write a sentence explaining what each means.

   b. It has been suggested that graduate students in business take fewer credits per quarter than the typical graduate student at this university. The mean for all graduate students is 9.1 credit hours per quarter, and the data are normally distributed. Set up the appropriate null and alternative hypotheses, and determine whether the null hypothesis can be rejected at a 95 percent confidence level.

4. Arbon Computer Corporation (ACC) produces a popular PC clone. The sales manager for ACC has recently read a report that indicated that sales per sales representative for other producers are normally distributed with a mean of $255,000. She is interested in knowing whether her sales staff is comparable. She picked a random sample of 16 salespeople and obtained the following results (c2p4):

Person	Sales	Person	Sales
1	177,406	9	110,027
2	339,753	10	182,577
3	310,170	11	177,707
4	175,520	12	154,096
5	293,332	13	236,083
6	323,175	14	301,051
7	144,031	15	158,792
8	279,670	16	140,891

At a 5 percent significance level, can you reject the null hypothesis that ACC's mean sales per salesperson was $255,000? Draw a diagram that illustrates your answer.

5. Assume that the weights of college football players are normally distributed with a mean of 205 pounds and a standard deviation of 30.
   a. What percentage of players would have weights greater than 205 pounds?
   b. What percentage of players would weigh less than 250 pounds?
   c. Ninety percentage of players would weigh more than what number of pounds?
   d. What percentage of players would weigh between 180 and 230 pounds?
6. Mutual Savings Bank of Appleton has done a market research survey in which people were asked to rate their image of the bank on a scale of 1 to 10, with 10 being the most favorable. The mean response for the sample of 400 people was 7.25, with a standard deviation of 2.51. On this same question, a state association of mutual savings banks has found a mean of 7.01.
   a. Clara Weston, marketing director for the bank, would like to test to see whether the rating for her bank is significantly greater than the norm of 7.01. Perform the appropriate hypothesis test for a 95 percent confidence level.
   b. Draw a diagram to illustrate your result.
   c. How would your result be affected if the sample size had been 100 rather than 400, with everything else being the same?
7. In a sample of 25 classes, the following numbers of students were observed (c2p7):

Class	Number of students	Class	Number of students
1	40	14	37
2	50	15	35
3	42	16	44
4	20	17	10
5	29	18	40
6	39	19	36
7	49	20	20
8	46	21	20
9	52	22	29
10	45	23	58
11	51	24	51
12	64	25	54
13	43		

   a. Calculate the mean, median, standard deviation, variance, and range for this sample.
   b. What is the standard error of the mean based on this information?

c. What would be the best point estimate for the population class size?

d. What is the 95 percent confidence interval for class size? What is the 90 percent confidence interval? Does the difference between these two make sense?

8. CoastCo Insurance, Inc., is interested in forecasting annual larceny thefts in the United States using the following data (c2p8):

Year	Larceny Thefts	Year	Larceny Thefts
1972	4,151	1984	6,592
1973	4,348	1985	6,926
1974	5,263	1986	7,257
1975	5,978	1987	7,500
1976	6,271	1988	7,706
1977	5,906	1989	7,872
1978	5,983	1990	7,946
1979	6,578	1991	8,142
1980	7,137	1992	7,915
1981	7,194	1993	7,821
1982	7,143	1994	7,876
1983	6,713		

a. Prepare a time-series plot of these data. On the basis of this graph, do you think there is a trend in the data? Explain.

b. Look at the autocorrelation structure of larceny thefts for lags of 1, 2, 3, 4, and 5. Do the autocorrelation coefficients fall quickly toward zero? Demonstrate that the critical value for r_k is 0.417. Explain what these results tell you about a trend in the data.

c. On the basis of what is found in parts *a* and *b*, suggest a forecasting method from Table 2.1 that you think might be appropriate for this series.

9. Use exploratory data analysis to determine whether there is a trend and/or seasonality in mobile home shipments (MHS). The data by quarter are shown in the following table (c2p9):

Period	MHS	Period	MHS	Period	MHS	Period	MHS
Mar-81	54.9	Dec-84	66.2	Sep-88	59.2	Jun-92	52.8
Jun-81	70.1	Mar-85	62.3	Dec-88	51.6	Sep-92	57
Sep-81	65.8	Jun-85	79.3	Mar-89	48.1	Dec-92	57.6
Dec-81	50.2	Sep-85	76.5	Jun-89	55.1	Mar-93	56.4
Mar-82	53.3	Dec-85	65.5	Sep-89	50.3	Jun-93	64.3
Jun-82	67.9	Mar-86	58.1	Dec-89	44.5	Sep-93	67.1
Sep-82	63.1	Jun-86	66.8	Mar-90	43.3	Dec-93	66.4
Dec-82	55.3	Sep-86	63.4	Jun-90	51.7	Mar-94	69.1
Mar-83	63.3	Dec-86	56.1	Sep-90	50.5	Jun-94	78.7
Jun-83	81.5	Mar-87	51.9	Dec-90	42.6	Sep-94	78.7
Sep-83	81.7	Jun-87	62.8	Mar-91	35.4	Dec-94	77.5
Dec-83	69.2	Sep-87	64.7	Jun-91	47.4	Mar-95	79.2
Mar-84	67.8	Dec-87	53.5	Sep-91	47.2	Jun-95	86.8
Jun-84	82.7	Mar-88	47	Dec-91	40.9	Sep-95	87.6
Sep-84	79	Jun-88	60.5	Mar-92	43	Dec-95	86.4

On the basis of your analysis, do you think there is a significant trend in MHS? Is there seasonality? What forecasting methods might be appropriate for MHS according to the guidelines in Table 2.1?

10. Home sales are often considered an important determinant of the future health of the economy. Thus, there is widespread interest in being able to forecast home sales (HS). Quarterly data for HS are shown in the following table in thousands of units (c2p10):

Date	Home sales (000) per Quarter	Date	Home sales (000) per Quarter	Date	Home sales (000) per Quarter
Mar-89	161	Jun-95	185	Sep-01	216
Jun-89	179	Sep-95	181	Dec-01	199
Sep-89	172	Dec-95	145	Mar-02	240
Dec-89	138	Mar-96	192	Jun-02	258
Mar-90	153	Jun-96	204	Sep-02	254
Jun-90	152	Sep-96	201	Dec-02	220
Sep-90	130	Dec-96	161	Mar-03	256
Dec-90	100	Mar-97	211	Jun-03	299
Mar-91	121	Jun-97	212	Sep-03	294
Jun-91	144	Sep-97	208	Dec-03	239
Sep-91	126	Dec-97	174	Mar-04	314
Dec-91	116	Mar-98	220	Jun-04	329
Mar-92	159	Jun-98	247	Sep-04	292
Jun-92	158	Sep-98	218	Dec-04	268
Sep-92	159	Dec-98	200	Mar-05	328
Dec-92	132	Mar-99	227	Jun-05	351
Mar-93	154	Jun-99	248	Sep-05	326
Jun-93	183	Sep-99	221	Dec-05	278
Sep-93	169	Dec-99	185	Mar-06	285
Dec-93	160	Mar-00	233	Jun-06	300
Mar-94	178	Jun-00	226	Sep-06	251
Jun-94	185	Sep-00	219	Dec-06	216
Sep-94	165	Dec-00	199	Mar-07	214
Dec-94	142	Mar-01	251	Jun-07	240
Mar-95	154	Jun-01	243		

a. Prepare a time-series plot of THS. Describe what you see in this plot in terms of trend and seasonality.

b. Calculate and plot the first 12 autocorrelation coefficients for HS. What does this autocorrelation structure suggest about the trend?

11. Exercise 12 of Chapter 1 includes data on the Japanese exchange rate (EXRJ) by month. On the basis of a time-series plot of these data and the autocorrelation structure of EXRJ, would you say the data are stationary? Explain your answer. (c2p11)

Period	EXRJ	Period	EXRJ
1	127.36	13	144.98
2	127.74	14	145.69
3	130.55	15	153.31
4	132.04	16	158.46
5	137.86	17	154.04
6	143.98	18	153.7
7	140.42	19	149.04
8	141.49	20	147.46
9	145.07	21	138.44
10	142.21	22	129.59
11	143.53	23	129.22
12	143.69	24	133.89