chapter 17
Hypothesis Testing

learning objectives

After reading this chapter, you should understand . . .

1 The nature and logic of hypothesis testing.
2 What a statistically significant difference is.
3 The six-step hypothesis testing procedure.
4 The differences between parametric and nonparametric tests and when to use each.
5 The factors that influence the selection of an appropriate test of statistical significance.
6 How to interpret the various test statistics.

Don't confuse "hypothesis" and "theory." The former is a possible explanation; the latter, the correct one. The establishment of theory is the very purpose of science.

Martin H. Fischer, professor emeritus physiology
University of Cincinnati
"Sara, I'd like to meet with you about verifying the gender and age differences on the alcohol issue for Center City Performing Arts Association," Jason makes his way through the cluttered outer office, stepping around piles of printouts, topped with sketched graphs or detailed cross-tabulated tables with handwritten notes.

Moments later, Sara enters, carrying the cross-tabulated data to which Jason had referred.

Jason smiles, waiting for her to settle. "So what have you got?"

"There is definitely a difference in attitude about serving alcohol during intermission at performances. But it doesn't appear to be quite what the CCPA board expected."

"How so?"

"Well, the younger patrons seem somewhat divided, while those between 35 and 54 are against and those 55 and over are strongly in favor."

"What was your original hypothesis?"

"Based on your meeting notes from the project session with the CCPA board, I formulated the hypothesis that there would be a difference on the alcohol issue based on age," shares Sara. "But I assumed that the younger the patrons, the more in favor of alcohol they would be. The numbers just aren't supporting that. And I'm not so sure that age is the right variable to look at."

Jason extends his hand across his desk. "Let me see the statistics on age."

"And I've got the stats on gender, too," offers Sara.

"Are those in line with your hypothesis?"

"Not really," shares Sara. "Men and women are all over the place. I hypothesized that men would be in favor while women would be against. That's just not panning out."

Jason glances at the printout, pleased to see that her interpretation of the statistics is correct. "Looks like you have some work yet, to determine the pockets of resistance. Since the sample split wasn't 57 percent in favor to 43 percent against—we don't want to recommend that CCPA proceed without being able to tell the board the likely direction of potential trouble."

"Sometimes our preliminary analysis plan can take us only so far," comments Jason. "Let's talk about what tests you plan to run now."
> Introduction

In Chapters 15 and 16, we discussed the procedures for data preparation and preliminary analysis. The next step for many studies is hypothesis testing.

Just as your understanding of scientific reasoning was important in the last two chapters, recollection of the specific differences between induction and deduction is fundamental to hypothesis testing. Inductive reasoning moves from specific facts to general, but tentative, conclusions. We can never be absolutely sure that inductive conclusions are flawless. With the aid of probability estimates, we can qualify our results and state the degree of confidence we have in them. Statistical inference is an application of inductive reasoning. It allows us to reason from evidence found in the sample to conclusions we wish to make about the population.

Inferential statistics is the second of two major categories of statistical procedures, the other being descriptive statistics. We used descriptive statistics in Chapter 14 when describing distributions. Under the heading inferential statistics, two topics are discussed in this book. The first, estimation of population values, was used with sampling in Chapter 14, but we will return to it here briefly. The second, testing statistical hypotheses, is the primary subject of this chapter. There are more examples of hypothesis tests in this chapter than most students will need for a term project or early assignments in their research careers. A section on nonparametric techniques in Appendix C provides further study for readers with a special interest in nominal and ordinal variables.

After you have detailed your hypotheses in your preliminary analysis planning, the purpose of hypothesis testing is to determine the accuracy of your hypotheses due to the fact that you have collected a sample of data, not a census. Exhibit 17-1 reminds you of the relationships among your design strategy, data collection activities, preliminary analysis, and hypothesis testing.

We evaluate the accuracy of hypotheses by determining the statistical likelihood that the data reveal true differences—not random sampling error. We evaluate the importance of a statistically significant difference by weighing the practical significance of any change that we measure.

Although there are two approaches to hypothesis testing, the more established is the classical or sampling-theory approach. Classical statistics are found in all of the major statistics books and are widely used in research applications. This approach represents an objective view of probability in which the decision making rests totally on an analysis of available sampling data. A hypothesis is established; it is rejected or fails to be rejected, based on the sample data collected.

The second approach is known as Bayesian statistics, which are an extension of the classical approach. They also use sampling data, but they go beyond to consider all other available information. This additional information consists of subjective probability estimates stated in terms of degrees of belief. These subjective estimates are based on general experience rather than on specific collected data. Various decision rules are established, cost and other estimates can be introduced, and the expected outcomes of combinations of these elements are used to judge decision alternatives.

Statistical Significance

Following classical statistics approach, we accept or reject a hypothesis on the basis of sampling information alone. Since any sample will almost surely vary somewhat from its population, we must judge whether the differences are statistically significant or insignificant. A difference has statistical significance if there is good reason to believe the difference does not represent random sampling fluctuations only. For example, Honda, Toyota, Chrysler, Nissan, Ford, and other auto companies produce hybrid vehicles using an advanced technology that combines a small gas engine with an electric motor. The vehicles run on an electric motor at slow speeds but shift to both the gasoline motor and the electric motor at city and higher freeway speeds. Their advertising strategies focus on fuel economy. Let’s say that the hybrid Toyota has maintained an average of about 60 miles per gallon (mpg) with a standard deviation of 10 mpg. Suppose researchers discover by analyzing all production vehicles that the mpg is now 61. Is this difference statistically significant from 60? Of course it is, because the difference is based on a census of the vehicles and there is no sampling involved. It has been demonstrated conclusively that the population average has moved from 60 to 61 mpg. Although it is of statistical
significance, whether it is of practical significance is another question. If a decision maker judges that this variation has no real importance, then it is of little practical significance.

Since it would be too expensive to analyze all of a manufacturer's vehicles frequently, we resort to sampling. Assume a sample of 25 cars is randomly selected and the average mpg is calculated to be 64. Is this statistically significant? The answer is not obvious. It is significant if there is good reason to believe the average mpg of the total population has moved up from 60. Since the evidence consists of only a sample, consider the second possibility: that this is only a random sampling error and thus is not significant. The task is to decide whether such a result from this sample is or is not statistically significant. To answer this question, one needs to consider further the logic of hypothesis testing.
The Logic of Hypothesis Testing

In classical tests of significance, two kinds of hypotheses are used. The null hypothesis is used for testing. It is a statement that no difference exists between the parameter (a measure taken by a census of the population or a prior measurement of a sample of the population) and the statistic being compared to it (a measure from a recently drawn sample of the population). Analysts usually test to determine whether there has been no change in the population of interest or whether a real difference exists. Why not state the hypothesis in a positive form? Why not state that any difference between the sample statistic and the population parameter is due to some reason? Unfortunately, this type of hypothesis cannot be tested definitively. Evidence that is consistent with a hypothesis stated in a positive form can almost never be taken as conclusive grounds for accepting the hypothesis. A finding that is consistent with this type of hypothesis might be consistent with other hypotheses too, and thus it does not demonstrate the truth of the given hypothesis.

For example, suppose a coin is suspected of being biased in favor of heads. The coin is flipped 100 times and the outcome is 52 heads. It would not be correct to jump to the conclusion that the coin is biased simply because more than the expected number of 50 heads resulted. The reason is that 52 heads is consistent with the hypothesis that the coin is fair. On the other hand, flipping 85 or 90 heads in 100 flips would seem to contradict the hypothesis of a fair coin. In this case there would be a strong case for a biased coin.

In the hybrid-vehicle example, the null hypothesis states that the population parameter of 60 mpg has not changed. A second, alternative hypothesis holds that there has been a change in average mpg (i.e., the sample statistic of 64 indicates the population value probably is no longer 60). The alternative hypothesis is the logical opposite of the null hypothesis.

The hybrid-car example can be explored further to show how these concepts are used to test for significance:

- The null hypothesis ($H_0$): There has been no change from the 60 mpg average.

The alternative hypothesis ($H_a$) may take several forms, depending on the objective of the researchers. The $H_a$ may be of the "not the same" or the "greater than" or "less than" form:

- The average mpg has changed from 60.
- The average mpg has increased (decreased) from 60.

These types of alternative hypotheses correspond with two-tailed and one-tailed tests. A two-tailed test, or nondirectional test, considers two possibilities: the average could be more than 60 mpg, or it could be less than 60. To test this hypothesis, the regions of rejection are divided into two tails of the distribution. A one-tailed test, or directional test, places the entire probability of an unlikely outcome into the tail specified by the alternative hypothesis. In Exhibit 17-2, the first diagram represents a nondirectional hypothesis, and the second is a directional hypothesis of the "greater than" variety.
Hypothesis Testing

Hypotheses for Exhibit 17-2 may be expressed in the following form:

Null  \( H_0: \mu = 60 \text{ mpg} \)
Alternative  \( H_A: \mu \neq 60 \text{ mpg} \) (not the same case)

Or

Null  \( H_0: \mu \leq 60 \text{ mpg} \)
Alternative  \( H_A: \mu > 60 \text{ mpg} \) (greater than case)

Or

Null  \( H_0: \mu \geq 60 \text{ mpg} \)
Alternative  \( H_A: \mu < 60 \text{ mpg} \) (less than case)

In testing these hypotheses, adopt this decision rule: take no corrective action if the analysis shows that one cannot reject the null hypothesis. Note the language "cannot reject" rather than "accept" the null hypothesis. It is argued that a null hypothesis can never be proved and therefore cannot be "accepted." Here, again, we see the influence of inductive reasoning. Unlike deduction, where the connections between premises and conclusions provide a legitimate claim of "conclusive proof," inductive conclusions do not possess that advantage. Statistical testing gives only a chance to (1) disprove (reject) or (2) fail to reject the hypothesis. Despite this terminology, it is common to hear "accept the null" rather than the clumsy "fail to reject the null." In this discussion, the less formal accept means "fail to reject" the null hypothesis.

If we reject a null hypothesis (finding a statistically significant difference), then we are accepting the alternative hypothesis. In either accepting or rejecting a null hypothesis, we can make incorrect decisions. A null hypothesis can be accepted when it should have been rejected or rejected when it should have been accepted.

These problems are illustrated with an analogy to the American legal system. In our system of justice, the innocence of an indicted person is presumed until proof of guilt beyond a reasonable doubt can be established. In hypothesis testing, this is the null hypothesis; there should be no difference between the presumption of innocence and the outcome unless contrary evidence is furnished. Once evidence establishes beyond reasonable doubt that innocence can no longer be maintained, a just conviction is required. This is equivalent to rejecting the null hypothesis and accepting the alternative hypothesis. Incorrect decisions or errors are the other two possible outcomes. We can unjustly convict an innocent person, or we can acquit a guilty person.

Toyota Prius, 2010 Car of the Year, is the preeminent hybrid gas-electric car. It inspires a cult-like devotion which has translated into unprecedented satisfaction rates in user studies. It produces an EPA fuel economy of 51 mpg city and 60 mpg highway. Its "harmony" ad campaign, featuring 100 actors representing 1,000,000 images of people portraying grass, leaves, sun, cloud, flowers, and butterflies created almost as big a sensation as the 3rd generation car.
Direct-to-consumer (D-to-C) pharmaceutical ads have drawn a lot of criticism since 1997 Food and Drug Administration (FDA) regulations permitted such tactics. Proponents of legislation to disallow the practice fear such ads "unfairly influence important health care decisions" by causing patients to pressure doctors and thus encourage doctors to prescribe unnecessary prescriptions. The chairman of the American Medical Association believes such advertising may create an adversarial relationship between doctor and patient. He wants to know if the ads "improve the quality of care enough to make it worth the increased costs of the medicines being advertised." One democratic legislator believes "taxpayers would not have to subsidize excessive advertising that leads to higher prices at the pharmacy counter."

Ipsos-NPD tracks this issue for the pharmaceutical industry with its monthly PharmTrends panel, comprising 16,000 U.S. households. Panel members are measured for ad recall, prescriptions filled, physician recommendations for over-the-counter (OTC) products, and OTC products purchased, as well as condition being treated. Panel findings reveal that advertising "has encouraged higher levels of script fulfillment per year among consumers who reported that they were aware of advertising." Additionally, such advertising is credited with reminding patients to refill prescriptions. In its February InstaVue omnibus mail survey of 20,000 adults, 47 percent had seen pharmaceutical advertising in the past year, 25 percent indicated D-to-C ads encouraged them to call/visit their doctors to discuss the pharmaceutical advertised, and 15 percent reported asking for the very drug advertised.

How would you determine if this research confirmed or refuted that "pharmaceutical ads undermine quality of care"?

Exhibit 17-3 compares the statistical situation to the legal one. One of two conditions exists in nature—either the null hypothesis is true or the alternative hypothesis is true. An indicted person is innocent or guilty. Two decisions can be made about these conditions; one may accept the null hypothesis or reject it (thereby accepting the alternative). Two of these situations result in correct decisions; the other two lead to decision errors.
When a **Type I error** ($\alpha$) is committed, a true null hypothesis is rejected; the innocent person is unjustly convicted. The value is called the **level of significance** and is the probability of rejecting the true null. With a **Type II error** ($\beta$), one fails to reject a false null hypothesis; the result is an unjust acquittal, with the guilty person going free. In our system of justice, it is more important to reduce the probability of convicting the innocent than that of acquitting the guilty. Similarly, hypothesis testing places a greater emphasis on Type I errors than on Type II errors. Next we shall examine each of these errors in more detail.

**Type I Error**

Assume the hybrid manufacturer’s problem is complicated by a consumer testing agency’s assertion that the average city mpg has changed. Assume the population mean is 50 mpg, the standard deviation of the population is 10 mpg, and the size of the sample is 25 vehicles. With this information, one can calculate the standard error of the mean ($\sigma_x$) (the standard deviation of the distribution of sample means). This hypothetical distribution is pictured in Exhibit 17-4. The standard error of the mean is calculated to be 2 mpg:

$$\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$$
If the decision is to reject $H_0$ with a 95 percent confidence interval ($\alpha = .05$), a Type I error of .025 in each tail is accepted (this assumes a two-tailed test). In part A of Exhibit 17-4, see the regions of rejection indicated by the shaded areas. The area between these two regions is known as the region of acceptance. The dividing points between the rejection and acceptance areas are called critical values. Since the distribution of sample means is normal, the critical values can be computed in terms of the standardized random variable, $Z$, where

$$Z = 1.96 \text{ (significance level } \alpha = .05)$$

$X_c$ = the critical value of the sample mean

$\mu$ = the population value stated in $H_0 = 50$

$\sigma_x$ = the standard error of a distribution of means of samples of 25

Thus, the critical values for the test of the null hypothesis (that the mpg has not changed) are computed as follows:

$$Z = \frac{\bar{X} - \mu}{\sigma_x}$$

$$-1.96 = \frac{\bar{X}_c - 50}{\sigma_x}$$

$$\bar{X}_c = 46.08$$

$$1.96 = \frac{\bar{X}_c - 50}{\sigma_x}$$

$$\bar{X}_c = 53.92$$

If the probability of a Type I error is 5 percent ($\alpha = .05$), the probability of a correct decision if the null hypothesis is true is 95 percent. By changing the probability of a Type I error, you move critical values either closer to or farther away from the assumed parameter of 50. This can be done if a smaller or larger $\alpha$ error is desired and critical values are moved to reflect this. You can also change the Type I error and the regions of acceptance by changing the size of the sample. For example, if you take a sample of 100, the critical values that provide a Type I error of .05 are 48.04 and 51.96.

The alternative hypothesis concerned a change in either direction from 50, but the manufacturer is interested only in increases in mpg. For this, one uses a one-tailed (greater than) $H_A$ and places the entire region of rejection in the upper tail of the distribution. One can accept a 5 percent $\alpha$ risk and compute a new critical value ($X_c$). (See Appendix D, Exhibit D-1, to find the $Z$ value of 1.645 for the area of .05 under the curve.) Substitute this in the $Z$ equation and solve for $X_c$:

$$Z = 1.645 = \frac{\bar{X}_c - 50}{\sigma_x}$$

$$\bar{X}_c = 53.29$$

This new critical value, the boundary between the regions of acceptance and rejection, is pictured in part B of Exhibit 17-4.

**Type II Error**

The manufacturer would commit a Type II error ($\beta$) by accepting the null hypothesis ($\mu = 50$) when in truth it had changed. This kind of error is difficult to detect. The probability of committing a Type II error depends on five factors: (1) the true value of the parameter, (2) the $\alpha$ level we have selected, (3) whether a one- or two-tailed test was used to evaluate the hypothesis, (4) the sample standard deviation, and (5) the size of the sample. We secure a different $\beta$ error if the new $\mu$ moves from 50 to 54 rather than only to 52. We must compute separate $\beta$ error estimates for each of a number of assumed new population parameters and $X_c$ values.
To illustrate, assume \( \mu \) has actually moved to 54 from 50. Under these conditions, what is the probability of our making a Type II error if the critical value is set at 53.29? (See Exhibit 17-5.) This may be expressed in the following fashion:

\[
P(A_1)S_1 = \alpha = .05 \text{ (assume a one-tailed alternative hypothesis)}
\]

\[
P(A_1)S_2 = \beta = ?
\]

\[
\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2
\]

\[
Z = \frac{\bar{X} - \mu}{\sigma_x} = \frac{53.29 - 54}{2} = -.355
\]

Using Exhibit D-1 in Appendix D, we interpolate between .35 and .36 Z scores to find the .355 Z score. The area between the mean and Z is .1387. \( \beta \) is the tail area, or the area below the Z, and is calculated as

\[
\beta = .50 - .1387 = .36
\]

This condition is shown in Exhibit 17-5. It is the percent of the area where we would not reject the null \( H_0: \mu = 50 \) when, in fact, it was false because the true mean was 54. With an \( \alpha \) of .05 and a sample of 25, there is a 36 percent probability of a Type II (\( \beta \)) error if the \( \mu \) is 54. We also speak of the power of the test—that is \( 1 - \beta \). For this example, the power of the test equals 64 percent \( (1 - .36) \)—that is, we will correctly reject the false null hypothesis with a 64 percent probability. A power of 64 percent is less than the 80 percent minimum percentage recommended by statisticians.

There are several ways to reduce a Type II error. We can shift the critical value closer to the original \( \mu \) of 50; but to do this, we must accept a bigger \( \alpha \). Whether to take this action depends on the evaluation of the relative \( \alpha \) and \( \beta \) risks. It might be desirable to enlarge the acceptable \( \alpha \) risk because a worsening of the mileage would probably call for increased efforts to stimulate efficiency. Committing a Type I error would mean only that we engaged in efforts to stimulate efficiency when the situation had not worsened. This act probably would not have many adverse effects even if mpg had not increased.

A second way to reduce Type II error is to increase sample size. For example, if the sample were increased to 100, the power of the test would be much stronger:

\[
\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1
\]

\[
Z = \frac{\bar{X} - \mu}{\sigma_x} = \frac{53.29 - 54}{1} = -.71
\]

\[
\beta = .50 - .2612 = .24
\]
This would reduce the Type II error to 24 percent and increase the power of the test to 76 percent.

A third method seeks to improve both $\alpha$ and $\beta$ errors simultaneously and is difficult to accomplish. We know that measuring instruments, observations, and recording produce error. By using a better measuring device, tightening the observation and recording processes, or devising a more efficient sample, we can reduce the variability of observations. This diminishes the standard error of estimate and in turn reduces the sampling distributions' spread. The net effect is that there is less tail area in the error regions.

**Statistical Testing Procedures**

Testing for statistical significance follows a relatively well-defined pattern, although authors differ in the number and sequence of steps. One six-stage sequence is as follows:

1. **State the null hypothesis.** Although the researcher is usually interested in testing a hypothesis of change or differences, the null hypothesis is always used for statistical testing purposes.
2. **Choose the statistical test.** To test a hypothesis, one must choose an appropriate statistical test. There are many tests from which to choose, and there are at least four criteria that can be used in choosing a test. One is the power efficiency of the test. A more powerful test provides the same level of significance with a smaller sample than a less powerful test. In addition, in choosing a test, one can consider how the sample is drawn, the nature of the population, and (importantly) the type of measurement scale used. For instance, some tests are useful only when the sequence of scores is known or when observations are paired; other tests are appropriate only if the population has certain characteristics; still other tests are useful only if the measurement scale is interval or ratio. More attention is given to test selection later in the chapter.
3. **Select the desired level of significance.** The choice of the level of significance should be made before we collect the data. The most common level is .05, although .01 is also widely used. Other levels such as .10, .025, or .001 are sometimes chosen. The exact level to choose is largely determined by how much risk one is willing to accept and the effect that this choice has on $\beta$ risk. The larger the $\alpha$, the lower is the $\beta$.
4. **Compute the calculated difference value.** After the data are collected, use the formula for the appropriate significance test to obtain the calculated value. Although the computation typically results from a software program, we illustrate the procedures in this chapter to help you visualize what is being done.
5. **Obtain the critical test value.** After we compute the calculated $t$, $\chi^2$, or other measure, we must look up the critical value in the appropriate table for that distribution (or it is provided with the software calculation). The critical value is the criterion that defines the region of rejection from the region of acceptance of the null hypothesis.
6. **Interpret the test.** For most tests if the calculated value is larger than the critical value, we reject the null hypothesis and conclude that the alternative hypothesis is supported (although it is by no means proved). If the critical value is larger, we conclude we have failed to reject the null.³

**Probability Values (p Values)**

According to the "interpret the test" step of the statistical test procedure, the conclusion is stated in terms of rejecting or not rejecting the null hypothesis based on a reject region selected before the test is conducted. A second method of presenting the results of a statistical test reports the extent to which the test statistic disagrees with the null hypothesis. This method has become popular because analysts want to know what percentage of the sampling distribution lies beyond the sample statistic on the curve, and most statistical computer programs report the results of statistical tests as probability values ($p$ values). The $p$ value is the probability of observing a sample value as extreme as, or more extreme than, the value actually observed, given that the null hypothesis is true. This area represents the probability of a Type I error that must be assumed if the null hypothesis is rejected. The $p$ value is compared to the
significance level ($\alpha$), and on this basis the null hypothesis is either rejected or not rejected. If the $p$ value is less than the significance level, the null hypothesis is rejected (if $p$ value < $\alpha$, reject the null). If $p$ is greater than or equal to the significance level, the null hypothesis is not rejected (if $p$ value > $\alpha$, don't reject the null).

Statistical data analysis programs commonly compute the $p$ value during the execution of a hypothesis test. The following example will help illustrate the correct way to interpret a $p$ value.

In part B of Exhibit 17-4 the critical value was shown for the situation in which the manufacturer was interested in determining whether the average mpg had increased. The critical value of 33.29 was computed based on a standard deviation of 10, sample size of 25, and the manufacturer's willingness to accept a 5 percent $\alpha$ risk. Suppose that the sample mean equaled 55. Is there enough evidence to reject...
the null hypothesis? If the p value is less than .05, the null hypothesis will be rejected. The p value is computed as follows.

The standard deviation of the distribution of sample means is 2. The appropriate Z value is

\[ Z = \frac{\bar{X} - \mu}{\sigma_x} \]

\[ Z = \frac{55 - 50}{2} \]

\[ Z = 2.5 \]

The p value is determined using the standard normal table. The area between the mean and a Z value of 2.5 is .4938. For this one-tailed test, the p value is the area above the Z value. The probability of observing a Z value at least as large as 2.5 is only .0062 (.5000 - .4938 = .0062) if the null hypothesis is true.

This small p value represents the risk of rejecting a true null hypothesis. It is the probability of a Type I error if the null hypothesis is rejected. Since the p value \( p = .0062 \) is smaller than \( \alpha = .05 \), the null hypothesis is rejected. The manufacturer can conclude that the average mpg has increased. The probability that this conclusion is wrong is .0062.

Tests of Significance

This section provides an overview of statistical tests that are representative of the vast array available to the researcher. After a review of the general types of tests and their assumptions, the procedures for selecting an appropriate test are discussed. The remainder of the section contains examples of parametric and nonparametric tests for one-sample, two-sample, and \( k \)-sample cases. Readers needing a comprehensive treatment of significance tests are referred to the suggested readings for this chapter.

Types of Tests

There are two general classes of significance tests: parametric and nonparametric. Parametric tests are more powerful because their data are derived from interval and ratio measurements. Nonparametric tests are used to test hypotheses with nominal and ordinal data. Parametric techniques are the tests of choice if their assumptions are met. Assumptions for parametric tests include the following:

- The observations must be independent—that is, the selection of any one case should not affect the chances for any other case to be included in the sample.
- The observations should be drawn from normally distributed populations.
- These populations should have equal variances.
- The measurement scales should be at least interval so that arithmetic operations can be used with them.

The researcher is responsible for reviewing the assumptions pertinent to the chosen test. Performing diagnostic checks on the data allows the researcher to select the most appropriate technique. The normality of a distribution may be checked in several ways. We have previously discussed the measures of location, shape, and spread for preliminary analysis and considered graphic techniques for exploring data patterns and examining distributions. Another diagnostic tool is the normal probability plot. This plot compares the observed values with those expected from a normal distribution. If the data display the characteristics of normality, the points will fall within a narrow band along a straight line. An example is shown in the upper left panel of Exhibit 17-6.

An alternative way to look at this is to plot the deviations from the straight line. These are shown in a...
Exhibit 17-6 Probability Plots and Tests of Normality

"detrended" plot in the upper right panel of the figure. Here we would expect the points to cluster without pattern around a straight line passing horizontally through 0. In the bottom two panels of Exhibit 17-6, there is neither a straight line in the normal probability plot nor a random distribution of points about 0 in the detrended plot. Visually, the bottom two plots tell us the variable is not normally distributed. In addition, two separate tests of the hypothesis that the data come from normal distributions are rejected at a significance level of less than .01.

If we wished to check another assumption—say, one of equal variance—a spread-and-level plot would be appropriate. Statistical software programs often provide diagnostic tools for checking assumptions. These may be nested within a specific statistical procedure, such as analysis of variance or regression, or provided as a general set of tools for examining assumptions.
Are American teens exposed to unrealistic drug usage or to unrealistic consequences from such use? Mediascope, a nonprofit organization concerned with responsible depictions of social and health issues in the media, recently completed for the Office of National Drug Control Policy a content analysis of the top 200 rental movies to determine their depiction of substance use. The researchers used the Entertainment Merchants Association's most popular (top 100) home video titles based on rental income during two sequential years. Movies were categorized as follows: action adventure, comedy, or drama. Data were also collected on each title's Motion Picture Association of America (MPAA) rating (G, PG, PG-13, or R). Although technically teens should have been excluded from R-rated titles (which made up 48 percent of the overall sample), the study included all 20 of the most popular teen movies as identified in a prior independent study.

Trained coders watched all 200 movies, paying particular attention to alcohol, tobacco, illicit drugs, over-the-counter medicines, prescription medicines, inhalants, and unidentified pills. Coders ignored substances administered by medical personnel in a hospital or health-related scenario. Substance use included explicit portrayals of consumption. Substance appearance included evidence of materials or paraphernalia without any indication of use. Coders identified dominant messages about substance use and the consequences of use. Coders also noted scenes depicting illicit drug use or those depicting use by characters known to be under 18. Prevalence of use was determined by counting the characters in each movie and determining not only the percentage of characters using drugs but also whether the character had a major or minor role. Coders profiled characters by age, gender, and ethnicity, as well as other characteristics. Frequency of substance abuse was determined for each five-minute interval of each movie, with the presence or absence of various substances noted, starting with the completion of the title credits and ending when the final credits began. How would the last movie you watched have fared under this scrutiny?

www.mediacampaign.org; www.mediascope.org; www.vada.org

Parametric tests place different emphasis on the importance of assumptions. Some tests are quite robust and hold up well despite violations. For others, a departure from linearity or equality of variance may threaten the validity of the results.

Nonparametric tests have fewer and less stringent assumptions. They do not specify normally distributed populations or equality of variance. Some tests require independence of cases; others are expressly designed for situations with related cases. Nonparametric tests are the only ones usable with nominal data; they are the only technically correct tests to use with ordinal data, although parametric tests are sometimes employed in this case. Nonparametric tests may also be used for interval and ratio data, although they waste much of the information available. Nonparametric tests are also easy to understand and use. Parametric tests have greater efficiency when their use is appropriate, but even in such cases nonparametric tests often achieve an efficiency as high as 95 percent. This means the nonparametric test with a sample of 100 will provide the same statistical testing power as a parametric test with a sample of 95.

How to Select a Test

In attempting to choose a particular significance test, the researcher should consider at least three questions:

- Does the test involve one sample, two samples, or k (more than two) samples?
- If two samples or k samples are involved, are the individual cases independent or related?
- Is the measurement scale nominal, ordinal, interval, or ratio?
Additional questions may arise once answers to these are known: What is the sample size? If there are several samples, are they of equal size? Have the data been weighted? Have the data been transformed? Often such questions are unique to the selected technique. The answers can complicate the selection, but once a tentative choice is made, standard statistics textbooks will provide further details.

Decision trees provide a more systematic means of selecting techniques. One widely used guide from the Institute for Social Research starts with questions about the number of variables, nature of the variables (continuous, discrete, dichotomous, independent, dependent, and so forth), and level of measurement. It goes through a tree structure asking detailed questions about the nature of the relationships being searched, compared, or tested. More than 130 solutions to data analysis problems are paired with commonly asked questions.\(^5\)

An expert system offers another approach to choosing appropriate statistics. Capitalizing on the power and convenience of personal computers, expert system programs provide a comprehensive search of the statistical terrain just as a computer search of secondary sources does. Most programs ask about your research objectives, the nature of your data, and the intended audience for your final report. When you are not 100 percent confident of your answers, you can bracket them with an estimate of the degree of your certainty. One such program, Statistical Navigator\(^\text{TM}\) covers various categories of statistics from exploratory data analysis through reliability testing and multivariate data analysis. In response to your answers, a report is printed containing recommendations, rationale for selections, references, and the statistical packages that offer the suggested procedure.\(^7\) SPSS and SAS include coaching and help modules with their software.

**Selecting Tests Using the Choice Criteria**

In the next section, we use the three questions discussed in the last section (see bullets) to develop a classification of the major parametric and nonparametric tests and measures. Because parametric tests are preferred for their power when their assumptions are met, we discuss them first in each of the subsections: one-sample tests, two-sample tests, \(k\)- (more-than-two) sample tests. This is shown in Exhibit 17-7.\(^8\) To illustrate the application of the criteria to test selection, consider that your testing situation involves two samples, the samples are independent, and the data are interval. The figure suggests the \(t\)-test of differences as the appropriate choice. The most frequently used of the tests listed in Exhibit 17-7 are covered next. For additional examples see Appendix C.

**Exhibit 17-7 Recommended Statistical Techniques by Measurement Level and Testing Situation**

<table>
<thead>
<tr>
<th>Measurement Scale</th>
<th>One-Sample Case</th>
<th>Two-Samples Tests</th>
<th>(k)-Samples Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>• Binomial</td>
<td>• McNemar</td>
<td>• Cochran Q</td>
</tr>
<tr>
<td>Ordinal</td>
<td>• Kolmogorov-Smirnov one-sample test</td>
<td>• Wilcoxon matched-pairs test</td>
<td>• Friedman two-way ANOVA</td>
</tr>
<tr>
<td>Interval and Ratio</td>
<td>• (t)-test</td>
<td>• (t)-test for paired samples</td>
<td>• (t)-test</td>
</tr>
<tr>
<td></td>
<td>• (Z) test</td>
<td>• (t)-test for paired samples</td>
<td>• (Z) test</td>
</tr>
</tbody>
</table>

\(^5\) Some examples of such techniques include decision trees and expert systems.

\(^7\) SPSS and SAS are examples of programs that include coaching and help modules.

\(^8\) The figure in Exhibit 17-7 illustrates the application of the criteria to test selection.
One-Sample Tests

One-sample tests are used when we have a single sample and wish to test the hypothesis that it comes from a specified population. In this case we encounter questions such as these:

- Is there a difference between observed frequencies and the frequencies we would expect, based on some theory?
- Is there a difference between observed and expected proportions?
- Is it reasonable to conclude that a sample is drawn from a population with some specified distribution (normal, Poisson, and so forth)?
- Is there a significant difference between some measures of central tendency (\( \bar{x} \)) and its population parameter (\( \mu \))?

A number of tests may be appropriate in this situation. The parametric test is discussed first.

Parametric Tests

The Z test or \( t \)-test is used to determine the statistical significance between a sample distribution mean and a parameter.

The Z distribution and \( t \) distribution differ. The \( t \) has more tail area than that found in the normal distribution. This is a compensation for the lack of information about the population standard deviation. Although the sample standard deviation is used as a proxy figure, the imprecision makes it necessary to go farther away from 0 to include the percentage of values in the \( t \) distribution necessarily found in the standard normal.

When sample sizes approach 120, the sample standard deviation becomes a very good estimate of the population standard deviation (\( \sigma \)); beyond 120, the \( t \) and Z distributions are virtually identical.

Some typical real-world applications of the one-sample test are:

- Finding the average monthly balance of credit card holders compared to the average monthly balance five years ago.
- Comparing the failure rate of computers in a 20-hour test of quality specifications.
- Discovering the proportion of people who would shop in a new district compared to the assumed population proportion.
- Comparing the average product revenues this year to last year's revenues.

Example

To illustrate the application of the \( t \)-test in the one-sample case, consider again the hybrid-vehicle problem mentioned earlier. With a sample of 100 vehicles, the researchers find that the mean miles per gallon for the car is 52.5 mpg, with a standard deviation of 14. Do these results indicate the population mean might still be 50?

In this problem, we have only the sample standard deviation (\( s \)). This must be used in place of the population standard deviation (\( \sigma \)). When we substitute \( s \) for \( \sigma \), we use the \( t \) distribution, especially if the sample size is less than 30. We define \( t \) as

\[
t = \frac{\bar{x} - \mu}{s/\sqrt{n}}
\]

This significance test is conducted by following the six-step procedure recommended earlier:

1. **Null hypothesis.**
   \[ H_0: \mu = 50 \text{ miles per gallon (mpg)} \]
   \[ H_1: \mu > 50 \text{ mpg (one-tailed test)} \]

2. **Statistical test.** Choose the \( t \)-test because the data are ratio measurements. Assume the underlying population is normal and we have randomly selected the sample from the population of production vehicles.
3. **Significance level.** Let $\alpha = .05$, with $n = 100$.

4. **Calculated value.**

$$ t = \frac{52.5 - 50}{2.5} = 1.786 $$

$$ \text{d.f.} = n - 1 = 99 $$

5. **Critical test value.** We obtain this by entering the table of critical values of $t$ (see Appendix C, Exhibit C-2, at the back of the book), with 99 degrees of freedom (d.f.) and a level of significance value of .05. We secure a critical value of about 1.66 (interpolated between d.f. = 60 and d.f. = 120 in Exhibit D-2).

6. **Interpretation.** In this case, the calculated value is greater than the critical value (1.786 > 1.66), so we reject the null hypothesis and conclude that the average mpg has increased.

### Nonparametric Tests

In a one-sample situation, a variety of nonparametric tests may be used, depending on the measurement scale and other conditions. If the measurement scale is nominal (classificatory only), it is possible to use either the binomial test or the chi-square ($\chi^2$) one-sample test. The binomial test is appropriate when the population is viewed as only two classes, such as male and female, buyer and nonbuyer, and successful and unsuccessful, and all observations fall into one or the other of these categories. The binomial test is particularly useful when the size of the sample is so small that the $\chi^2$ test cannot be used.

### Chi-Square Test

Probably the most widely used nonparametric test of significance is the chi-square ($\chi^2$) test. It is particularly useful in tests involving nominal data but can be used for higher scales. Typical are cases where persons, events, or objects are grouped in two or more nominal categories such as “yes-no,” “favor-undecided-against,” or class “A, B, C, or D.”

Using this technique, we test for significant differences between the **observed** distribution of data among categories and the **expected** distribution based on the null hypothesis. Chi-square is useful in cases of one-sample analysis, two independent samples, or $k$ independent samples. It must be calculated with actual counts rather than percentages.

In the one-sample case, we establish a null hypothesis based on the expected frequency of objects in each category. Then the deviations of the actual frequencies in each category are compared with the hypothesized frequencies. The greater the difference between them, the less is the probability that these differences can be attributed to chance. The value of $\chi^2$ is the measure that expresses the extent of this difference. The larger the divergence, the larger is the $\chi^2$ value.

The formula by which the $\chi^2$ test is calculated is

$$ \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} $$

in which

- $O_i = \text{observed number of cases categorized in the } i\text{th category}$
- $E_i = \text{expected number of cases in the } i\text{th category under } H_0$
- $k = \text{the number of categories}$

There is a different distribution for $\chi^2$ for each number of degrees of freedom (d.f.), defined as $(k - 1)$ or the number of categories in the classification minus 1:

$$ \text{d.f.} = k - 1 $$

With chi-square contingency tables of the two-samples or $k$-samples variety, we have both rows and columns in the cross-classification table. In that instance, d.f. is defined as rows minus 1 ($r - 1$) times columns minus 1 ($c - 1$):

$$ \text{d.f.} = (r - 1)(c - 1) $$
In a 2 × 2 table there is 1 d.f., and in a 3 × 2 table there are 2 d.f. Depending on the number of degrees of freedom, we must be certain the numbers in each cell are large enough to make the $\chi^2$ test appropriate. When d.f. = 1, each expected frequency should be at least 5 in size. If d.f. > 1, then the $\chi^2$ test should not be used if more than 20 percent of the expected frequencies are smaller than 5 or when any expected frequency is less than 1. Expected frequencies can often be increased by combining adjacent categories. Four categories of freshmen, sophomores, juniors, and seniors might be classified into upper class and lower class. If there are only two categories and still there are too few in a given class, it is better to use the binomial test.

Assume a survey of student interest in the Metro University Dining Club (discussed in Chapter 15) is taken. We have interviewed 200 students and learned of their intentions to join the club. We would like to analyze the results by living arrangement (type and location of student housing and eating arrangements). The 200 responses are classified into the four categories shown in the accompanying table.

<table>
<thead>
<tr>
<th>Living Arrangement</th>
<th>$O_i$</th>
<th>Intend to Join</th>
<th>Number Interviewed</th>
<th>Percent (no. interviewed/200)</th>
<th>$E_i$ = Expected Frequencies (percent × 60)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dorm/fraternity</td>
<td>16</td>
<td>90</td>
<td>45</td>
<td></td>
<td>27</td>
</tr>
<tr>
<td>Apartment/rooming house, nearby</td>
<td>13</td>
<td>40</td>
<td>20</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>Apartment/rooming house, distant</td>
<td>16</td>
<td>40</td>
<td>20</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>Live at home</td>
<td>15</td>
<td>30</td>
<td>15</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>60</td>
<td>200</td>
<td>100</td>
<td></td>
<td>60</td>
</tr>
</tbody>
</table>

Do these variations indicate there is a significant difference among these students, or are they sampling variations only? Proceed as follows:

1. **Null hypothesis.** $H_0$: $O_i = E_i$. The proportion in the population who intend to join the club is independent of living arrangement. In $H_A$: $O_i \neq E_i$, the proportion in the population who intend to join the club is dependent on living arrangement.

2. **Statistical test.** Use the one-sample $\chi^2$ to compare the observed distribution to a hypothesized distribution. The $\chi^2$ test is used because the responses are classified into nominal categories and there are sufficient observations.

3. **Significance level.** Let $\alpha = .05$.

4. **Calculated value.**

   $$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

   Calculate the expected distribution by determining what proportion of the 200 students interviewed were in each group. Then apply these proportions to the number who intend to join the club. Then calculate the following:

   $$\chi^2 = \frac{(16 - 27)^2}{27} + \frac{(13 - 12)^2}{12} + \frac{(16 - 12)^2}{12} + \frac{(15 - 9)^2}{9}$$

   $$= 4.48 + 0.08 + 1.33 + 4.0$$

   $$= 9.89$$

   d.f. = $(4 - 1)(2 - 1) = 3$

5. **Critical test value.** Enter the table of critical values of $\chi^2$ (see Exhibit D-3), with 3 d.f., and secure a value of 7.82 for $\alpha = .05$.

6. **Interpretation.** The calculated value (9.89) is greater than the critical value (7.82), so the null hypothesis is rejected and we conclude that intending to join is dependent on living arrangement.
Research beyond the Clip

You're McDonald's and you just announced that you are closing 300 stores in the United States. How is this playing in newspapers across America? Is the reporting balanced? Or maybe you're BP (formerly British Petroleum), and you've just invested millions to change your corporate identification and logo and to reposition your firm as the "environmentally friendly" energy conglomerate through executive presentations and advertising. How is the press treating the story? Is the spin positive or negative? Where is the story appearing? Is your key message getting through? The Burrelle's Information Services division of Burrelle's/Luce (B/L) offers one of the longest established "clipping" services used by public relations managers to answer questions like these.

"The most basic research we provide clients is the ad-equivalent value of the news clips," shared Sharon Miller, account executive for B/L. The client notifies B/L that it plans to distribute a press release. Staff at B/L scan the desired print and Internet sources for news of the client—"clips." For print and online publications, the actual space that the story occupies is physically measured. Then, that space is multiplied by the ad rate for identical space in that medium—the ad equivalent." Burrelle's also delivers an assessment of the coverage of key messages the client tried to convey, as well as the tone of the story and the firm's prominence in any story—did the firm get mentioned in the headline or the lead paragraph, or was it the focus of more than half of the article? Managers can obtain comparative analysis evaluating their firm's news coverage against that of other firms in their industry or against ROI investment criteria through online reporting via B/L's secure Insight platform.

If you were a public relations professional, how would you test the hypothesis that your coverage for any given event was more positive than that of your competition? www.burrellesluce.com

Two-Independent-Samples Tests

The need to use two-independent-samples tests is often encountered in business research. We might compare the purchasing predispositions of a sample of subscribers from two magazines to discover if they are from the same population. Similarly, a test of distribution methods from two channels or the market share movements from two competing products could be compared.

Parametric Tests

The Z and t-tests are frequently used parametric tests for independent samples, although the F test also can be used.

The Z test is used with large sample sizes (exceeding 30 for both independent samples) or with smaller samples when the data are normally distributed and population variances are known. The formula for the Z test is

\[ Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \]

With small sample sizes, normally distributed populations, and the assumption of equal population variances, the t-test is appropriate:

\[ t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S^2}{n_1} + \frac{1}{n_2}}} \]

where

\[ \mu_1 - \mu_2 \] is the difference between the two population means.

\[ S^2 \] is associated with the pooled variance estimate:

\[ S^2_p = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \]
To illustrate this application, consider a problem that might face a manager at KDL, a media firm that is evaluating account executive trainees. The manager wishes to test the effectiveness of two methods for training new account executives. The company selects 22 trainees, who are randomly divided into two experimental groups. One receives type A and the other type B training. The trainees are then assigned and managed without regard to the training they have received. At the year’s end, the manager reviews the performances of employees in these groups and finds the following results:

<table>
<thead>
<tr>
<th></th>
<th>A Group</th>
<th>B Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average hourly sales</td>
<td>$1,500</td>
<td>$1,300</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>225</td>
<td>251</td>
</tr>
</tbody>
</table>

Following the standard testing procedure, we will determine whether one training method is superior to the other:

1. **Null hypothesis.**
   - $H_0$: There is no difference in sales results produced by the two training methods.
   - $H_1$: Training method A produces sales results superior to those of method B.

2. **Statistical test.** The t-test is chosen because the data are at least interval and the samples are independent.

3. **Significance level.** $\alpha = .05$ (one-tailed test).

4. **Calculated value.**
   
   $t = \frac{(1,500 - 1,300) - 0}{\sqrt{\frac{(10)(225)^2 + (10)(251)^2}{20}(\frac{1}{11} + \frac{1}{11})}}$

   $= \frac{200}{101.63} = 1.97$

   There are $n - 1$ degrees of freedom in each sample, so total d.f. is

   $\text{d.f.} = (11 - 1) + (11 - 1) = 20$

5. **Critical test value.** Enter Appendix D, Exhibit D-2 with d.f. = 20, one-tailed test, $\alpha = .05$. The critical value is 1.725.

6. **Interpretation.** Since the calculated value is larger than the critical value ($1.97 > 1.725$), reject the null hypothesis and conclude that training method A is superior.

### Nonparametric Tests

The chi-square ($\chi^2$) test is appropriate for situations in which a test for differences between samples is required. It is especially valuable for nominal data but can be used with ordinal measurements. When parametric data have been reduced to categories, they are frequently treated with $\chi^2$ although this results in a loss of information. Preparing to solve this problem is the same as presented earlier although the formula differs slightly:

$\chi^2 = \sum \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

in which

- $O_{ij}$ = observed number of cases categorized in the $ij$th cell
- $E_{ij}$ = expected number of cases under $H_0$ to be categorized in the $ij$th cell

Suppose MindWriter is implementing a smoke-free workplace policy and is interested in whether smoking affects worker accidents. Since the company has complete reports on on-the-job accidents, a sample of names of workers is drawn from those who were involved in accidents during the last year. A similar sample from among workers who had no reported accidents in the last year is drawn.
of both groups are interviewed to determine if each is a nonsmoker or smoker and, if a smoker, whether the person classifies himself or herself as a heavy or moderate smoker. The results appear in the following table, with expected values calculated as shown.

<table>
<thead>
<tr>
<th>Smoker</th>
<th>Expected Values</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heavy smoker</td>
<td>1,1</td>
<td>1.2</td>
<td>127</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>127</td>
<td>4</td>
<td>824</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8.24</td>
<td>7.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>2,1</td>
<td>2.2</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>6</td>
<td>7.73</td>
<td>7.27</td>
</tr>
<tr>
<td></td>
<td>7.73</td>
<td>7.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonsmoker</td>
<td>3,1</td>
<td>3.2</td>
<td>13</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>18.03</td>
<td>16.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Column Total</td>
<td>34</td>
<td>32</td>
<td>66</td>
<td></td>
</tr>
</tbody>
</table>

The testing procedure is:
1. **Null hypothesis.**
   - $H_0$: There is no relationship in on-the-job accident occurrences between smokers and nonsmokers.
   - $H_A$: There is a relationship in on-the-job accident occurrences between smokers and nonsmokers.

2. **Statistical test.** $\chi^2$ is appropriate, but it may waste some of the data because the measurement appears to be ordinal.

3. **Significance level.** $\alpha = .05$, with d.f. $= (3 - 1)(2 - 1) = 2$

4. **Calculated value.** The expected distribution is provided by the marginal totals of the table. If there is no relationship between accidents and smoking, there will be the same proportion of smokers in both accident and nonaccident groups. The numbers of expected observations in each cell are calculated by multiplying the two marginal totals common to a particular cell and dividing this product by $n$. For example,

   \[
   \frac{34 \times 16}{66} = 8.24, \text{ the expected value in cell (1,1)}
   \]

   \[
   \chi^2 = \frac{(12 - 8.24)^2}{8.24} + \frac{(4 - 7.75)^2}{7.75} + \frac{(9 - 7.73)^2}{7.73} + \frac{(6 - 7.72)^2}{7.72} + \frac{(13 - 18.03)^2}{18.03} + \frac{(22 - 16.97)^2}{16.97}
   \]

   \[
   = 7.01
   \]

5. **Critical test value.** Turn to Appendix D, Exhibit D-3, and find the critical value 7.01 with $\alpha = .05$ and d.f. = 2.

6. **Interpretation.** Since the calculated value is greater than the critical value, the null hypothesis is rejected.

For chi-square to operate properly, data must come from random samples of multinomial distributions, and the expected frequencies should not be too small. We previously noted the traditional cautions that expected frequencies ($E_i$) below 5 should not compose more than 20 percent of the cells, and that no cell should have an $E_i$ of less than 1. Some research has argued that these restrictions are too severe.
In another type of $\chi^2$, the 2 × 2 table, a correction known as *Yates' correction for continuity* is applied when sample sizes are greater than 40 or when the sample is between 20 and 40 and the values of $E_i$ are 5 or more. (We use this correction because a continuous distribution is approximating a discrete distribution in this table. When the $E_i$'s are small, the approximation is not necessarily a good one.) The formula for this correction is

$$\chi^2 = \frac{\left|AD - BC\right|^2}{(A + B)(C + D)(A + C)(B + D)}$$

where the letters represent the cells designated as

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

When the continuity correction is applied to the data shown in Exhibit 17-8, a $\chi^2$ value of 5.25 is obtained. The observed level of significance for this value is .02192. If the level of significance had been set at .01, we would accept the null hypothesis. However, had we calculated $\chi^2$ without correction, the value would have been 6.25, which has an observed level of significance of .01287. Some researchers may be tempted to reject the null at this level. (But note that the critical value of $\chi^2$ at .01 with 1 d.f. is 6.64. See Appendix D, Exhibit D-3.) The literature is in conflict regarding the merits of Yates' correction, but if nothing else, this example suggests one should take care when interpreting 2 × 2 tables. To err on the conservative side would be in keeping with our prior discussion of Type I errors.

The Mantel-Haenszel test and the likelihood ratio also appear in Exhibit 17-8. The former is used with ordinal data; the latter, based on maximum likelihood theory, produces results similar to Pearson's chi-square.

### Two-Related-Samples Tests

The two-related-samples tests concern those situations in which persons, objects, or events are closely matched or the phenomena are measured twice. One might compare the consumption of husbands and
wives, the performance of employees before and after vacations, or the effects of a marketing test stimulus when persons were randomly assigned to groups and given pretests and posttests. Both parametric and nonparametric tests are applicable under these conditions.

Parametric Tests

The *t*-test for independent samples would be inappropriate for this situation because one of its assumptions is that observations are independent. This problem is solved by a formula in which the difference is found between each matched pair of observations, thereby reducing the two samples to the equivalent of a one-sample case—that is, there are now several differences, each independent of the other, for which one can compute various statistics.

In the following formula, the average difference $\bar{D}$ corresponds to the normal distribution when the difference is known and the sample size is sufficient. The statistic $t$ with $(n - 1)$ degrees of freedom is defined as

$$t = \frac{\bar{D}}{S_{D}/\sqrt{n}}$$

where

$$\bar{D} = \frac{\sum D}{n}$$

$$S_{D} = \sqrt{\frac{\sum D^2 - (\sum D)^2}{n-1}}$$

To illustrate, we use two years of *Forbes* sales data (in millions of dollars) from 10 companies, as listed in Exhibit 17-9.

1. Null hypothesis.
   $H_0: \mu = 0$; there is no difference between year 1 and year 2 sales.
   $H_1: \mu \neq 0$; there is a difference between year 1 and year 2 sales.

2. Statistical test. The matched- or paired-samples *t*-test is chosen because there are repeated measures on each company, the data are not independent, and the measurement is ratio.

<table>
<thead>
<tr>
<th>Company</th>
<th>Sales Year 2</th>
<th>Sales Year 1</th>
<th>Difference $D$</th>
<th>$D^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM</td>
<td>126932</td>
<td>123505</td>
<td>3427</td>
<td>11744329</td>
</tr>
<tr>
<td>GE</td>
<td>54874</td>
<td>49662</td>
<td>4912</td>
<td>24127744</td>
</tr>
<tr>
<td>Exxon</td>
<td>86566</td>
<td>78344</td>
<td>7712</td>
<td>59474944</td>
</tr>
<tr>
<td>IBM</td>
<td>62710</td>
<td>59512</td>
<td>3198</td>
<td>10227204</td>
</tr>
<tr>
<td>Ford</td>
<td>95146</td>
<td>92300</td>
<td>3846</td>
<td>14791716</td>
</tr>
<tr>
<td>AT&amp;T</td>
<td>36112</td>
<td>35173</td>
<td>939</td>
<td>881721</td>
</tr>
<tr>
<td>Mobil</td>
<td>50220</td>
<td>48111</td>
<td>2109</td>
<td>4447831</td>
</tr>
<tr>
<td>DuPont</td>
<td>35999</td>
<td>32427</td>
<td>3572</td>
<td>7139584</td>
</tr>
<tr>
<td>Sears</td>
<td>53704</td>
<td>49975</td>
<td>3129</td>
<td>14584761</td>
</tr>
<tr>
<td>Amoco</td>
<td>23066</td>
<td>20779</td>
<td>3187</td>
<td>10156969</td>
</tr>
</tbody>
</table>

$\Sigma D = 35821$  $\Sigma D^2 = 157576853$
3. Significance level. Let $\alpha = .01$, with $n = 10$ and d.f. = $n - 1$.

4. Calculated value.

$$t = \frac{\bar{D}}{S_D / \sqrt{n}} = \frac{3582.10}{570.98} = 6.27 \quad \text{d.f.} = 9$$

5. Critical test value. Enter Appendix D, Exhibit D-2, with d.f. = 9, two-tailed test, $\alpha = .01$. The critical value is 3.25.

6. Interpretation. Since the calculated value is greater than the critical value ($6.27 > 3.25$), reject the null hypothesis and conclude there is a statistically significant difference between the two years of sales.

A computer solution to the problem is illustrated in Exhibit 17-10. Notice that an observed significance level is printed for the calculated $t$ value (highlighted). With SPSS, this is often rounded and would be interpreted as significant at the .0005 level. The correlation coefficient, to the left of the $t$ value, is a measure of the relationship between the two pairs of scores. In situations where matching has occurred (such as husbands' and wives' scores), it reveals the degree to which the matching has been effective in reducing the variability of the mean difference.

Nonparametric Tests

The McNemar test may be used with either nominal or ordinal data and is especially useful with before-after measurement of the same subjects. Test the significance of any observed change by setting up a fourfold table of frequencies to represent the first and second set of responses:

<table>
<thead>
<tr>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Favor</td>
<td>Do Not Favor</td>
</tr>
<tr>
<td>Do Not Favor</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>C</td>
</tr>
</tbody>
</table>

Since $A + D$ represents the total number of people who changed ($B$ and $C$ are no-change responses), the expectation under a null hypothesis is that $1/2 (A + D)$ cases change in one direction and the same proportion in the other direction. The McNemar test uses a transformation of the $\chi^2$ test:

$$\chi^2 = \frac{(A - D - 1)^2}{A + D} \text{ with d.f.} = 1$$

The "minus 1" in the equation is a correction for continuity since the $\chi^2$ is a continuous distribution and the observed frequencies represent a discrete distribution.
To illustrate this test's application, we use survey data from SteelShelf Corporation, whose researchers decided to test a new concept in office seating with employees at the company's headquarters facility. Managers took a random sample of their employees before the test, asking them to complete a questionnaire on their attitudes toward the design concept. On the basis of their responses, the employees were divided into equal groups reflecting their favorable or unfavorable views of the design. After the campaign, the same 200 employees were asked again to complete the questionnaire. They were again classified as to favorable or unfavorable attitudes. The testing process is:

1. **Null hypothesis.**
   
   \[ H_0: P(A) = P(D) \]

2. **Statistical test.** The McNemar test is chosen because nominal data are used and the study involves before-after measurements of two related samples.

3. **Significance level.** Let \( \alpha = .05 \), with \( n = 200 \).

4. **Calculated value.**
   
   \[ \chi^2 = \frac{(110 - 40 - 1)^2}{10 + 40} = \frac{29^2}{50} = 16.82 \]
   
   d.f. = 1

5. **Critical test value.** Enter Appendix D, Exhibit D-3, and find the critical value to be 3.84 with \( \alpha = .05 \) and d.f. = 1.

6. **Interpretation.** The calculated value is greater than the critical value (16.82 > 3.84), indicating one should reject the null hypothesis, and conclude that the new concept had a significant positive effect on employees' attitudes. In fact, \( \chi^2 \) is so large that it would have surpassed an \( \alpha \) of .001.

### k-Independent-Samples Tests

We often use *k*-independent-samples tests in research when three or more samples are involved. Under this condition, we are interested in learning whether the samples might have come from the same or identical populations. When the data are measured on an interval-ratio scale and we can meet the necessary assumptions, analysis of variance and the \( F \) test are used. If preliminary analysis shows the assumptions cannot be met or if the data were measured on an ordinal or nominal scale, a nonparametric test should be selected.

As with the two-samples case, the samples are assumed to be independent. This is the condition of a completely randomized experiment when subjects are randomly assigned to various treatment groups. It is also common for an ex post facto study to require comparison of more than two independent sample means.

### Parametric Tests

The statistical method for testing the null hypothesis that the means of several populations are equal is **analysis of variance** (ANOVA). **One-way analysis of variance** is described in this section. It uses a single-factor, fixed-effects model to compare the effects of one treatment or factor (brands of coffee, varieties of residential housing, types of retail stores) on a continuous dependent variable (coffee consumption, hours of TV viewing, shopping expenditures). In a fixed-effects model, the levels of
the factor are established in advance, and the results are not generalizable to other levels of treatment. For example, if coffee were Jamaican-grown, Colombian-grown, and Honduran-grown, we could not extend our inferences to coffee grown in Guatemala or Mexico.

To use ANOVA, certain conditions must be met. The samples must be randomly selected from normal populations, and the populations should have equal variances. In addition, the distance from one value to its group’s mean should be independent of the distances of other values to that mean (independence of error). ANOVA is reasonably robust, and minor variations from normality and equal variance are tolerable. Nevertheless, the analyst should check the assumptions with the diagnostic techniques previously described.

Analysis of variance, as the name implies, breaks down or partitions total variability into component parts. Unlike the t-test, which uses sample standard deviations, ANOVA uses squared deviations of the variance so that computation of distances of the individual data points from their own mean or from the grand mean can be summed (recall that standard deviations sum to zero).

In an ANOVA model, each group has its own mean and values that deviate from that mean. Similarly, all the data points from all of the groups produce an overall grand mean. The total deviation is the sum of the squared differences between each data point and the overall grand mean.

The total deviation of any particular data point may be partitioned into between-groups variance and within-groups variance. The between-groups variance represents the effect of the treatment, or factor. The differences of between-groups means imply that each group was treated differently, and the treatment will appear as deviations of the sample means from the grand mean. Even if this were not so, there would still be some natural variability among subjects and some variability attributable to sampling. The within-groups variance describes the deviations of the data points within each group from the sample mean. This results from variability among subjects and from random variation. It is often called error.

Intuitively, we might conclude that when the variability attributable to the treatment exceeds the variability arising from error and random fluctuations, the viability of the null hypothesis begins to diminish. And this is exactly the way the test statistic for analysis of variance works.

The test statistic for ANOVA is the $F$ ratio. It compares the variance from the last two sources:

$$F = \frac{\text{mean square}_{\text{between}}}{\text{mean square}_{\text{within}}}$$

where

$$\text{Mean square}_{\text{between}} = \frac{\text{sum of squares}_{\text{between}}}{\text{degrees of freedom}_{\text{between}}}$$

$$\text{Mean square}_{\text{within}} = \frac{\text{sum of squares}_{\text{within}}}{\text{degrees of freedom}_{\text{within}}}$$

To compute the $F$ ratio, the sum of the squared deviations for the numerator and denominator are divided by their respective degrees of freedom. By dividing, we are computing the variance as an average or mean; thus the term mean square. The degrees of freedom for the numerator, the mean square between groups, are one less than the number of groups ($k - 1$). The degrees of freedom for the denominator, the mean square within groups, are the total number of observations minus the number of groups ($n - k$).

If the null hypothesis is true, there should be no difference between the population means, and the ratio should be close to 1. If the population means are not equal, the numerator should manifest this difference, and the $F$ ratio should be greater than 1. The $F$ distribution determines the size of ratio necessary to reject the null hypothesis for a particular sample size and level of significance.

To illustrate one-way ANOVA, consider Travel Industry Magazine’s reports from international travelers about the quality of in-flight service on various carriers from the United States to Asia. Before writing a feature story coinciding with a peak travel period, the magazine decided to retain a researcher to secure a more balanced perspective on the reactions of travelers. The researcher selected passengers...
who had current impressions of the meal service, comfort, and friendliness of a major carrier. Three airlines were chosen and 20 passengers were randomly selected for each airline. The data, found in Exhibit 17-11, are used for this and the next two examples. For the one-way analysis of variance problem, we are concerned only with the columns labeled “Flight Service Rating 1” and “Airline.” The factor, airline, is the grouping variable for three carriers.
>Exhibit 17-12 Summary Tables for One-Way ANOVA Example*

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model (airline)</td>
<td>2</td>
<td>11644.033</td>
<td>5822.017</td>
<td>28.304</td>
<td>0.0001</td>
</tr>
<tr>
<td>Residual (error)</td>
<td>57</td>
<td>11724.550</td>
<td>205.694</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>59</td>
<td>23368.583</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Count</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lufthansa</td>
<td>20</td>
<td>39.950</td>
<td>14.006</td>
</tr>
<tr>
<td>Malaysia Airlines</td>
<td>20</td>
<td>58.900</td>
<td>15.089</td>
</tr>
<tr>
<td>Cathay Pacific</td>
<td>20</td>
<td>72.900</td>
<td>13.902</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vs.</th>
<th>Diff.</th>
<th>Crit. Diff.</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lufthansa</td>
<td>Malaysia</td>
<td>19.950</td>
<td>11.400</td>
</tr>
<tr>
<td>Cathay</td>
<td>Malaysia</td>
<td>33.950</td>
<td>11.400</td>
</tr>
<tr>
<td>Malaysia</td>
<td>Cathay</td>
<td>14.000</td>
<td>11.400</td>
</tr>
</tbody>
</table>

*All data are hypothetical.

1. **Null hypothesis.**
   \[ H_0: \mu_{A1} = \mu_{A2} = \mu_{A3} \]
   \[ H_a: \mu_{A1} \neq \mu_{A2} \neq \mu_{A3} \] (The means are not equal.)

2. **Statistical test.** The F test is chosen because we have \( k \) independent samples, accept the assumptions of analysis of variance, and have interval data.

3. **Significance level.** Let \( \alpha = .05 \), and d.f. = [numerator \((k - 1) = (3 - 1) = 2\)], [denominator \((n - k) = (60 - 3) = 57\)] = (2, 57).

4. **Calculated value.**
   \[ F = \frac{MS_A}{MS_E} = \frac{5822.017}{205.694} = 28.304 \] 
   d.f. (2, 57)

See summary in Exhibit 17-12.

5. **Critical test value.** Enter Appendix D, Exhibit D-8, with d.f. (2, 57), \( \alpha = .05 \). The critical value is 3.16.

6. **Interpretation.** Since the calculated value is greater than the critical value \((28.3 > 3.16)\), we reject the null hypothesis and conclude there are statistically significant differences between two or more pairs of means. Note in Exhibit 17-12 that the \( p \) value equals .0001. Since the \( p \) value \((.0001)\) is less than the significance level \(.05\), we have a second method for rejecting the null hypothesis.

The ANOVA model summary in Exhibit 17-12 is a standard way of summarizing the results of analysis of variance. This table contains the sources of variation, degrees of freedom, sum of squares,
mean squares, and calculated $F$ value. The probability of rejecting the null hypothesis is computed up to 100 percent $\alpha$—that is, the probability value column reports the exact significance for the $F$ ratio being tested.

**A Priori Contrasts**

When we compute a $t$-test, it is not difficult to discover the reasons why the null is rejected. But with one-way ANOVA, how do we determine which pairs are not equal? We could calculate a series of $t$-tests, but they would not be independent of each other and the resulting Type I error would increase substantially. This is not recommended. If we decided in advance that a comparison of specific populations was important, a special class of tests known as *a priori contrasts* could be used after the null was rejected with the $F$ test (it is *a priori* because the decision was made before the test).

A modification of the $F$ test provides one approach for computing contrasts:

$$F = \frac{MS_{CON}}{MS_w}$$

The denominator, the within-groups mean square, is the same as the error term of the one-way's $F$ ratio (recorded in the summary table, Exhibit 17-12). We have previously referred to the denominator of the $F$ ratio as the error variance estimator. The numerator of the contrast test is defined as

$$MS_{CON} = SS_{CON} = \frac{\left(\sum \frac{C_jX_j}{n_j}\right)^2}{\sum \frac{C_j^2}{n_j}}$$

where

$C_j$ = the contrast coefficient for the group $j$

$n_j$ = the number of observations recorded for group $j$

A contrast is useful for experimental and quasi-experimental designs when the researcher is interested in answering specific questions about a subset of the factor. For example, in a comparison of coffee products, we have a factor with six levels. The levels, blends of coffee, are meaningfully ordered. Assume we are particularly interested in two Central American—grown blends and one Colombian blend. Rather than looking at all possible combinations, we can channel the power more effectively by stating the comparisons of interest. This increases our likelihood of detecting differences if they really exist.

**Multiple Comparison Tests**

For the probabilities associated with the contrast test to be properly used in the report of our findings, it is important that the contrast strategy be devised ahead of the testing. In the airline study, we had no theoretical reason for an *a priori* contrast. However, when we examine the table of mean ratings (Exhibit 17-12), it is apparent that the airline means were quite different. Comparisons after the results are compared require *post hoc* tests or pairwise *multiple comparison tests* (or *range tests*) to determine which means differ. These tests find homogeneous subsets of means that are not different from each other. Multiple comparison tests test the difference between each pair of means and indicate significantly different group means at an $\alpha$ level of .05, or another level that you specify. Multiple comparison tests use group means and incorporate the $MS_{err}$ term of the $F$ ratio. Together they produce confidence intervals for the population means and a criterion score. Differences between the mean values may be compared.

There are more than a dozen such tests with different optimization goals: maximum number of comparisons, unequal cell size compensation, cell homogeneity, reduction of Type I or Type II errors, and so forth. The merits of various tests have produced considerable debate among statisticians, leaving the researcher without much guidance for the selection of a test. In Exhibit 17-13, we provide a general guide. For the example in Exhibit 17-12, we chose Scheffé's $S$. It is a conservative test that is robust to violations of assumptions. The computer calculated the critical difference criterion as 11.4; all the differences between the pairs of means exceed this. The null hypothesis for the Scheffé was tested at the .05 level. Therefore, we can conclude that all combinations of flight service mean scores differ from each other.
>Exhibit 17-13 Selection of Multiple Comparison Procedures

<table>
<thead>
<tr>
<th>Test</th>
<th>Pairwise Comparisons</th>
<th>Complex Comparisons</th>
<th>Equal n's Only</th>
<th>Unequal n's Assumed</th>
<th>Unequal Variances Not Assumed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisher LSD</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Bonferroni</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Tukey HSD</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Tukey-Kramer</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Games-Howell</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Tamhane T2</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Scheffé S</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Brown-Forsythe</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Newman-Keuls</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Duncan</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Dunnett's T3</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dunnett's C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

While the table in Exhibit 17-12 provides information for understanding the rejection of the one-way null hypothesis and the Scheffé null, in Exhibit 17-14 we use plots for the comparisons. The means plot shows relative differences among the three levels of the factor. The means by standard deviations plot reveals lower variability in the opinions recorded by the hypothetical Lufthansa and Cathay Pacific passengers. Nevertheless, these two groups are sharply divided on the quality of in-flight service, and that is apparent in the plot.

Exploring the Findings with Two-Way ANOVA

Is the airline on which the passengers traveled the only factor influencing perceptions of in-flight service? By extending the one-way ANOVA, we can learn more about the service ratings. There are

>Exhibit 17-14 One-Way Analysis of Variance Plots
Chapter 17  Hypothesis Testing

Exhibit 17-15 Summary Table for Two-Way ANOVA Example*

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airline</td>
<td>2</td>
<td>11644.033</td>
<td>5822.017</td>
<td>39.178</td>
<td>0.0001</td>
</tr>
<tr>
<td>Seat selection</td>
<td>1</td>
<td>3182.817</td>
<td>3182.817</td>
<td>21.418</td>
<td>0.0001</td>
</tr>
<tr>
<td>Airline by seat selection</td>
<td>2</td>
<td>517.033</td>
<td>258.517</td>
<td>1.740</td>
<td>0.1853</td>
</tr>
<tr>
<td>Residual</td>
<td>54</td>
<td>8024.700</td>
<td>148.606</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Means Table Effect: Airline by Seat Selection

<table>
<thead>
<tr>
<th>Count</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>35.600</td>
<td>12.140</td>
<td>3.839</td>
</tr>
<tr>
<td>10</td>
<td>42.300</td>
<td>15.550</td>
<td>4.917</td>
</tr>
<tr>
<td>10</td>
<td>48.500</td>
<td>12.601</td>
<td>3.953</td>
</tr>
<tr>
<td>10</td>
<td>69.300</td>
<td>9.166</td>
<td>2.898</td>
</tr>
<tr>
<td>10</td>
<td>64.800</td>
<td>13.037</td>
<td>4.123</td>
</tr>
<tr>
<td>10</td>
<td>81.000</td>
<td>9.603</td>
<td>3.037</td>
</tr>
</tbody>
</table>

*All data are hypothetical.
*Dependent: Flight service rating 1.

many possible explanations. We have chosen to look at the seat selection of the travelers in the interest of brevity.

Recall that in Exhibit 17-11, data were entered for the variable seat selection: economy and business-class travelers. Adding this factor to the model, we have a two-way analysis of variance. Now three questions may be considered with one model:

* Are differences in flight service ratings attributable to airlines?
* Are differences in flight service ratings attributable to seat selection?
* Do the airline and the seat selection interact with respect to flight service ratings?

The third question reveals a distinct advantage of the two-way model. A separate one-way model on airlines averages out the effects of seat selection. Similarly, a single-factor test of seat selection averages out the effects of the airline choice. But an interaction test of airline by seat selection considers them jointly.

Exhibit 17-15 reports a test of the hypotheses for these three questions. The significance level was established at the .01 level. We first inspect the interaction effect, airline by seat selection, since the individual main effects cannot be considered separately if the factors interact. The interaction was not significant at the .01 level, and the null is accepted. Now the separate main effects, airline and seat selection, can be verified. As with the one-way ANOVA, the null hypothesis for the airline factor was rejected, and seat selection was also rejected (statistically significant at .0001).

Means and standard deviations listed in the table are plotted in Exhibit 17-16. We note a band of similar deviations for economy-class travelers and a band of lower variability for business class—with the exception of one carrier. The plot of cell means confirms visually what we already know from the summary table: there is no interaction between airline and seat selection (p = .185). If an interaction had occurred, the lines connecting the cell means would have crossed rather than displaying a parallel pattern.

Analysis of variance is an extremely versatile and powerful method that may be adapted to a wide range of testing applications. Discussions of further extensions in n-way and experimental designs may be found in the list of suggested readings.
Nonparametric Tests

When there are $k$ independent samples for which nominal data have been collected, the chi-square test is appropriate. It can also be used to classify data at higher measurement levels, but metric information is lost when reduced. The $k$-samples $\chi^2$ test is an extension of the two-independent-samples cases treated earlier. It is calculated and interpreted in the same way.

The Kruskal-Wallis test is appropriate for data that are collected on an ordinal scale or for interval data that do not meet $F$-test assumptions, that cannot be transformed, or that for another reason prove to be unsuitable for a parametric test. Kruskal-Wallis is a one-way analysis of variance by ranks. It assumes random selection and independence of samples and an underlying continuous distribution.

Data are prepared by converting ratings or scores to ranks for each observation being evaluated. The ranks range from the highest to the lowest of all data points in the aggregated samples. The ranks are then tested to decide if they are samples from the same population. An application of this technique is provided in Appendix C.

$k$-Related-Samples Tests

Parametric Tests

A $k$-related-samples test is required for situations where (1) the grouping factor has more than two levels, (2) observations or subjects are matched or the same subject is measured more than once, and (3) the data are at least interval. In test marketing experiments or ex post facto designs with $k$ samples, it is often necessary to measure subjects several times. These repeated measurements are called trials. For example, multiple measurements are taken in studies of stock prices, products evaluated by reliability, inventory, sales, and measures of product performance. Hypotheses for these situations may be tested with a univariate or multivariate general linear model. The latter is beyond the scope of this discussion.

The repeated-measures ANOVA is a special type of $n$-way analysis of variance. In this design, the repeated measures of each subject are related just as they are in the related $t$-test when only two measures are present. In this sense, each subject serves as its own control requiring a within-subjects variance effect to be assessed differently than the between-groups variance in a factor like airline or seat selection. The effects of the correlated measures are removed before calculation of the $F$ ratio.
This model is an appropriate solution for the data presented in Exhibit 17-11. You will remember that the one-way and two-way examples considered only the first rating of in-flight service. Assume a second rating was obtained after one week by reinterviewing the same respondents. We now have two trials for the dependent variable, and we are interested in the same general question as with the one-way ANOVA, with the addition of how the passage of time affects perceptions of in-flight service.

Following the testing procedure, we state:

1. **Null hypotheses.**
   - (1) Airline: \( H_0: \mu_{A1} = \mu_{A2} = \mu_{A3} \)
   - (2) Ratings: \( H_0: \mu_{R1} = \mu_{R2} \)
   - (3) Ratings \( \times \) airline: \( H_0: \left( \mu_{R1A1} - \mu_{R1A2} - \mu_{R1A3} \right) = \left( \mu_{R1A1} - \mu_{R1A2} - \mu_{R1A3} \right) \)

   For the alternative hypotheses, we will generalize to the statement that not all the groups have equal means for each of the three hypotheses.

2. **Statistical test.** The \( F \) test for repeated measures is chosen because we have related trials on the dependent variable for \( k \) samples, accept the assumptions of analysis of variance, and have interval data.

3. **Significance level.** Let \( \alpha = 0.05 \) and d.f. = [airline (2, 57), ratings (1, 57), ratings by airline (2, 57)].

4. **Calculated values.** See summary in Exhibit 17-17.

**Exhibit 17-17 Summary Tables for Repeated-Measures ANOVA**

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airline</td>
<td>2</td>
<td>35527.550</td>
<td>17763.775</td>
<td>67.199</td>
<td>0.0001</td>
</tr>
<tr>
<td>Subject (group)</td>
<td>57</td>
<td>15067.650</td>
<td>254.345</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ratings</td>
<td>1</td>
<td>625.633</td>
<td>625.633</td>
<td>14.318</td>
<td>0.0004</td>
</tr>
<tr>
<td>Ratings by air</td>
<td>2</td>
<td>2061.717</td>
<td>1030.858</td>
<td>23.592</td>
<td>0.0001</td>
</tr>
<tr>
<td>Ratings by subj</td>
<td>57</td>
<td>2490.650</td>
<td>43.696</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Means Table Ratings by Airline**

<table>
<thead>
<tr>
<th>Count</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating 1, Lufthansa</td>
<td>20</td>
<td>38.950</td>
<td>14.006</td>
</tr>
<tr>
<td>Rating 1, Malaysia</td>
<td>20</td>
<td>58.900</td>
<td>15.038</td>
</tr>
<tr>
<td>Rating 1, Cathay</td>
<td>20</td>
<td>72.900</td>
<td>13.902</td>
</tr>
<tr>
<td>Rating 2, Lufthansa</td>
<td>20</td>
<td>32.400</td>
<td>8.268</td>
</tr>
<tr>
<td>Rating 2, Malaysia</td>
<td>20</td>
<td>72.250</td>
<td>10.572</td>
</tr>
<tr>
<td>Rating 2, Cathay</td>
<td>20</td>
<td>79.600</td>
<td>11.285</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Count</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating 1</td>
<td>60</td>
<td>56.917</td>
<td>19.902</td>
</tr>
<tr>
<td>Rating 2</td>
<td>60</td>
<td>61.433</td>
<td>23.208</td>
</tr>
</tbody>
</table>

*All data are hypothetical.

*Dependent: flight service ratings 1 and 2.
Exhibit 17-18 Repeated-Measures ANOVA Plot

Interaction Plot Effect: Ratings by Airline

*All data are hypothetical.

5. Critical test value. Enter Appendix D, Exhibit D-8, with d.f. (2, 57), $\alpha = .05$ and (1, 57), $\alpha = .05$. The critical values are 3.16 (2, 57) and 4.01 (1, 57).

6. Interpretation. The statistical results are grounds for rejecting all three null hypotheses and concluding there are statistically significant differences between means in all three instances. We conclude the perceptions of in-flight service were significantly affected by the different airlines, the interval between the two measures had a significant effect on the ratings, and the measures' time interval and the airlines interacted to a significant degree.

The ANOVA summary table in Exhibit 17-17 records the results of the tests. A means table provides the means and standard deviations for all combinations of ratings by airline. A second table of means reports the differences between flight service ratings 1 and 2. In Exhibit 17-18, there is an interaction plot for these data. Note that the second in-flight service rating was improved in two of the three groups after one week, but for the third carrier, there was a decrease in favorable response. The intersecting lines in the interaction plot reflect this finding.

Nonparametric Tests

When the $k$ related samples have been measured on a nominal scale, the Cochran $Q$ test is a good choice. This test extends the McNemar test, discussed earlier, for studies having more than two samples. It tests the hypothesis that the proportion of cases in a category is equal for several related categories.

When the data are at least ordinal, the Friedman two-way analysis of variance is appropriate. It tests matched samples, ranking each case and calculating the mean rank for each variable across all cases. It uses these ranks to compute a test statistic. The product is a two-way table where the rows represent subjects and the columns represent the treatment conditions. See Appendix C for additional nonparametric tests.
In classical statistics we make inferences about a population based on evidence gathered from a sample. Although we cannot state unequivocally what is true about the entire population, representative samples allow us to make statements about what is probably true and how much error is likely to be encountered in arriving at a decision. The Bayesian approach also employs sampling statistics but has an additional element of prior information to improve the decision maker's judgment.

A difference between two or more sets of data is statistically significant if it actually occurs in a population. To have a statistically significant finding based on sampling evidence, we must be able to calculate the probability that some observed difference is large enough that there is little chance it could result from random sampling. Probability is the foundation for deciding on the acceptability of the null hypothesis, and sampling statistics facilitate acquiring the estimates.

Hypothesis testing can be viewed as a six-step procedure:

1. Establish a null hypothesis as well as the alternative hypothesis. It is a one-tailed test of significance if the alternative hypothesis states the direction of difference. If no direction of difference is given, it is a two-tailed test.

2. Choose the statistical test on the basis of the assumption about the population distribution and measurement level. The form of the data can also be a factor. In light of these considerations, one typically chooses the test that has the greatest power efficiency or ability to reduce decision errors.

3. Select the desired level of confidence. While $\alpha = .05$ is the most frequently used level, many others are also used. The $\alpha$ is the significance level that we desire and is typically set in advance of the study. Alpha or Type I error is the risk of rejecting a true null hypothesis and represents a decision error. The $\beta$ or Type II error is the decision error that results from accepting a false null hypothesis. Usually, one determines a level of acceptable $\alpha$ error and then seeks to reduce the $\beta$ error by increasing the sample size, shifting from a two-tailed to a one-tailed significance test, or both.

4. Compute the actual test value of the data.

5. Obtain the critical test value, usually by referring to a table for the appropriate type of distribution.

6. Interpret the result by comparing the actual test value with the critical test value.

Parametric and nonparametric tests are applicable under the various conditions described in the chapter. They were also summarized in Exhibit 17-6. Parametric tests operate with interval and ratio data and are preferred when their assumptions can be met. Diagnostic tools examine the data for violations of those assumptions. Nonparametric tests do not require stringent assumptions about population distributions and are useful with less powerful nominal and ordinal measures.

In selecting a significance test, one needs to know, at a minimum, the number of samples, their independence or relatedness, and the measurement level of the data. Statistical tests emphasized in the chapter were the $Z$ and $t$-tests; analysis of variance, and chi-square. The $Z$ and $t$-tests may be used to test for the difference between two means. The $t$-test is chosen when the sample size is small. Variations on the $t$-test are used for both independent and related samples.

One-way analysis of variance compares the means of several groups. It has a single grouping variable, called a factor, and a continuous dependent variable. Analysis of variance (ANOVA) partitions the total variation among scores into between-groups (treatment) and within-groups (error) variance. The $F$ ratio, the test statistic, determines if the differences are large enough to reject the null hypothesis. ANOVA may be extended to two-way, n-way, repeated-measures, and multivariate applications.

Chi-square is a nonparametric statistic that is used frequently for cross-tabulation or contingency tables. Its applications include testing for differences between proportions in populations and testing for independence. Corrections for chi-square were discussed.
Terms in Review
1. Distinguish between the following:
   a. Parametric tests and nonparametric tests.
   b. Type I error and Type II error.
   c. Null hypothesis and alternative hypothesis.
   d. Acceptance region and rejection region.
   e. One-tailed tests and two-tailed tests.
   f. Type II error and the power of the test.

2. Summarize the steps of hypothesis testing. What is the virtue of this procedure?

3. In analysis of variance, what is the purpose of the mean square between and the mean square within? If the null hypothesis is accepted, what do these quantities look like?

4. Describe the assumptions for ANOVA, and explain how they may be diagnosed.

Making Research Decisions
5. Suggest situations where the researcher should be more concerned with Type II error than with Type I error.
   a. How can the probability of a Type I error be reduced? A Type II error?
   b. How does practical significance differ from statistical significance?
   c. Suppose you interview all the members of the freshman and senior classes and find that 65 percent of the freshmen and 62 percent of the seniors favor a proposal to send Help Centers offshore. Is this difference significant?
   d. A company has three categories of marketing analysts: (1) with professional qualifications but without work experience, (2) with professional qualifications and with work experience, and (3) without professional qualifications but with work experience. A study exists that measures each analyst's motivation level (classified as high, normal, and low). A hypothesis of no relation between analyst category and motivation is to be tested.
   e. A company has 24 salespersons. The test must evaluate whether their sales performance is unchanged or has improved after a training program.

6. You conduct a survey of a sample of 25 members of this year's graduating marketing students and find that the average GPA is 3.2. The standard deviation of the sample is 0.4. Over the last 10 years, the average GPA has been 3.0. Is the GPA of this year's students significantly different from the long-run average? At what alpha level would it be significant?

7. You are curious about whether the professors and students at your school are of different political persuasions, so you take a sample of 20 professors and 20 students drawn randomly from each population. You find that 10 professors say they are conservative and 6 students say they are conservative. Is this a statistically significant difference?

8. You contact a random sample of 36 graduates of Western University and learn that their starting salaries averaged $28,000 last year. You then contact a random sample of 40 graduates from Eastern University and find that their average starting salary was $28,800. In each case, the standard deviation of the sample was $1,000.
   a. Test the null hypothesis that there is no difference between average salaries received by the graduates of the two schools.
   b. What assumptions are necessary for this test?
10 A random sample of students is interviewed to determine if there is an association between class and attitude toward corporations. With the following results, test the hypothesis that there is no difference among students on this attitude.

<table>
<thead>
<tr>
<th>Class</th>
<th>Favorable</th>
<th>Neutral</th>
<th>Unfavorable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freshmen</td>
<td>100</td>
<td>50</td>
<td>70</td>
</tr>
<tr>
<td>Sophomores</td>
<td>80</td>
<td>60</td>
<td>70</td>
</tr>
<tr>
<td>Juniors</td>
<td>50</td>
<td>50</td>
<td>80</td>
</tr>
<tr>
<td>Seniors</td>
<td>40</td>
<td>60</td>
<td>90</td>
</tr>
</tbody>
</table>

Test the hypothesis that there are no differences between the means of these products ($\alpha = .05$).

11 You do a survey of marketing students and liberal arts school students to find out how many times a week they read a daily newspaper. In each case, you interview 100 students. You find the following:

- Marketing students: $\overline{X}_m = 4.5$ times per week, $S_m = 1.5$
- Liberal arts students: $\overline{X}_n = 5.6$ times per week, $S_n = 2.0$

Test the hypothesis that there is no significant difference between these two samples.

12 One-Koat Paint Company has developed a new type of porch paint that it hopes will be the most durable on the market. The R&D group tests the new product against the two leading competing products by using a machine that scrubs until it wears through the coating. One-Koat runs five trials with each product and secures the following results (in thousands of scrubs):

<table>
<thead>
<tr>
<th>Trial</th>
<th>One-Koat</th>
<th>Competitor A</th>
<th>Competitor B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>37</td>
<td>34</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>19</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>34</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>28</td>
<td>31</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>29</td>
<td>27</td>
<td>20</td>
</tr>
</tbody>
</table>

Test the hypothesis that there is no difference between the means of these products ($\alpha = .05$).

a. Test the hypothesis that there is no difference between the means of the retailers ($\alpha = .05$).

b. Select a multiple comparison test, if necessary, to determine which groups differ in mean sales ($\alpha = .05$).

From the Headlines

14 Researchers at the University of Aberdeen found that when people were asked to recall past events or imagine future ones, the participants' bodies subliminally acted out the metaphors we commonly conceptualized with the flow of time. With past years, the participants leaned backward, while when imagining the future, they leaned forward. The leanings were small, but the directionality was clear and dependable. Using this research as a base, if you have two groups (group A holds a cup of hot coffee, and group B holds iced coffee), what statistical hypothesis would you propose to test the groups' perceptions of the personality of an imaginary individual holding coffee based on its temperature?

Proofpoint: Capitalizing on a Reporters Love of Statistics.

Yahoo! Consumer Direct Marries Purchase Metrics to Banner Ads