Part V

The Part of Tens

Enjoy an additional Business Statistic Part of Tens chapter online at `www.dummies.com/extras/businessstatistics`.

In this part…

- ✓ See how statistical tests are based on the assumption of normality, and review several techniques available for testing whether a particular set of data is normally distributed.
- ✓ Check out several types of problems that may arise when the assumptions of regression analysis are not met; two problems that can plague simple regression analysis are *autocorrelation* and *heteroscedasticity*.

Chapter 18

Ten Common Errors That Arise in Statistical Analysis

In This Chapter

- Understanding logical fallacies that may arise in statistical analysis
- Avoiding drawing incorrect conclusions from statistical results
- Understanding the types of errors that can result in regression analysis
- Understanding forecasting errors
- Realizing how information may be presented incorrectly

In the *For Dummies* Part of Tens fashion, this chapter discusses ten ways people may draw incorrect conclusions from statistical tests. These erroneous conclusions can result from several sources, including incorrect assumptions, misunderstanding the meaning of a statistical test, use of inappropriate data, and measurement error.

Any one of these mistakes can lead to erroneous conclusions being drawn, no matter how sophisticated the techniques being used. Part of the art of statistics is knowing which techniques to use under different circumstances and how to correctly interpret them. The following sections discuss different types of errors that may result from the incorrect application of statistical techniques.

Designing Misleading Graphs

Graphs may give a misleading picture of a sample or population if they're not well designed. For example, if you use scales on a graph that are substantially different from the values in the data you're analyzing, you may end up with a highly distorted view of the data.

Figures 18-1 and 18-2 represent the same data with two different histograms (see Chapter 2 for an overview of histograms).

In this example, the data consist of the distribution of a bank's branches scattered throughout the four regions of the United States — North, South, East, and West.

Region	*Branches*
North	1,213
South	1,415
East	1,199
West	1,098

In Figure 18-1, the values on the vertical axis are separated by only 20 branches.

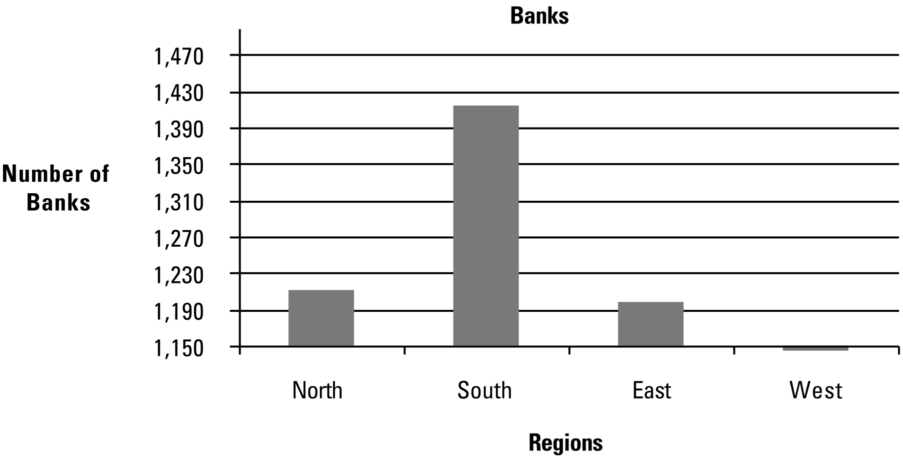

Figure 18-1: Distribution of bank branches by geographical region.

With such closely spaced values on the vertical axis, the differences between the number of branches in each region appear to be very large. But, in fact, the difference between the largest number and the smallest number is only 317 (about 29 percent).

In Figure 18-2, the spacing of the values on the vertical axis is much wider, separated by 500 branches, making it appear that the differences between the numbers of branches are quite minimal.

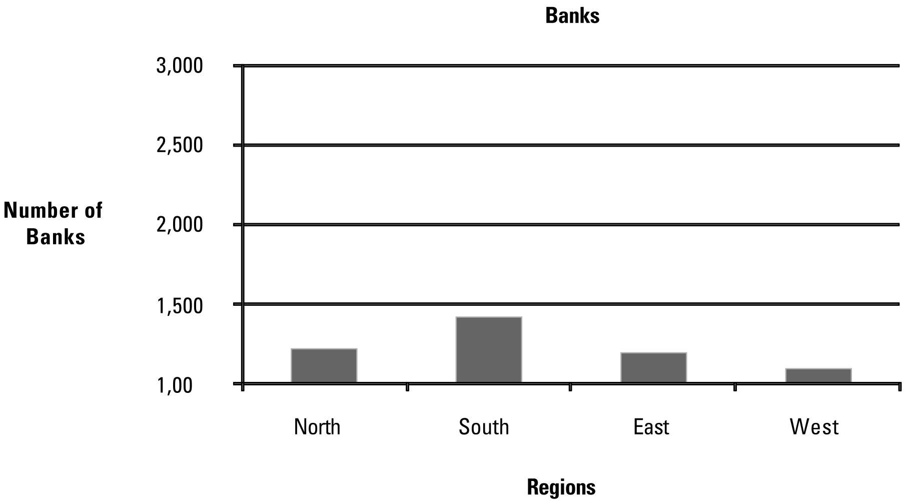

Figure 18-2: Another look at the distribution of bank branches by geographical region.

These figures show how easy it is to give a distorted view of data through poor design.

Drawing the Wrong Conclusion from a Confidence Interval

When constructing a confidence interval, you can easily draw the wrong conclusion from the results. (Confidence intervals are covered in Chapter 11.) For example, suppose that a university constructs a 95 percent confidence interval for the mean GPA of its students. The sample mean is estimated to be 3.10; the 95 percent confidence interval is (2.95, 3.25).

It's tempting to conclude that the probability of the population mean being in the interval (2.95, 3.25) is 95 percent. Instead, this result indicates that for every confidence interval that's constructed from this population, in 95 cases out of 100, the confidence interval will contain the true population mean.

Misinterpreting the Results of a Hypothesis Test

One potential problem that may arise in hypothesis testing is confusing what it means when the null hypothesis isn't rejected. It's important to distinguish between accepting the null hypothesis and failing to reject the null hypothesis.

For example, suppose that a jury trial is in progress. For this hypothesis test, the following null and alternative hypotheses are used:

- Null hypothesis (H_0): The defendant is innocent.
- Alternative hypothesis (H_1): The defendant is guilty.

If the null hypothesis is rejected, the defendant is guilty. If the null hypothesis isn't rejected, the defendant isn't necessarily innocent. There's simply insufficient evidence to show that he's guilty. There's a world of difference between being "innocent" and "not guilty!"

The proper procedure in a hypothesis test is to conclude that a null hypothesis fails to be rejected unless strong contrary evidence exists against it. The conclusion should never be that the null hypothesis is accepted.

Placing Too Much Confidence in the Coefficient of Determination (R^2)

With regression analysis, researchers sometimes use the coefficient of determination to figure out whether one model "fits" the data better than another. The coefficient of determination assumes a value between 0 and 1; the closer it is to 1, presumably the better the regression model explains the relationship between *X* and *Y*. One of the drawbacks to the coefficient of determination is that it can be very close to 1 even for a model that makes no economic sense, such as a regression between two unrelated variables.

Another issue that arises with the coefficient of determination is that it automatically increases when new independent variables are added to a regression equation, even if the variables don't contribute any additional explanatory power to the regression. For this reason, the adjusted coefficient of determination is the preferred measure with multiple regression analysis because it increases only when newly added independent variables add at least some explanatory power.

Assuming Normality

Many statistical tests are based on the assumption of normality. For example, residuals are assumed to be normally distributed in regression analysis, enabling confidence intervals to be constructed for the slope coefficients.

For example, it's often assumed that the returns to stocks are normally distributed. In fact, although they're close to being normally distributed, they exhibit a property known as *fat tails,* where the actual probability of extreme outcomes (large positive returns and large negative returns) is greater than under the normal distribution. The assumption of normality causes investors to underestimate the true riskiness of their portfolios.

Several techniques are available for testing whether a particular set of data is normally distributed. For example, a Q-Q plot can be used to visually inspect data for normality. (You can read more about QQ plots at `http://en.wikipedia.org/wiki/Q-Q_plot`.)

A formal hypothesis test of normality is available; it's known as the Jarque-Bera test. (You can read more about the Jarque-Bera test at `http://en.wikipedia.org/wiki/Jarque%E2%80%93Bera_test`.)

These types of techniques should be used before jumping to any conclusions about normality.

Thinking Correlation Implies Causality

One common error in statistical analysis is to assume that if two variables are correlated, one *causes* the other. Correlation simply indicates the tendency of two variables to move in the same or opposite directions. For example, new car sales tend to rise at the same time as new home sales, but no one would suggest that new home sales *cause* new car sales. (Equivalently, no one would suggest that new car sales are *caused* by new home sales.) These variables are positively correlated because they're both directly influenced by the economy. During an expansion, both new car sales and new home sales rise; during a recession, both fall.

One particularly well-known example of the dangers of assuming that correlation implies causality comes from the 19th century British economist William Stanley Jevons. Jevons was interested in applying statistical methods to the measurement of business cycles. He noticed that the business cycle had a tendency to follow changes in sunspot activity. Sunspots went through a cycle that lasted for about 11 years, while business cycles tended to last for just under 11 years. From his studies, Jevons concluded that the sunspots were actually responsible for the business cycle. (It's not as crazy as it sounds; sunspots can lead to changes in weather patterns, which would have a huge influence on the business cycles of a primarily agriculture-based economy. In spite of this, sunspots do *not* directly cause changes in the business cycle.)

Drawing Conclusions from a Regression Equation when the Data do not Follow the Assumptions

Several types of problems may arise when the assumptions of regression analysis are not met. (Simple regression analysis is covered in Chapter 15; multiple regression analysis is covered in Chapter 16.) Two problems that can plague simple regression analysis are known as *autocorrelation* and *heteroscedasticity*.

Autocorrelation occurs when the error terms are correlated with each other (they are related to each other). It violates the assumption of independence. Two independent variables have a correlation of 0 between them. Autocorrelated error terms can cause understating the standard errors of the regression coefficients, thus increasing the risk that coefficients are incorrectly found to be statistically significant (for example, different from zero).

Heteroscedasticity occurs when the error terms don't have a constant variance. This problem can cause understating the standard errors of the regression coefficients, increasing the risk that coefficients are incorrectly found to be statistically significant (for example, different from zero).

When these problems are present, it is important to correct for them; otherwise, all results will be deceptive.

Including Correlated Variables in a Multiple Regression Equation

One potential difficulty with multiple regression analysis is *multicollinearity*. Multicollinearity occurs when two or more of the independent variables are highly correlated with each other, causing the correlated variables to have large standard errors, so they appear to be statistically insignificant even if they're not. (In other words, there's a risk that independent variables will be removed from the regression equation that should be included.)

Multicollinearity is unique to multiple regression because it has multiple independent variables (simple regression has only one independent variable so that multicollinearity cannot occur).

A statistic known as the variance inflation factor (VIF) may be used to check for multicollinearity. As a rule of thumb, if the VIF is 10 or more, it's a sign that multicollinearity is present. (You can find more information about the variance inflation factor at `http://en.wikipedia.org/wiki/Variance_inflation_factor`.) If multicollinearity is present, one of the highly correlated variables should be removed from the regression equation.

Placing Too Much Confidence in Forecasts

Many techniques are used to forecast future values of economic variables, such as stock prices, GDP growth, corporate sales, the demand for new products, and so on. Many of these techniques are highly sophisticated, which may give the false impression that they're extremely accurate. One major difficulty with forecasting techniques is that they're based on historical data that may not be repeated in the future. For example, if an economist is attempting to forecast future interest rates, his results don't capture any structural changes that occur in the economy during the forecast period, such as the selection of a new chairman of the Federal Reserve Board. In this case, future interest rates are unlikely to behave in exactly the same way that they have in the past, and the results of the forecast are inaccurate.

Two types of errors that may arise in forecasting are bias error and random error. *Bias error* occurs when a forecast is consistently greater than or less than actual values of a variable. *Random error* refers to unpredictable factors that can distort the results of a factor. These include earthquakes, strikes, sudden increases in oil prices, political turmoil, and so on.

With so much uncertainty surrounding forecasts, it would be a mistake to assume a high degree of accuracy.

Using the Wrong Distribution

In many situations, a variable is assumed to follow a specific probability distribution. For example, a computer chip manufacturer may assume that the number of defective chips produced by a specific process follows the binomial distribution. (The binomial distribution is covered in Chapter 8.) The binomial distribution is based on several assumptions, one of which is that the trials are independent of each other. Suppose that in this process, one defective chip is highly likely to be followed by another defective chip (for example, repairs to the process are needed). In this case, the trials (chips) aren't truly independent of each other. As a result, any conclusions drawn about the distribution of defective chips are likely inaccurate. The manufacturer needs to find another distribution that more accurately reflects the distribution of the chips.

Chapter 19

Ten Key Categories of Formulas for Business Statistics

In This Chapter

- Keeping the most important statistical concepts fresh in your memory
- Seeing how key statistical formulas are related

This chapter provides a brief overview of many key formulas encountered in the text. This provides a handy reference guide so that you can quickly find the formulas that you need without having to search through the entire book.

Summary Measures of a Population or a Sample

Summary measures are used to describe key properties of a sample or a population. These measures can be classified as:

- **Measures of central tendency** identify the *center* of a data set. Three of the most widely used measures of central tendency are the mean, median, and mode.
    - The *mean* is another word for average.
    - The *median* is a value that divides a sample or a population in half: Half of the elements in the data are below the median, and half of the elements in the data are above the median.
    - The *mode* is the most frequently occurring value in a sample or a population.

- **Measures of dispersion** are used to measure how spread out, or disperse, are the values of a sample or a population. Some of the most important measures of dispersion are the variance, standard deviation, percentiles, quartiles, and the interquartile range (IQR).
    - **Variance:** The variance is calculated as the size of the average *squared* difference between the elements of a data set (a sample or a population) and the mean of the data set. The greater is the variance, the further the elements of the data set tend to be from the mean.
    - **Standard deviation:** The square root of the variance. The standard deviation is more convenient to use than the variance due to the units in which these measures are calculated. As an example, if a sample consists of dollar prices, the sample standard deviation is measured in dollars, while the sample variance is measured in dollars *squared*, which is difficult to make sense of.
    - **Percentiles:** Percentiles split a data set into 100 equal parts, each consisting of 1 percent of the values in the data set. For example, the 80th percentile represents the value in a sample or a population where 20 percent of the observations are above this value, and 80 percent are below this value.
    - **Quartiles:** Special types of percentiles, where the first quartile (Q_1) is the 25th percentile, the second quartile (Q_2) is the 50th percentile, and the third quartile (Q_3) is the 75th percentile.
    - **Interquartile range (IQR):** The difference between the third and first quartile.
- **Measures of association** provide a measure of how closely two samples or populations are related to each other. The two most important measures of association are:
    - **Covariance** is a measure of the tendency for two variables to move in the same direction or in opposite directions. If two variables increase or decrease under the same circumstances, the covariance between them is positive. If two variables move in opposite directions, the covariance between them is negative. If two variables are unrelated to each other, the covariance between them is zero (or very close to zero).
    - **Correlation** is closely related to covariance; it has more convenient properties than covariance. For example, correlation always assumes a value between -1 and 1, whereas covariance has no lower or upper limits. As a result, it is easier to tell if the relationship between two variables is very strong or very weak with correlation than with covariance.

Probability

You use probability theory to model a large number of events in business applications. Probability theory is based on *set algebra,* and the important rules are

- **Addition rule:** The formula for the Addition rule is:

  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

The addition rule is designed to compute the probability of a *union* of two sets. In general, the union of sets A and B contains all the elements that are in set A, set B or both.

- **Multiplication rule:** The Multiplication rule has two forms:

  $P(A \cap B) = P(A \mid B)P(B)$

  $P(A \cap B) = P(B \mid A)P(A)$

The multiplication rule is designed to compute the probability of the *intersection* of two sets. In general, the intersection of sets A and B contains all the elements that are in *both* set A and set B.

- **Complement rule:** The Complement rule has two forms:

  $P\left(A^C\right) = 1 - P(A)$

  $P(A) = 1 - P\left(A^C\right)$

The complement rule tells you the probability of all elements that are *not* in a set. For example, suppose set A contains all the black cards in a standard deck; the complement of A (written as A^C) is a set containing all the red cards. The probability of A^C can be computed with the complement rule.

Discrete Probability Distributions

A discrete probability distribution occurs where only a finite number of different outcomes may occur. The properties of a probability distribution may be summarized by a set of *moments*. Moments are numerical values that describe key properties of a probability distribution. Some of the most important are as follows:

- The **expected value** is the first moment of a probability distribution. You compute it as

  $$E(X)=\sum_{i=1}^{n} X_i P(X_i)$$

The expected value tells you the average value of X.

- The **variance** is the second moment of a probability distribution. You compute it as

  $$\sigma^2=\sum_{i=1}^{n}\left[X_i-E(X)\right]^2 P(X_i)$$

  σ^2 represents the variance of *X*.

The variance tells you how much the different possible values of X are scattered around the expected value.

- The **standard deviation** isn't a separate moment; it's the square root of the variance. The formula is

  $$\sigma=\sqrt{\sum_{i=1}^{n}\left[X_i-E(X)\right]^2 P(X_i)}$$

The standard deviation is preferred to the variance since the variance is measured in squared units, which are difficult to interpret.

Following are three of the most widely used discrete probability distributions in business applications:

- **Binomial distribution:** The binomial distribution is defined for a random process consisting of a series of trials in which only two different outcomes can occur on each trial. It enables you to determine the probability of a specified number of events occurring during a series of trials.

  $$P(X=x)=\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

- **Geometric distribution:** The geometric distribution is related to the binomial distribution; it is used to determine how many trials are needed before a specified event occurs.

  $$P(X=x)=(1-p)^{x-1}p$$

- **Poisson distribution:** The Poisson distribution is used to determine the probability that a specified number of events will occur during an interval of time.

  $$P(X=x)=e^{-\lambda}\frac{\lambda^x}{x!}$$

Continuous Probability Distributions

A continuous probability distribution is defined for an infinite number of possible values. The uniform distribution and the normal distribution are two of the most widely used continuous probability distributions in business applications.

- The **uniform distribution** is defined over an interval (*a, b*); in other words, all values between a and b. For example, the uniform distribution may be defined over the interval (1, 10). This means that the distribution is defined for all values between 1 and 10. You can compute probabilities for the uniform distribution with the following equation, known as a *probability density function (pdf)*:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{b} \leq \text{x} \leq \text{a} \\ 0 & \text{otherwise} \end{cases}$$

- The **normal distribution** is by far the most important continuous probability distribution for business applications. You can get probabilities for this distribution from normal tables, specialized calculators, and spreadsheet programs. The normal distribution is defined by the following probability density function:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-0.5\left(\frac{x-\mu}{\sigma}\right)^2}$$

The normal distribution is important because many business situations may be accurately modeled with the normal distribution. For example, returns to stock prices are often assumed to follow the normal distribution.

Sampling Distributions

A sampling distribution is a special type of probability distribution defined for *sample statistics.* A sample statistic is a measure that describes the properties of a sample. Three of the most important sample statistics are the sample mean ($\bar{X}$), sample variance (s^2), and sample standard deviation (s). For more details about sampling distributions, see Chapter 10.

Based on a key result in statistics known as the *central limit theorem,* the sampling distribution of the sample mean is *normal* as long as the underlying population is normal or if you choose sample sizes of at least 30 from the population. To compute a probability for the sample mean, convert it into a standard normal random variable as follows:

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}}$$

- $\overline{X}$ is the sample mean
- $\mu_{\overline{X}}$ is the mean of the sampling distribution of $\overline{X}$
- $\sigma_{\overline{X}}$ is the standard deviation (also known as the *standard error)* of the sampling distribution of $\overline{X}$

Confidence Intervals for the Population Mean

A *confidence interval* is a range of numbers that is expected to contain the true value of the population mean with a specified probability.

The formula you use to compute a confidence interval for the population mean depends on whether you know the population standard deviation (σ).

If you know the population standard deviation, the appropriate formula is

$$\overline{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$\overline{X}$ is the sample mean

$Z_{\alpha/2}$ is a quantile which represents the location of the right tail under the standard normal distribution with area $\alpha/2$

σ is the population standard deviation

n is the sample size

α is the level of significance

If you don't know the population standard deviation, you replace the population standard deviation with the sample standard deviation:

$$\overline{X} \pm t_{\alpha/2}^{n-1} \frac{s}{\sqrt{n}}$$

- $t_{\alpha/2}^{n-1}$ is a quantile (critical value) which represents the location of the right tail of the t-distribution with n-1 degrees of freedom with an area of $\alpha/2$
- s is the sample standard deviation

Testing Hypotheses about Population Means

Testing hypotheses about population means is a multi-step process, consisting of the null and alternative hypotheses, the level of significance, test statistic, critical value(s), and decision. (I walk you through all the steps of hypothesis testing in Chapter 12.)

You write the null hypothesis for testing the value of a single population mean as

$$H_0: \mu = \mu_0$$

where H_0 stands for the null hypotheses, μ is the true population mean and μ_0 is the hypothesized value of the population, or the value that you *think* is true.

The alternative hypothesis can assume one of three forms:

$H_1: \mu > \mu_0$ (known as a "right-tailed" test)

$H_1: \mu < \mu_{0\,(}$known as a "left-tailed" test)

$H_1: \mu \neq \mu_0$ (known as a "two-tailed" test)

To test a hypothesis, you must specify a *level of significance* — the probability of rejecting the null hypothesis when it's actually true.

When you're testing hypotheses about the population mean, the test statistic and the critical value (or values) depend on the size of the sample drawn from the population and whether you know the population standard deviation.

- For a *small* sample (less than 30), the appropriate test statistic is

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

- $\bar{X}$ is the sample mean
- μ_0 is the hypothesized value of the population mean
- s is the sample standard deviation
- n is the sample size
- For a *large* sample (30 or more) when you know the population standard deviation (σ), the appropriate test statistic is

  $$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

- For a *large* sample when you don't know the population standard deviation, use the sample standard deviation (s) instead:

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

For small samples (the sample size is less than 30), the critical values are drawn from the t-distribution with $n - 1$ degrees of freedom. For large samples, the critical values are drawn from the standard normal distribution.

To test hypotheses about the equality of two population means, the test statistic and critical values are different, but the basic process remains unchanged. In this case, though, you write the null hypothesis as H_0: $\mu_1 = \mu_2$, where μ_1 is the mean of population 1, and μ_2 is the mean of population 2.

- For independent samples with equal population variances, the test statistic is

  $$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

  s^2_p is the estimated common "pooled" variance of the two populations — which you calculate with this formula:

  $$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

  The critical values of independent samples with equal population variances are based on the t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

- If the independent samples are drawn from populations that don't have the same variance, the test statistic depends on the sizes of the two samples. If at least one sample is small, the test statistic becomes

  $$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

  Here, the critical values are also drawn from the t-distribution, but the degrees of freedom calculation is much more complex:

  $$df = \frac{\left[\left(s_1^2 / n_1\right) + \left(s_2^2 / n_2\right)\right]^2}{\frac{\left(s_1^2 / n_1\right)^2}{(n_1 - 1)} + \frac{\left(s_2^2 / n_2\right)^2}{(n_2 - 1)}}$$

- If the independent samples are drawn from populations that don't have the same variance and both samples are large, the test statistic becomes

  $$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

  In this case, the critical values are drawn from the standard normal distribution.

- If the two samples aren't independent, they're known as *paired samples.* The test statistic is then based on the differences between the samples:

  $$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

  $\bar{d}$ is the average difference between paired samples, and s_d is the standard deviation of the sample differences.

  In this case, the critical values are taken from the t-distribution with $n - 1$ degree of freedom.

Testing Hypotheses about Population Variances

Testing hypotheses about population variances follows the same six-step procedure as testing hypotheses about population means (see previous section and Chapter 12 for details).

For testing hypotheses about the variance of a single population, the appropriate test statistic is

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

- n is the sample size)
- s^2 is the sample variance
- σ_0^2 is the hypothesized value of the population variance

The critical values are drawn from the chi-square distribution with $n - 1$ degree of freedom.

For testing hypotheses about the equality of variances of two populations, the appropriate test statistic is

$$F = \frac{s_1^2}{s_2^2}$$

s_1^2 is the variance of the sample drawn from population 1; s_2^2 is the variance of the sample drawn from population 2. The populations are assigned a number of 1 or 2 in such a way as to ensure that s_1^2 is greater than or equal to s_2^2.

The critical values are drawn from the F-distribution, which has two different types of degrees of freedom: numerator and denominator. In this case, the numerator degrees of freedom equal $n_1 - 1$, and the denominator degrees of freedom equal $n_2 - 1$.

Using Regression Analysis

You use regression analysis to estimate the relationship between a dependent variable (*Y*) and one or more independent variables (*X*s).

- Use **simple regression analysis** to estimate the relationship between a dependent variable (*Y*) and one independent variable (*X*).
- Use **multiple regression analysis** to estimate the relationship between a dependent variable (*Y*) and two or more independent variables (*X*s).

Several tests allow you to validate the results of a regression equation. For example, if the coefficient of an independent variable equals 0, the variable doesn't belong in the regression. A hypothesis test helps you determine whether this coefficient equals 0. In the case of multiple regression, it may make sense to test the hypothesis that the slope coefficients all equal 0; if this hypothesis can't be rejected, then the regression equation is completely invalid.

It's also important to ensure that the underlying assumptions of regression analysis aren't being violated. Three potential problems can result if the assumptions aren't true:

- **Autocorrelation** indicates that the error terms aren't independent of each other.
- **Heteroscedasticity** indicates that the error terms don't have a common variance.
- **Multicollinearity** indicates that two or more of the independent variables are highly correlated with each other. (This can only affect the results with multiple regression.)

Forecasting Techniques

There are many different forecasting techniques that can be used to predict the future values of variables, such as stock prices, gas prices, and so on. (Forecasting techniques are covered in Chapter 17.)

One widely used technique for forecasting is known as *time series regression analysis.* A time series is a set of values for a single variable collected over a period of time. For example, the daily prices of Apple stock from 2010 to 2013 would constitute a time series.

As an example, the following regression equation may be used to forecast the *trend* of a time series. (The trend shows how a time series grows over time.)

$$y_t = TR_t + \varepsilon_t$$

The trend may take several different forms, including

- No trend
- Linear trend
- Quadratic trend
- Higher-order trend

Suppose that a time series is collected for the average price of gasoline in New York State over the past ten years. If the time series does not have a trend, this would indicate that gas prices do not grow at a steady rate over time. If the time series has a linear trend, then gasoline prices grow at a constant rate over time. If the time series has a quadratic or higher-order trend, then gasoline prices grow at a rate that changes over time.

Other techniques to forecast a time series include simple moving averages, centered moving averages, and exponentially weighted moving averages. Simple and centered moving averages "smooth" out the values of a time series to produce an estimate of the trend of the series. An exponentially weighted moving average is a more sophisticated version of these techniques and is designed to place less weight on older observations to reflect their diminishing relevance.