CHAPTER 17

Hypothesis Testing

Learning Objectives

After reading this chapter, you should understand . . .

1. The distinction between the two approaches to hypothesis testing.
2. The distinction between a statistically significant difference and one that is of practical importance for a manager.
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.

Bringing Research to Life

"Are you and Dorrie entertaining ghosts and goblins tonight?" Myra asked as she and Jason wrapped up their meeting.

"Absolutely. Our new neighborhood has a tradition of elaborate celebrations. I've been assigned the task of telling scary stories near the cider and donut table."

"You're too much the scientist to believe in things that go bump in the night. What story will you tell to the diminutive apparitions toting trick-or-treat bags?"

"Well back in my grad student days, a fellow student, Don Ticker, was a constant annoyance. No matter how often I explained to him, he never caught on to the idea of statistical post hoc analysis . . ."

"You're sure to be a hit with diminutive mermaids and aspiring Barneys," drawled Myra sarcastically. "What does this have to do with Halloween?"

Ignoring her interruption, Jason continued. "Don would take a set of data with 100 interval variables, say, and calculate the Pearson correlation coefficients among all the variables, which leaves you with 4,950 actual computations for the statistics program to print out—100 variables times 99 then divide by two (because the correlation of *X* with *Y* is the same as *Y* with *X*). The program will give you 4,950 Pearson correlation coefficients, therefore 4,950 statistical tests of significance to interpret."

"If you say so, Jason."

He saw her interest was flagging. "Do you follow me so far? Whether a relationship is statistically significant is a very big issue."

"No doubt it is . . . to a statistician. Your point is what, Jason? Something to do with a Halloween story, remember?"

"My point is, if you instruct your statistics program to flag all the results that appear to be significant at the level of $p = .05$, on average 250 of the calculations will be flagged as significant, though they may signify nothing more than sampling error."

"By now you've surely lost your 3- to 5-year-old audience. Get on with the Halloween story, would you?"

"In a moment. Choosing .05 as the probability level of statistical significance means you have consciously decided to risk that .05 or 1/20 of the results flagged as significant may actually be spurious. But you would have no way of telling which of the flagged results were spurious and which were due to real effects.

"Ticker would take his Day-Glo™ yellow marker and circle all the ostensibly significant results, and whip himself into a frenzy. He would grab his printout and run down the hallway with the printout streaming along behind, whooping and wanting to show his results around. But, of course, if he had preplanned his analysis, based on logical hypotheses, he would have saved countless hours and reams of computer paper."

"How all this relates to a scary Halloween story, I fail to see."

"It was on the afternoon before Halloween that I was about to break through to a solution on my own research when I was jolted by one of Don's dashes down the hall, breaking my attention. I reviewed all of my favorite writers of the bizarre, Edgar Allan Poe, H. P. Lovecraft, Stephen King, all masters of the macabre, but found no suitable punishment for the untimely interruption.

"That afternoon, late, the sky was dark, Myra, and there was a chill. Later, there was a moon but it had hid itself behind leaden clouds, and when it appeared, it threw shadows of the most bizarre shapes that I would only reveal to a counselor. No matter how you try, on that one night of the year, you had to wonder if retributive justice were not available. For it was Halloween."

"The truth is," Myra said, "it was just a little bit of fun for a graduate student with not much of a life, it was just a bit of a tingle to believe that on this night powerful

things were loose in the world, and that the Don Tickers of the world, with their post hoc analyses and spurious results, might encounter one of them . . . "

"Quite right. On arriving home, I tumbled onto the sofa, intending to catch a few winks. Yet I descended into one of those oppressively lucid dreams in which one knows one is dreaming but cannot gain control or float upward to consciousness. I was back in Burley Hall confronting a specter dressed in a black shroud, his face covered by a black veil. He was tall, at least seven feet, and he held a shepherd's crook, which he wielded effortlessly despite its length and thickness.

"Down the hall came Donald Ticker. His printout trailed behind him, by 25 or 30 feet. The specter thrust his shepherd's crook between Ticker's knees and hooked the near knee. Ticker became fully airborne six feet off the floor.

"At this point, I snapped awake. It was 6 A.M. I was cold, but sweating. I was filled with trepidation, not reassured by rationality, so I sped off to Burley Hall.

"Even at 6:15 Burley Hall is a busy place, what with students who have stayed up all night in the computer lab and the assistant professors eager to arrive before the senior faculty. The premonition was awful but the guilt was worse. Surely post hoc data analysis was a practice to be discouraged but a venial sin at worst.

"A knot of students and junior faculty were gathered outside my cubbyhole. To one side leaned the crook as I remembered it, long and sturdy."

"Was it Ticker's mangled corpse?"

"No, but there in the carpet, outside my office, exactly where the dream took place . . . right there . . . as clearly as could be, we could see embossed in the carpet the face—the nose and eyebrows, the lips, even—and two handprints—all 10 digits showing distinctly—of. Don Ticker, imprinted in the carpet of Burley Hall! And from that day to this, Ticker has not reappeared."

Myra's pupils had dilated, revealing only a narrow rim of blue iris.

Jason chuckled. "Got ya, didn't I?" he asked. "What do you think of my story, Myra?"

"You are a lying little sadist, Jason, and a prevaricating sociopath. But a word of advice: You may have captured my attention toward the end, but resurrect a Stephen King tale and save the little ones from boredom!"

Introduction

Induction and deduction were discussed in Chapter 2.

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 an important foundation 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 15 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 7, but we will return to it here briefly. The second, testing statistical hypotheses, is the primary subject of this chapter.

In the next few sections, we will refresh your memory of hypothesis testing and look at selected statistical tests. Many are basic, but they illustrate the diverse types of data and situations a researcher may encounter. A section on nonparametric techniques in Appendix E provides further study for readers with a special interest in nominal and ordinal variables.

Hypothesis Testing

Having 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.

Testing Approaches

There are two approaches to hypothesis testing. The more established is the classical or sampling-theory approach; the second is known as the Bayesian 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.

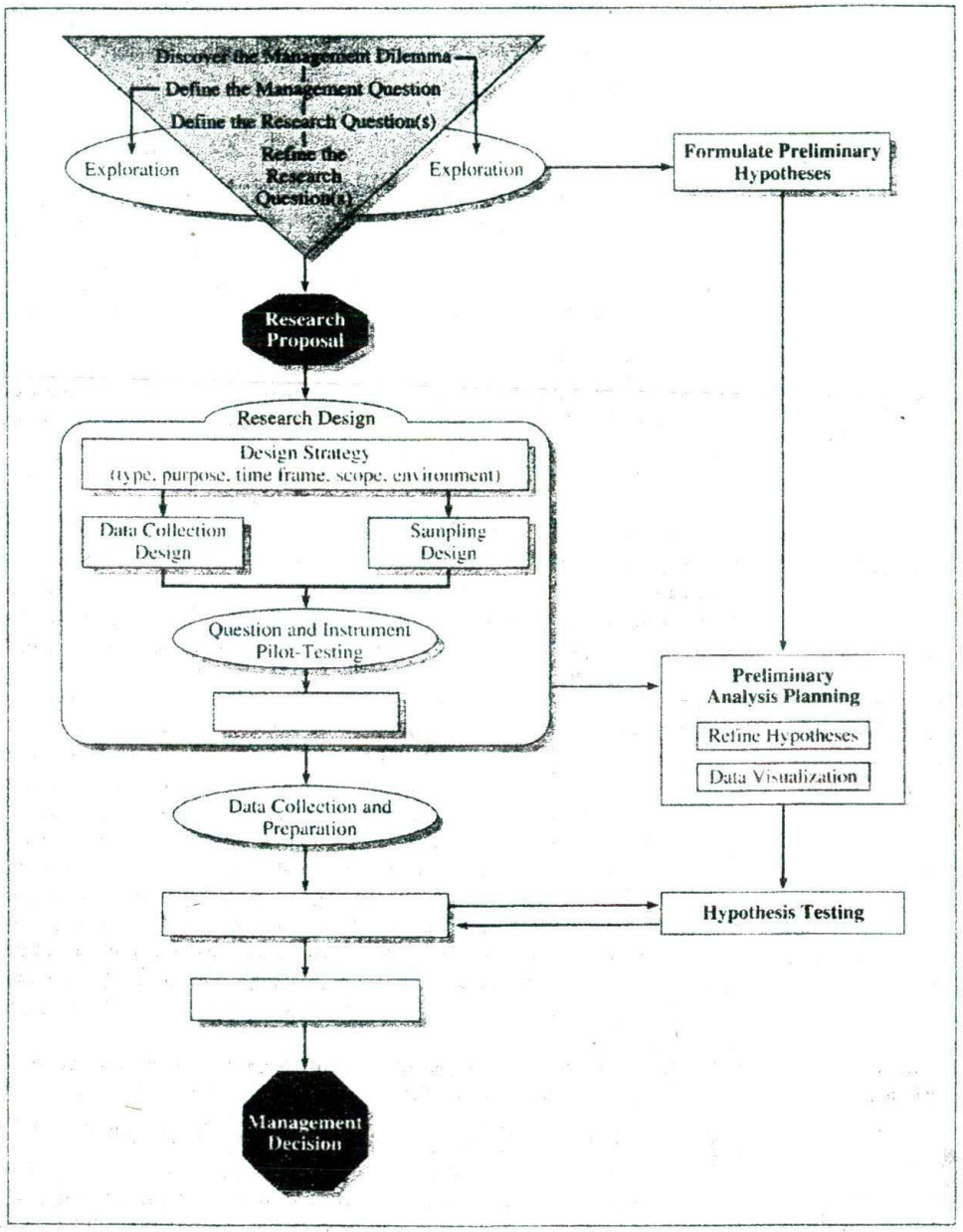
An example of Bayesian decision making is presented in Appendix B on the topic of valuing research information. The reader interested in learning more about Bayesian statistics is referred to the suggested readings at the end of this chapter.

Bayesian statistics are an extension of the classical approach. They also use sampling data for making decisions, but they go beyond them 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. They are expressed as a prior distribution that can be revised after sample information is gathered. The revised estimate, known as a *posterior distribution,* may be further revised by additional information, and so on. 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. The Bayesian approach, based on the centuries-old Bayes theorem, has emerged as an alternative hypothesis-testing procedure since the mid-1950s.

Statistical Significance

Following the sampling-theory 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 these 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, the controller of e-WEAR, an e-commerce division of a large retail chain, may be concerned about a possible slowdown in payments by the company's customers. She measures the rate of payment in terms of the average age of receivables outstanding. Generally, the company has maintained an average of about 50 days with a standard deviation of 10 days. Suppose the controller has all of the customer accounts analyzed and finds the average is now 51 days. Is this difference statistically significant from 50?

EXHIBIT 17–1 Hypothesis Testing and the Research Process



Of course it is, because the difference is based on a census of the accounts and there is no sampling involved. It is a fact that the population average has moved from 50 to 51 days. While it is of statistical significance, whether it is of **practical significance** is another question. If the controller 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 e-WEAR's receivables frequently, we normally resort to sampling. Assume a sample of 25 accounts is randomly selected and the average number of days outstanding is calculated to be 54. Is this statistically significant? The answer is not obvious. It is significant if there is good reason to believe the average age of the total group of receivables has moved up from 50. 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 judge 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 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. It would not be surprising to flip a fair coin 100 times and observe 52 heads. 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 e-WEAR example, the null hypothesis states that the population parameter of 50 days has not changed. A second, **alternative hypothesis** holds that there has been a change in average days outstanding (i.e., the sample statistic of 54 indicates the population value probably is no longer 50). The alternative hypothesis is the logical opposite of the null hypothesis.

The e-WEAR example can be explored further to show how these concepts are used to test for significance.

MANAGEMENT



- The null hypothesis (H_0) is: There has been no change from the 50 days average age of accounts outstanding.

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 age of accounts has changed from 50 days.
- The average age of receivables has increased (decreased) from 50 days.

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 50 days, or it could be less than 50 days. 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. Hypotheses for the example may be expressed in the following form:

Null	H_0:μ	=	50 days
Alternative	H_A:μ	≠	50 days (not the same case)
or	H_A:μ	>	50 days (greater than case)
or	H_A:μ	<	50 days (less than case)

MANAGEMENT Tip

In testing these hypotheses, adopt this decision rule: Take no corrective action if the analysis shows that one cannot reject the null hypothesis. Note "not to 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.[1] 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 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

EXHIBIT 17–2 One- and Two-Tailed Tests at the 5% Level of Significance

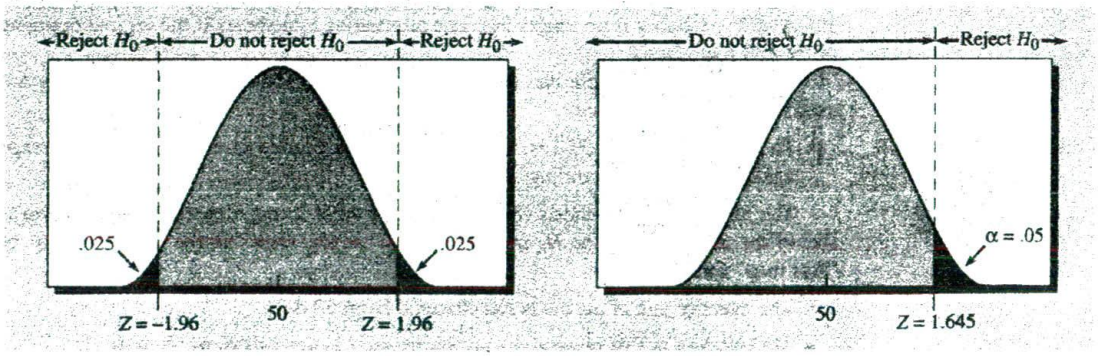
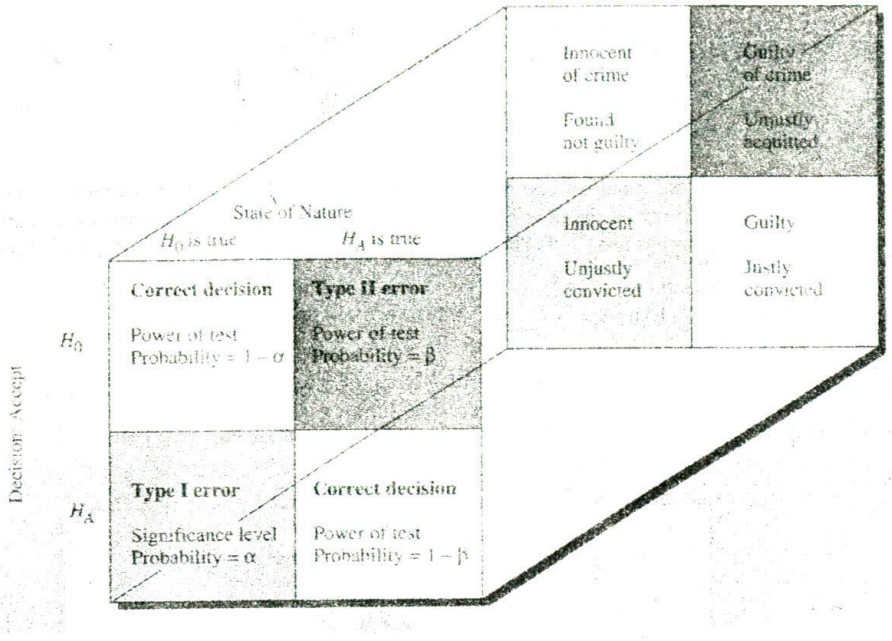


EXHIBIT 17–3 Comparison of Statistical Decisions to Legal Analogy



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.

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** (α) is committed, a true null hypothesis is rejected; the innocent person is unjustly convicted. The α value is called the level of significance and is the probability of rejecting the true null. With a **Type II error** (β), 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 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 e-WEAR controller's problem is deciding whether the average age of accounts receivable has changed. Assume the population mean is 50 days, the standard deviation of the population is 10 days, and the size of the sample is 25 accounts.

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}}$$

$$-1.96 = \frac{\bar{X}_c - 50}{2}$$

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

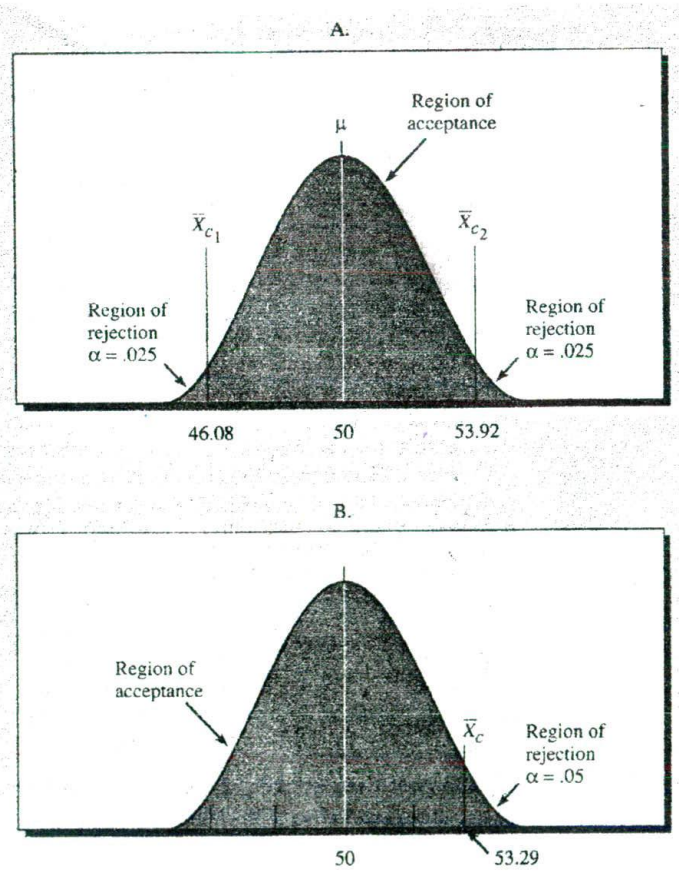
$$1.96 = \frac{\bar{X}_c - 50}{2}$$

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

With this information, one can calculate the standard error of the mean ($\sigma_{\bar{x}}$ (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 days.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$$

EXHIBIT 17–4 Probability of Making a Type I Error Given H_0 Is True



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 (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 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,[2] where

$Z = 1.96$ (significance level = .05)

$\bar{X}_c$ = The critical value of the sample mean

μ = The population value stated in $H_0 = 50$

$\sigma_{\bar{X}}$ = The standard error of a distribution of means of samples of 25

The probability of a Type I error is

$$\alpha = .05, \text{ or } 5\%$$

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 α 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 controller may be interested only in increases in the age of receivables. 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 α risk and compute a new critical value (X_c). (See Appendix Table F-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 $\bar{X}_c$.

$$Z = 1.645 = \frac{\bar{X}_c - 50}{2}$$

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

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 controller would commit a Type II error (β) 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 β error depends on five factors: (1) the true value of the parameter, (2) the α 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 β error if the new β moves from 50 to 54 than if it moves only to 52. We must compute separate β error estimates for each of a number of assumed new population parameters and $\bar{X}_c$ values.

To illustrate, assume μ 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? This may be expressed in the following fashion:

$P(A_2|S_1) = \alpha = .05$ (assume a one-tailed alternative hypothesis)

$P(A_1|S_2) = \beta = ?$

If the new μ is 54, then

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$$

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{53.29 - 54}{2} = -.355$$

Using Table F-1 in Appendix F, we interpolate between .35 and .36 Z scores to find the .355 Z score. The area between the mean and Z is .1387. β is the tail area, or the area below the Z, and is calculated as

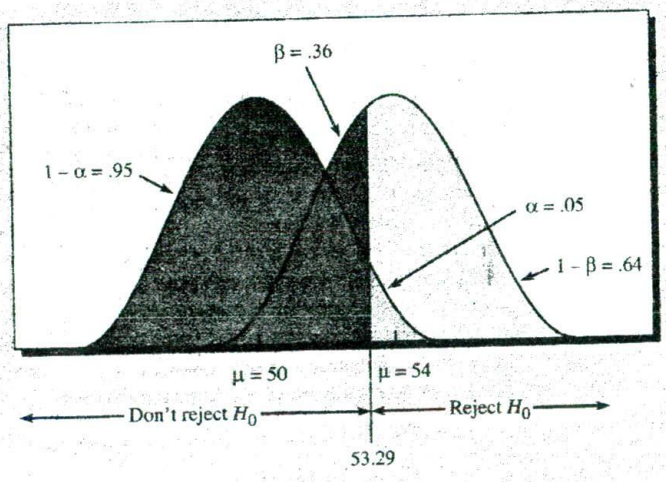
$$\beta = .50 - .1387 = .36$$

This condition is shown in Exhibit 17–5. With an α of .05 and a sample of 25, there is a 36 percent probability of a Type II (β) error if the μ 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 usually needed.

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

EXHIBIT 17–5 Probability of Making a Type II Error



$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1$$

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{53.29 - 54}{1} = -.71$$

$$\beta = .50 - .2612 = .24$$

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 α and β 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.** While the researcher is usually interested in testing a hypothesis of change or differences, the null hypothesis is always used for statistical testing purposes.

MANAGEMENT Tip

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 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 α levels such as .10, .025, or .001 are sometimes chosen. The exact level to choose is largely determined by how much α risk one is willing to accept and the effect that this choice has on β risk. The larger the α, the lower is the β.
4. **Compute the calculated difference value.** After the data are collected, use the formula for the appropriate significance test to obtain the calculated value.
5. **Obtain the critical test value.** After we compute the calculated t, χ^2, or other measure, we must look up the critical value in the appropriate table for that distribution. 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 (α), 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 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 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 where the controller was interested in determining whether the average age of accounts receivable had increased. The critical value of 53.29 was computed based on a standard deviation of 10, sample size of 25, and the contoller's willingness to accept a 5 percent α 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

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

$$Z = \frac{55 - 50}{2}$$

$$Z = 2.5$$

The *p* value is determined using the standard normal table. The area between the mean and a *Z* value of 2.5 is .4938. The *p* value is the area above the *Z* value (shown in Part B of Exhibit 17–4). The probability of observing a *Z* value at least as large as 2.5 is only .0062 (.5000 – .4932 = .0062) if the null hypothesis is true.

This small *p* value represents the risk of rejecting the 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 controller can conclude that the average age of the accounts receivable 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-

SNAPSHOT

Content Analysis Unreels Drug Use in Movies

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 Video Software Dealers 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* was noted when evidence of materials or paraphernalia was noted 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.americacares.org

www.vsda.org

sample, two-sample, and *k*-sample cases. Readers needing a comprehensive treatment of significance tests are referred to the suggested readings at the end of 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:

MANAGEMENT Tip

- 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.[4] 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 "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.[5]

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.

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 homogeneity 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 some 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

MANAGEMENT



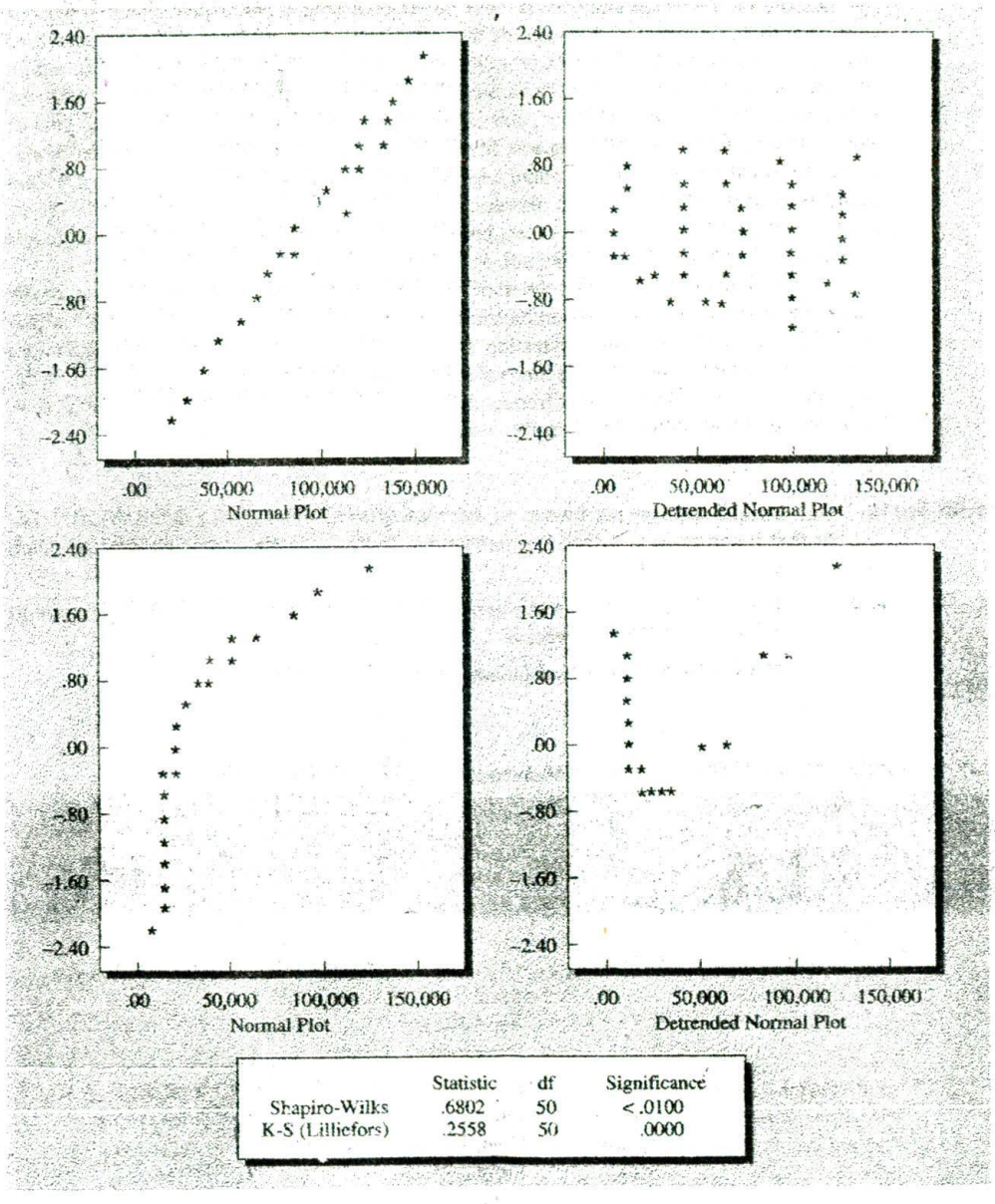
In attempting to choose a particular significance test, the researcher should consider at least three questions:

- Does the test involve one sample, two samples, or k 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, most 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

EXHIBIT 17–6 Probability Plots and Tests of Normality



	Statistic	df	Significance
Shapiro-Wilks	.6802	50	< .0100
K-S (Lilliefors)	.2558	50	.0000

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. Over 130 solutions to data analysis problems are paired with commonly asked questions.[6]

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, covers eight 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.

In this chapter, we used the above three criteria to develop a classification of the major parametric and nonparametric tests and measures. This is shown in Exhibit 17–7.[8] For example, if 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 E.

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?

EXHIBIT 17–7 Recommended Statistical Techniques by Measurement Level and Testing Situation

Measurement Level	*One-Sample Case*	*Two-Samples Case*		*k-Samples Case*	
		Related Samples	Independent Samples	Related Samples	Independent Samples
Nominal	• Binomial • χ^2 One-sample	• McNemar	• Fisher exact test • χ^2 Two-samples test	• Cochran Q	• χ^2 for *k* samples
Ordinal	• Kolmogorov-Smirnov one-sample test • Runs test	• Sign test • Wilcoxon matched-pairs test	• Median test • Mann-Whitney U • Kolmogorov-Smirnov • Wald-Wolfowitz	• Friedman two-way ANOVA	• Median extension • Kruskal-Wallis one-way ANOVA
Interval and ratio	• *t*-test • *Z* test	• *t*-test for paired samples	• *t*-test • *Z* test	• Repeated-measures ANOVA	• One-way ANOVA • *n*-way ANOVA

- 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 (μ)?

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

Parametric Tests The Z 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 σ; 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 income taxes collected this year to last year's income tax revenues.

Example. To illustrate the application of the *t*-test to the one-sample case, consider again the controller's problem mentioned earlier. With a sample of 100 accounts, she finds that the mean age of outstanding receivables is 52.5 days, with a standard deviation of 14. Do these results indicate the population mean might still be 50 days?

In this problem, we have only the sample standard deviation (s). This must be used in place of the population standard deviation (σ). When we substitute s for σ, 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: = 50 days.
   H_A: > 50 days (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 customer accounts.
3. **Significance level.** Let $\alpha = .05$, with $n = 100$.
4. **Calculated value.**

$$t = \frac{52.5 - 50}{14/\sqrt{100}} = \frac{2.5}{1.4} = 1.786; \qquad \text{d.f.} = n - 1 = 99$$

5. **Critical test value.** We obtain this by entering the table of critical values of t (see Appendix Table F–2 at back of 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 Table F–2).
6. **Interpret.** 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 accounts receivable outstanding has increased.

Nonparametric Tests A variety of nonparametric tests may be used in a one-sample situation, depending on the measurement scale used and other conditions. If the measurement scale is nominal (classificatory only), it is possible to use either the binomial test or the chi-square (χ^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 sample is so small that the χ^2 test cannot be used.

Chi-Square Test Probably the most widely used nonparametric test of significance is the **chi-square (χ^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 χ^2 is the measure that expresses the extent of this difference. The larger the divergence, the larger is the χ^2 value.

The formula by which the χ^2 test is calculated is

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

in which

O_i = Observed number of cases categorized in the ith category.

E_i = Expected number of cases in the ith category under H_0.

k = The number of categories.

There is a different distribution for χ^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-sample or k-sample 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 χ^2 test appropriate. When d.f. = 1, each expected frequency should be at least 5 in size. If d.f. > 1, then the χ^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 7 is taken. We have interviewed 200 students and learned of their intentions to join such a 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. Do these variations indicate there is a significant difference among these students, or are these sampling variations only? Proceed as follows:

Living Arrangement	Intend to Join	Number Interviewed	Percent (No. Interviewed/200)	Expected Frequencies (Percent × 60)
Dorm/fraternity	16	90	45	27
Apartment/rooming house, nearby	13	40	20	12
Apartment/rooming house, distant	16	40	20	12
Live at home	15	30	15	9
Total	60	200	100	60

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 χ^2 to compare the observed distribution to a hypothesized distribution. The χ^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:

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

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

$$= 9.89$$

$$\text{d.f.} = (4-1)(2-1) = 3$$



SNAPSHOT Ramping Up Skate Sales

If you were trying to compete with RollerBlade, a brand synonymous with the in-line skating phenomenon, you'd be tempted to challenge conventional wisdom in seeking a pathway to success. That's what K2Sports's marketing guru Amy Buckalter did when she discontinued the brand's advertising program.

In-line skates are traditionally sold in sporting goods stores, an environment characterized by high employee turnover. K2Sports developed a national ambassador program, designed to demonstrate skate sales techniques to young, part-time salespeople. The teach-by-example ambassador program started by sending young skating enthusiasts to sporting goods retailers in southern California. These ambassadors essentially serve as part-time employees for a relatively short period of time in any one store. On days when an ambassador is in the store, a retailer's sales of in-line skates triple, with no less than 70 percent of skates sold being the K2Sports brand. As the ambassador program rolls nationwide, the impressive sales results continue.

Is the repeated sales success of the ambassador program an experiment that proves a causal relationship between K2Sports's ambassador program and retailers' sales increase?

www.k2sports.com

5. **Critical test value.** Enter the table of critical values of χ^2 (see Appendix Table F–3), with 3 d.f., and secure a value of 7.82 for $\alpha = .05$.
6. **Interpret.** The calculated value is greater than the critical value, so the null hypothesis is rejected.

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 output methods from two production lines or the price movements of common stock from two samples could be compared. A study of worker productivity from two groups or different samples from a public opinion poll would also use this method.

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

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

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

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)0}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

SNAPSHOT

Ergonomics or Economics: Toothbrushes Go on Steroids

Toothbrush design underwent radical change when Colgate introduced its Wave brush in 1997. Once narrow, straight handled appliances that fit the ubiquitous toothbrush holder in every bathroom, they are now "on steroids," too fat to fit the holders. Tom Mintel, associate director of toothbrush development at Colgate-Palmolive Co., said its research revealed that some people change their grip more than 100 times during brushing. To accommodate the need for improved ergonomics, Colgate's research found that thicker handles made teeth cleaning significantly easier and increased brushing time.

Procter & Gamble Co. makes a whale-shaped brush while Johnson & Johnson's Reach model, with a thick handle, is marketed to youngsters. In 1998 Gillette Co.'s Oral-B division released a product aimed at a segment it refers to as "power grippers," composed of 28 percent of the population. Its videotaped research in family bathrooms showed differences among styles of brushing. It seems that many people don't place their thumb on the handle but grip it like a hammer. For this group, the ultrawide Cross Action was designed.

So what do people do with large toothbrushes that won't fit the traditional holder? Although Colgate received over a thousand calls about holder problems the first year its Wave was introduced, Robert Blanchard, head of Colgate's toothbrush division, said that follow-ups showed fewer defections than expected because most people won't sacrifice dental health for a holder. David Weber, Oral-B's head of research, claims that the holder problem applies to only 19 percent of U.S. households—particularly in the Northeast and older sections of the country where holders are built in. Approximately 34 percent keep them in countertop holders, according to Oral-B. Until the holder manufacturers adjust, the rest of us will have to keep those Leviathans in cups, drawers, and medicine cabinets.

www.wsj.com

www.colgate.com

www.oralb.com

www.pg.com

www.jnj.com

where

$(\mu_1 - \mu_2)$ is the difference between the two population means

S_p^2 is associated with the pooled variance estimate:

$$S_p^2 = \frac{(n_1 - 1)\,S_1^2 + (n_2 - 1)\,S_2^2}{n_1 + n_2 - 2}$$

To illustrate this application, consider a problem that might face a manager at Dean Merrill Brokerage who 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:

	A Group	B Group
Average hourly sales	$\bar{X}_1 = \$1{,}500$	$\bar{X}_2 = \$1{,}300$
Standard deviation	$s_1 = 225$	$s_2 = 251$

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_A: 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.**

$$t = \frac{(1{,}500 - 1{,}300) - 0}{\sqrt{\frac{(10)(225)^2 + (10)(251)^2}{20}\left(\frac{1}{11} + \frac{1}{11}\right)}}$$

$$= \frac{200}{101.63} = 1.97, \qquad \text{d.f.} = 20$$

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 Table F–2 with d.f. = 20, one-tailed test, $\alpha = .05$. The critical value is 1.725.
6. **Interpret.** 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 (χ^2) test is appropriate f 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 χ^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:

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

in which

O_{ij} = Observed number of cases categorized in the ijth cell.

E_{ij} = Expected number of cases under H_0 to be categorized in the ijth cell.

Suppose Containers Inc. is implementing a smoke-free workplace policy and is interested in whether smoking affects worker accidents. Since the company has complete reports of on-the-job accidents, a sample of names of workers were drawn 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. Members of both groups are interviewed to determine if they are smokers or not. The results appear in the table at the top of the next page.

The expected values have been calculated and are shown. The testing procedure is:

1. **Null hypothesis.** H_0: There is no difference in on-the-job accident occurrences between smokers and nonsmokers.
   H_A: There is a difference in on-the-job accident occurrences between smokers and nonsmokers.

		On-the-Job Accident		
Smoker	Count Expected values	Yes	No	Row Total
	Heavy	12 8.24	4 7.75	16
	Moderate	9 7.73	6 7.27	15
	Nonsmoker	13 18.03	22 16.97	35
	Column Total	34	32	66

2. **Statistical test.** χ^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 classes. 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,

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

$$\chi^2 = \frac{(12 - 8.24)^2}{8.24} + \frac{(4 - 7.75)^2}{7.75} + \frac{(9 - 7.73)^2}{7.73} + \frac{(6 - 7.72)^2}{7.772} + \frac{(13 - 18.03)^2}{18.03} + \frac{(22 - 16.97)^2}{16.97}$$

$$= 6.86$$

5. **Critical test value.** Enter Appendix Table F–3 and find the critical value 5.99 with $\alpha = .05$ and d.f. $= 2$.

6. **Interpret.** 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 caution that expected frequencies below 5 should not compose more than 20 percent of the cells, and no cell should have an E_i of less than 1. Some research has argued that these restrictions are too severe.[9]

In another type of χ^2, the 2×2 table, a correction known as *Yates' correction for continuity* is often 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. The formula for this correction is

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

where the letters represent the cells designated as

A	B
C	D

When the continuity correction is applied to the data shown in Exhibit 17–8, a χ^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 χ^2 without correction, the value would have been 6.25, which has an observed level of significance of .01242. Some researchers may be tempted to reject the null at this level. (But note that the critical value of χ^2 at .01 with 1 d.f. is 6.64. See Appendix Table F–3.) The literature is in conflict regarding the merits of Yates' correction, but this example suggests one should take care when interpreting 2×2 tables.[10] 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, so it does not apply; the latter, based on maximum likelihood theory, produces results similar to Pearson's chi-square.

Two Related Samples Test

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 output of specific workers before and after vacations, the performance of the same stocks at two intervals, or the effects of an experimental stimulus when persons were

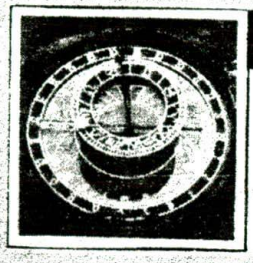
EXHIBIT 17–8 Comparison of Corrected and Noncorrected Chi-Square Results Using SPSS Procedure Crosstab

INCOME BY POSSESSION OF CPA

INCOME	Count	CPA Yes 1	CPA No 2	Row Total
	High 1	30	30	60 60.0
	Low 2	10	30	40 40.0
	Column Total	40 40.0	60 60.0	100 100.0

Chi-Square	Value	D.F.	Significance
Pearson	6.25000	1	.01242
Continuity Correction	5.25174	1	.02192
Likelihood Ratio	6.43786	1	.01117
Mantel-Haenszel	6.18750	1	.01287

Minimum Expected Frequency: 16.000



SNAPSHOT

So, You Are a Gemini? Maybe I Should Drive.

Star signs are religiously followed by some and are a source of amusement to others. In a lighthearted study conducted by the Sydney-based insurance company Suncorp Metway, the number of car accident claims over a three-year period were compared with star signs. More than 14,500 drivers born between May 21 and June 21 (Geminis) had crashed their cars. Warren Duke, national manager of personal insurance, said, ". . . it was interesting that Geminis, typically described as restless, easily bored, and frustrated by things moving slowly, had more car accidents than any other sign." Those with the fewest were Capricorns, said to be patient and careful. Women also had more claims than men in 2001, according to the study. Duke added, "Women make significantly fewer claims than men until their late twenties, but after that women, aged 29 and over, edge ahead of men, making slightly more claims." Suncorp Metway has no intention of using astrology as a rating factor in determining a customer's motor insurance premium. But it had fun looking for trends. How might you construct a chi-square test of star sign by gender? What variables would you use as controls?

Most likely to file an accident claim by star sign

1. Gemini (May 21–June 21)
2. Taurus (April 20–May 20)
3. Pisces (February 19–March 20)
4. Virgo (August 23–September 22)
5. Cancer (June 22–July 22)
6. Aquarius (January 20–February 18)
7. Aries (March 21–April 19)
8. Leo (July 23–August 22)
9. Libra (September 23–October 22)
10. Sagittarius (November 22–December 21)
11. Scorpio (October 23–November 21)
12. Capricorn (December 22–January 19)

www.suncorpmetway.com.au

randomly assigned to groups and given pretests and post-tests. Both parametric and nonparametric tests are applicable under these conditions.

Parametric Tests The *t*-test for independent samples would normally be inappropriate for this situation because one of its assumptions is that observations are independent. This problem is solved by a formula where 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 α 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

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

$$S_D = \sqrt{\frac{\Sigma D^2 - \frac{(\Sigma D)^2}{n}}{n-1}}$$

To illustrate this application, we use two years of *Forbes* sales data (in millions of dollars) from 10 companies found in Exhibit 17–9.

1. **Null hypothesis.** H_0: $\mu = 0$; there is no difference between the two years' sales records. H_A: $\neq 0$; there is a difference between sales for Year 1 and Year 2.
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.
3. **Significance level.** Let $\alpha = .01$, with $n = 10$ and d.f. $= n - 1$.
4. **Calculated value.**

$$t = \frac{\bar{D}}{S_D/\sqrt{n}} = \frac{3{,}578.10}{570.98} = 6.28; \quad \text{d.f.} = 9$$

5. **Critical test value.** Enter Appendix Table F–2, with d.f. = 9, two-tailed test, $\alpha = .01$. The critical value is 3.25.
6. **Interpret.** Since the calculated value is greater than the critical value (6.28 > 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. 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.

EXHIBIT 17–9 Sales Data for Paired-Samples *t*-Test (dollars in millions)

Company	Sales Year 2	Sales Year 1	Difference D	D^2
GM	126932	123505	3427	11744329
GE	54574	49662	4912	24127744
Exxon	86656	78944	7712	59474944
IBM	62710	59512	3192	10227204
Ford	96146	92300	3846	14791716
AT&T	36112	35173	939	881721
Mobil	50220	48111	2109	4447881
DuPont	35099	32427	2632	6927424
Sears	53794	49975	3819	14584761
Amoco	23966	20779	3187	10156969
Totals			$\Sigma D = 35781$	$\Sigma D^2 = 157364693$

EXHIBIT 17–10 SPSS Output for Paired-Samples *t*-Test

---*t*-tests for paired samples---

Variable	Number of Cases	Mean	Standard Deviation	Standard Error
Year 2 Sales	10	62620.9	31777.649	10048.975
Year 1 Sales	10	59038.8	31072.871	9836.104

(Difference Mean)	Standard Deviation	Standard Error	Corr.	2-tail Prob.	*t* Value	Degrees of Freedom	2-tail Prob.
3582.1000	1803.159	570.209	.999	.000	6.28	9	.000

Test the significance of any observed change by setting up a fourfold table of frequencies to represent the first and second set of responses:

	After	
Before	Do Not Favor	Favor
Favor	*A*	*B*
Do not favor	*C*	*D*

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 the following transformation of the χ^2 test:

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

The "minus 1" in the equation is a correction for continuity since the χ^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 management decided to tell employees of the "values of teamwork" in an internal education campaign. Managers took a random sample of their employees before the campaign, asking them to complete a questionnaire on their attitudes on this topic. On the basis of their responses, the workers were divided into equal groups reflecting their favorable or unfavorable views of teamwork. 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: $P(A) = P(D)$.
   H_A: $P(A) \neq P(D)$.
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.**

$$\chi^2 = \frac{(|10 - 40| - 1)^2}{10 + 40} = \frac{29^2}{50} = 16.82; \text{ d.f.} = 1$$

	After	
Before	Unfavorable	Favorable
Favorable	$A = 10$	$B = 90$
Unfavorable	$C = 60$	$D = 40$

5. **Critical test value.** Enter Appendix Table F–3, and find the critical value to be 3.84 with $\alpha = .05$ and d.f. = 1.
6. **Interpret.** The calculated value is greater than the critical value ($16.82 > 3.84$), indicating one should reject the null hypothesis. In fact, χ^2 is so large that it would have surpassed an α of .001.

k Independent Samples Tests

In management and economic research, we often use ***k* independent samples tests** 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 factor (brands of coffee, varieties of residential housing, types of retail stores) on a continuous dependent variable. In a fixed-effects model, the levels of the factor are established in advance, and the results are not generizable 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 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-group 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:

$$F = \frac{\text{Between-groups variance}}{\text{Within-groups variance}} = \frac{\text{Mean square}_{\text{between}}}{\text{Mean square}_{\text{within}}}$$

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, is one less than the number of groups $(k - 1)$. The degrees of freedom for the denominator, the mean square within groups, is 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 populations, 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 Europe. 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,[11] are used for this and the next two examples. For the one-way

EXHIBIT 17–11 Data Table: Analysis of Variance Examples

	Flight Service					Flight Service			
	Rating 1	Rating 2	Airline	Seat Selection		Rating 1	Rating 2	Airline	Seat Selection
1	40	36	1	1	32	70	80	2	2
2	28	28	1	1	33	73	79	2	2
3	36	30	1	1	34	72	88	2	2
4	32	28	1	1	35	73	89	2	2
5	60	40	1	1	36	71	72	2	2
6	12	14	1	1	37	55	58	2	2
7	32	26	1	1	38	68	67	2	2
8	36	30	1	1	39	81	85	2	2
9	44	38	1	1	40	78	80	2	2
10	36	35	1	1	41	92	95	3	1
11	40	42	1	2	42	56	60	3	1
12	68	49	1	2	43	64	70	3	1
13	20	24	1	2	44	72	78	3	1
14	33	35	1	2	45	48	65	3	1
15	65	40	1	2	46	52	70	3	1
16	40	36	1	2	47	64	79	3	1
17	51	29	1	2	48	68	81	3	1
18	25	24	1	2	49	76	69	3	1
19	37	23	1	2	50	56	78	3	1
20	44	41	1	2	51	88	92	3	2
21	56	67	2	1	52	79	85	3	2
22	48	58	2	1	53	92	94	3	2
23	64	78	2	1	54	88	93	3	2
24	56	68	2	1	55	73	90	3	2
25	28	69	2	1	56	68	67	3	2
26	32	74	2	1	57	81	85	3	2
27	42	55	2	1	58	95	95	3	2
28	40	55	2	1	59	68	67	3	2
29	61	80	2	1	60	78	83	3	2
30	58	78	2	1					
31	52	65	2	2					

Airline: 1 = Delta; 2 = Lufthansa; 3 = KLM.

Seat Selection: 1 = Economy; 2 = Business.

All data are hypothetical.

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.

Again, we follow the procedure:

1. **Null hypothesis.** H_0: $\mu_{A1} = \mu_{A2} = \mu_{A3}$.
   H_A: 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.**

$$F = \frac{MS_b}{MS_W} = \frac{5822.017}{205.695} = 28.304, \text{ d.f. } (2, 57)$$

   See summary in Exhibit 17–12.
5. **Critical test value.** Enter Appendix Table F–9, with d.f. (2, 57), $\alpha = .05$. The critical value is 3.16.
6. **Interpret.** 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. It 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 α—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. Obviously, 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 (*a priori* because the decision was made before the test).[12]

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

$$MS_{CON} = SS_{CON} = \frac{\left(\sum_j C_j \bar{X}_j\right)^2}{\sum_j \frac{C_j^2}{n}}$$

EXHIBIT 17–12 Summary Tables for One-Way ANOVA Example

Model Summary						
Source		d.f.	Sum of Squares	Mean Square	*F* Value	*p* Value
Model	Airline	2	11644.033	5822.017	28.304	.0001
Residual	Error	57	11724.550	205.694		
Total		59	23368.583			

Factor: Airline.

Dependent: Flight Service Rating 1.

Means Table				
	Count	Mean	Std. Dev.	Std. Error
Delta	20	38.950	14.006	3.132
Lufthansa	20	58.900	15.089	3.374
KLM	20	72.900	13.902	3.108

Scheffé's S Multiple Comparison Procedure					
	Vs.	Diff.	Crit. Diff.	*p* Value	
Delta	Lufthansa	19.950	11.400	.0002	S
	KLM	33.950	11.400	.0001	S
Lufthansa	KLM	14.000	11.400	.0122	S

S = Significantly different at the .05 level.

Significance level: .05.

All data are hypothetical.

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 of the test into fewer degrees of freedom 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

EXHIBIT 17–13 Selection of Multiple Comparison Procedures

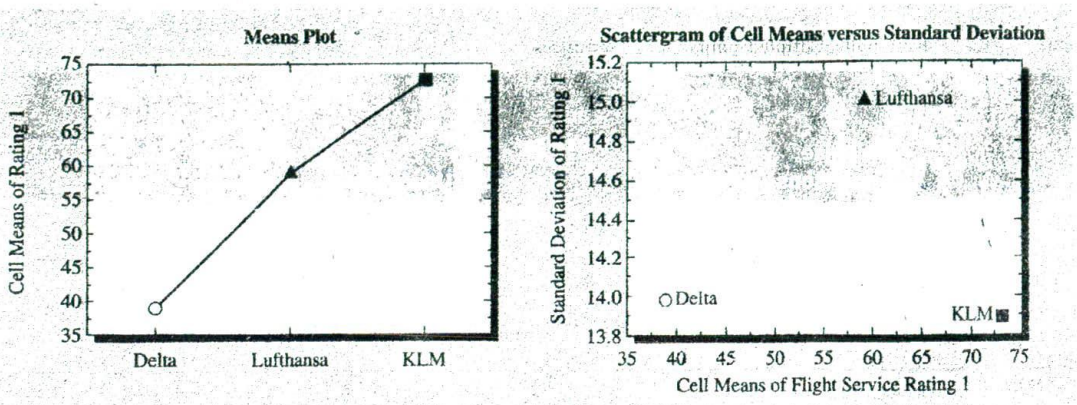
Test	Pairwise Comparisons	Complex Comparisons	Equal *n*'s Only	Unequal *n*'s	Equal Variances Assumed	Unequal Variances Not Assumed
Fisher LSD	x			x	x	
Bonferroni	x		x	x		
Tukey HSD	x		x		x	
Tukey-Kramer	x			x	x	
Games-Howell	x			x		x
Tamhane T2	x			x		x
Scheffé S		x		x	x	
Brown-Forsythe		x		x		x
Newman-Keuls	x		x		x	
Duncan	x		x		x	
Dunnett's T3						x
Dunnett's C						x

a priori contrast. However, examining the means table (Exhibit 17–12) revealed that the airline means were quite disparate. Comparisons after the results are compared require post hoc tests or pairwise **multiple comparison procedures** to determine which means differ. Range tests find homogeneous subsets of means that are not different from each other. Multiple comparisons test the difference between each pair of means and indicate significantly different group means at an α level of .05, or another level that you specify. Multiple comparison tests use group means and incorporate the MS_{error} 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, Type I or Type II error reduction, 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.[13] 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.

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 Delta and KLM passengers. Nevertheless, these two groups are sharply divided on the quality of in-flight service, and that is apparent in the upper plot.

EXHIBIT 17–14 One-Way Analysis of Variance Plots



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 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. 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 factors operate jointly. 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 for the airline factor was rejected, and seat selection was also found 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.

EXHIBIT 17–15 Summary Table for Two-Way ANOVA Example

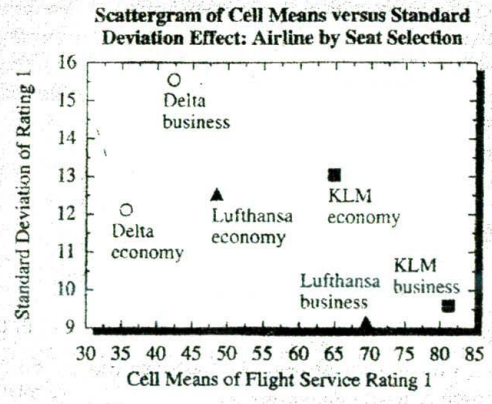
Model Summary					
Source	d.f.	Sum of Squares	Mean Square	*F* Value	*p* Value
Airline	2	11644.033	5822.017	39.178	.0001
Seat selection	1	3182.817	3182.817	21.418	.0001
Airline by seat selection	2	517.033	258.517	1.740	.1853
Residual	54	8024.700	148.606		

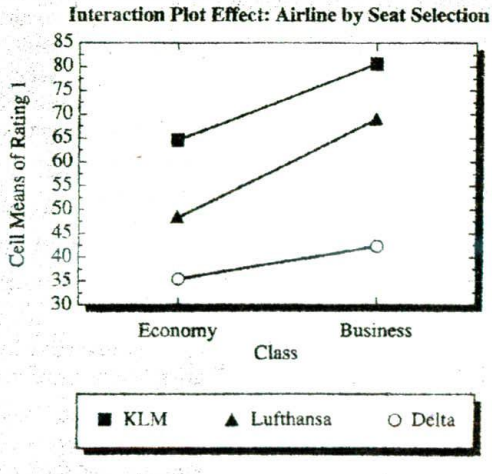
Dependent: Flight Service Rating 1.

Means Table Effect: Airline by Seat Selection				
	Count	Mean	Std. Dev.	Std. Error
Delta economy	10	35.600	12.140	3.839
Delta business	10	42.300	15.550	4.917
Lufthansa economy	10	48.500	12.501	3.953
Lufthansa business	10	69.300	9.166	2.898
KLM economy	10	64.800	13.037	4.123
KLM business	10	81.000	9.603	3.037

All data are hypothetical.

EXHIBIT 17–16 Two-Way Analysis of Variance Plots





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-sample χ^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 E.

k Related Samples Case

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 experimental 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 quality assurance, inventory, sales, and measures of human 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 a 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 × Airline: H_0: $(\mu_{R2A1} - \mu_{R2A2} - \mu_{R2A3}) = (\mu_{R1A1} - \mu_{R1A2} - \mu_{R1A3})$

   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 $\alpha = .05$ and d.f. = [airline (2, 57), ratings (1, 57), ratings by airline (2, 57)].
4. **Calculated values.** See summary in Exhibit 17–17.
5. **Critical test value.** Enter Appendix Table F–9, 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. **Interpret.** 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

EXHIBIT 17–17 Summary Tables for Repeated-Measures ANOVA

Model Summary

Source	d.f.	Sum of Squares	Mean Square	*F* Value	*p* Value
Airline	2	35527.550	17763.775	67.199	.0001
Subject (group)	57	15067.650	264.345		
Ratings	1	625.633	625.633	14.318	.0004
Ratings by air. . . .	2	2061.717	1030.858	23.592	.0001
Ratings by subj. . . .	57	2490.650	43.696		

Dependent: Flight service ratings 1 and 2.

Means Table Ratings by Airline

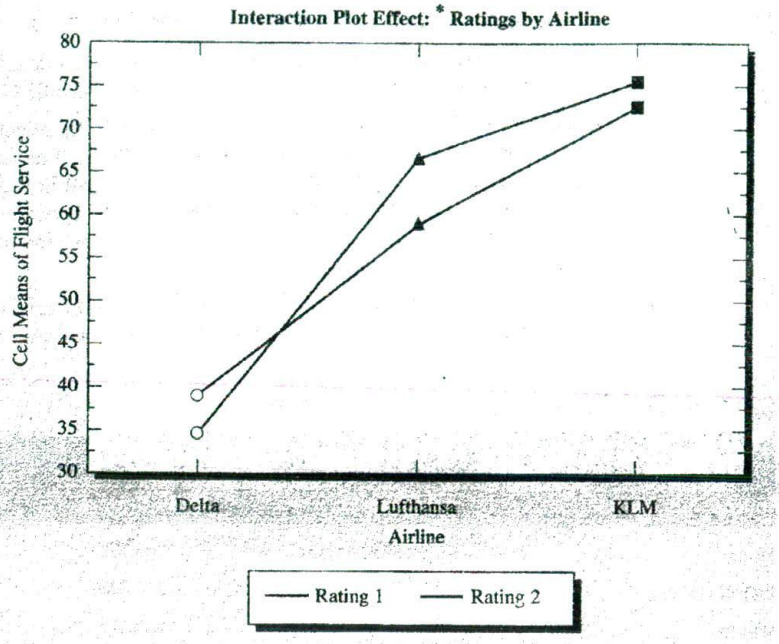
	Count	Mean	Std. Dev.	Std. Error
Rating 1, Delta	20	38.950	14.006	3.132
Rating 1, Lufthansa	20	58.900	15.089	3.374
Rating 1, KLM	20	72.900	13.902	3.108
Rating 2, Delta	20	32.400	8.268	1.849
Rating 2, Lufthansa	20	72.250	10.572	2.364
Rating 2, KLM	20	79.800	11.265	2.519

Means Table Effect: Ratings

	Count	Mean	Std. Dev.	Std. Error
Rating 1	60	56.917	19.902	2.569
Rating 2	60	61.483	23.208	2.996

All data are hypothetical.

EXHIBIT 17–18 Repeated-Measures ANOVA Plot



All data are hypothetical.

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, and 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.[14] 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.[15]

SUMMARY

1 There are two approaches to hypothesis testing—classical or sampling theory statistics and the Bayesian approach. With classical statistics, we make inferences about a population based on evidence gathered from a sample. Although we cannot state unequivo-

cally 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. With prudent use of prior probabilities, the Bayesian approach will also provide good results.

2 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.

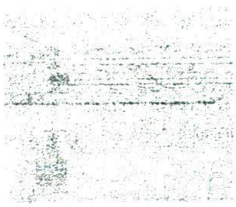
3 Hypothesis testing can be viewed as a six-step procedure:

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 α 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 β or Type II error is the decision error that results from accepting a false null hypothesis. Usually, one determines a level of acceptable α error and then seeks to reduce the β 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.

4 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.

5 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.

KEY TERMS

alternative hypothesis 523
analysis of variance (ANOVA) 546
a priori contrasts 549
Bayesian statistics 521
chi-square (χ^2) test 536
classical statistics 521
critical values 527
F ratio 547
inferential statistics 520
k independent samples test 546
k related samples test 554
level of significance 529
multiple comparison (post hoc) procedures 551
nonparametric tests 531
normal probability plot 532
null hypothesis 523
one-sample tests 534
one-tailed test 524
parametric tests 531
power of the test 528
practical significance 523
region of acceptance 527
region of rejection 527
statistical significance 521
t distribution 535
treatment (factor) 547
trials (repeated measures) 554
t-test 535
two independent samples tests 538
two related samples tests 542
two-tailed test 524
Type I error (α) 525
Type II error (β) 525
Z distribution 535
Z test 538

EXAMPLES

Company	Scenario	Page
American legal system	Correct and incorrect decisions about hypothesis testing are compared to the U.S. legal system's presumption of innocence until guilt is proven.	524
Colgate-Palmolive Co.	Study of toothbrushing ergonomics revealed that some users change their grip more than 100 times during brushing and that thicker-handled brushes promote longer brushing.	539
Containers Inc.	Implementing a smoke-free workplace policy by evaluating the relationship between accidents and smoking.	540
Dean Merrill Brokerage*	Testing the effectiveness of two sales training methods for account executive trainees.	539
e-WEAR*	Controller of e-commerce division of large retail chain evaluates receivables.	521
Forbes sales data	Testing for significances between two years of sales in 10 companies.	544
The Gillette Co.	Its fatter Oral-B toothbrush for "power grippers" earned complaints from users, as it wouldn't fit in toothbrush holders; further research revealed this was a problem in less than 19 percent of households with built-in holders.	539
K2Sports	A study reveals the effectiveness of a retailers sales associate training program for an in-line skates manufacturer.	538

Mediascope	Conducted a content analysis of the top 200 movies to determine the level of substance abuse in movies targeting teens.	531
Metro University*	A study conducted to determine the feasibility of a membership dining club near campus.	537
Office of National Drug Control Policy	Commissioned Mediascope's study of substance abuse in movies targeting teens.	531
South American coffee growers	Illustrations of ANOVA factors and *a priori* contrasts tests.	546
SteelShelf Corporation*	Office furniture manufacturer evaluates an internal teamwork values campaign with before-after testing.	545
Suncorp Metway	Insurance company studies risk factors and includes astrology in a lighthearted way.	543
Travel Industry Magazine*	Comparison of in-flight service differences among three carriers; seat selection and repeated measures of service are also examined.	547
Video Software Dealers Association	Provided the database to determine the top 200 movies used for the Office of National Drug Control Policy's study of substance abuse in movies.	531

*Due to the confidential and proprietary nature of most research, the names of some companies have been changed.

DISCUSSION QUESTIONS

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 certain ecological proposal. Is this difference significant?
6. What hypothesis-testing procedure would you use in the following situations?
   a. A test classifies applicants as accepted or rejected. On the basis of data on 200 applicants, we test the hypothesis that success is not related to gender.

b. A production batch of 26 gaskets must be evaluated on thickness specifications: A 3-mm thickness is specified by the quality control department.

c. A company manufactures automobiles at two different facilities. We want to know if the gas mileage is the same for vehicles from both facilities. There are samples of 45 units from each facility.

d. A company has three categories of managers: (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 manager's motivation level (classified as high, normal, and low). A hypothesis of no relation between manager 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.

f. A company has to evaluate whether it should attribute increased sales to product quality, advertising, or an interaction of product quality and advertising.

7. You conduct a survey of a sample of 25 members of this year's graduating class 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 class significantly different from the long-run average? At what alpha level would it be significant?

8. 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 six students say they are conservative. Is this a statistically significant difference?

9. You contact a random sample of 36 graduates of Western University and learn that their starting salaries were $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 attitudes toward corporations. With the following results, test the hypothesis that there is no difference among students on this attitude.

	Favorable	Neutral	Unfavorable
Freshmen	100	50	70
Sophomores	80	60	70
Juniors	50	50	80
Seniors	40	60	90

11. You do a survey of business school 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:

$X_b = 4.5$ times per week

$s_b = 1.5$

$X_{la} = 5.6$ times per week

$s_{la} = 2.0$

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):

Trial	One-Koat	Competitor A	Competitor B
1	37	34	24
2	30	19	25
3	34	22	23
4	28	31	20
5	29	27	20

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

13. A computer manufacturer is introducing a new product specifically targeted at the home market and wishes to compare the effectiveness of three sales strategies: computer stores, home electronics stores, and department stores. Numbers of sales by 15 salespeople are recorded below:

Electronics store: 5, 4, 3, 3, 3

Department store: 9, 7, 8, 6, 5

Computer store: 7, 4, 8, 4, 3

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

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

14. A financial analyst is interested in whether there was a significant change in profits for utilities from one period to another. A random sample of 11 companies from the Forbes 500 contributed the following data:

Company	Profits Year 1	Profits Year 2
Ohio Edison	218.9	361.0
Kentucky Utilities	79.4	82.3
PSI Holdings	99.1	125.2
Idaho Power	49.0	84.7
NY State E & G	171.5	157.8
Northeast Utilities	224.8	203.2
Southwestern Public Service	105.0	124.9
Pacific Corp.	446.8	465.6
Scana	120.7	122.6
Puget Sound Power & Light	128.2	117.7
Public Service Colorado	124.9	148.8

a. Should a test of independent or related samples be used?

b. Is there a difference in profits between the two years?

15. The *Fortune* magazine annual list of the 40 richest self-made Americans under the age of 40 (*Fortune*, September 17, 2001), revealed some interesting changes. With the collapse of the dot-coms, the new super-rich from sports and entertainment joined the list.

a. Devise a grouping variable to classify the companies in the accompanying table (e.g., Internet, computers, celebrities).

b. Using one-way analysis of variance, test the hypothesis that there is no difference in net worth among the groups.

Rank	Name	Company	Net Worth ($ in millions)
1	Michael Dell	Dell Computer	16,300
2	Pierre Omidyar	eBay	4,390
3	Jeff Skoll	eBay	2,630
4	Ted Waitt	Gateway, Inc.	1,870
5	Jeff Bezos	Amazon.Com, Inc.	1,230
6	Vinny Smith	Quest Software	780
7	David Filo	Yahoo	730
8	Jerry Yang	Yahoo	721
9	Rob Glaser	RealNetworks	635
10	Dan Snyder	Washington Redskins	604
11	Greg Reyes	Brocade Communications Systems	518
12	Jen-Hsun Huang	Nvidia	507
13	Michael Jordan	Washington Wizards	398
14	Joe Liemandt	Trilogy Software	390
15	Jeanette Symons	Zhone Technologies	374
16	Pantas Sutardja	Marvell Technology	363
17	John Schnatter	Papa John's International	293
18	Sanjay Kumar	Computer Associates International, Inc.	270
19	Tom Cruise	Cruise/Wagner Productions	251
20	Percy Miller (Master P)	No Limit	249
21	James T. Demetriades	SeeBeyond Technology	239
22	Sean Combs (P. Diddy)	Bad Boy Entertainment	231
23	Jerry Greenberg	Sapient	225
24	J. Stuart Moore	Sapient	224
25	Sudhakar Ravi	SonicWALL	219
26	Sreekanth Ravi	SonicWALL	219

Rank	Name	Company	Net Worth ($ in millions)
27	David Hitz	Network Appliance	202
28	John L. MacFarlane	Openwave Systems Inc.	198
29	Jeffrey Citron	Vonage	194
30	Raul Fernandez	Dimension Data North America	188
31	Eric Greenberg	Innovation Investments	187
32	Chris Klaus	Internet Security Systems	187
33	Anousheh Ansari	Sonus	180
34	Halsey Minor	12 Entrepreneuring	180
35	Michael Saylor	Microstrategy	180
36	Jim Carrey	Pit Bull Productions	171
37	Jonathan M. Rothberg	CuraGen	168
38	Marc Andreesser	Loudcloud	166
39	Nav Sooch	Silicon Laboratories	162
40	Tiger Woods	ETW	160

16. A consumer testing firm is interested in testing two competing antivirus products for personal computers. It wants to know how many strains of virus will be removed. The data are:

	Removed by Q-Cure?	
Removed by Anti-V?	Yes	No
Yes	45	33
No	58	20

Are Anti-V and Q-Cure equally effective ($\alpha = .05$)?

17. A researcher for an auto manufacturer is examining preferences for styling features on larger sedans. Buyers were classified as "first-time" and "repeat," resulting in the following table.

	Preference	
	European Styling	Japanese Styling
Repeat	40	20
First-time	8	32

a. Test the hypothesis that buying characteristic is independent of styling preference ($\alpha = .05$).

b. Should the statistic be adjusted?

From Concept to Practice

18. Using the data in Exhibit 17–11 for the variables Flight Service Rating 2 and Airline (2, 3), test the hypothesis of no difference between means.

WWW Exercises

Visit our website for Internet exercises related to this chapter at www.mhhe.com/business/cooper8

CASES

HITECH ENGINEERING

INQUIRING MINDS WANT TO KNOW—NOW!

MASTERING TEACHER LEADERSHIP

PERFORMANCE EVALUATIONS

RUBBERGATE

VIOLENCE ON TV

WASTE PAPER

*All cases indicating a video icon are located on the Instructor's Videotape Supplement. All nonvideo cases are in the case section of the textbook. All cases indicating a CD icon offer a data set, which is located on the accompanying CD.

REFERENCE NOTES

1. A more detailed example is found in Amir D. Aczel and Jayavel Sounderpandian, *Complete Business Statistics*, 5th ed. (Chicago: Irwin/McGraw-Hill, 2001).
2. The standardized random variable, denoted by Z, is a deviation from expectancy and is expressed in terms of standard deviation units. The mean of the distribution of a standardized random variable is 0, and the standard deviation is 1. With this distribution, the deviation from the mean by any value of X can be expressed in standard deviation units.
3. Procedures for hypothesis testing are reasonably similar across authors. This outline was influenced by Sidney Siegel, *Nonparametric Statistics for the Behavioral Sciences* (New York: McGraw-Hill, 1956), Chapter 2.
4. Marija J. Norusis/SPSS, Inc., *SPSS for Windows Base System User's Guide, Release 6.0* (Chicago: SPSS, Inc., 1993), pp. 601–06.
5. For further information on these tests, see ibid., pp. 187–88.
6. F. M. Andrews, L. Klem, T. N. Davidson, P. M. O'Malley, and W. L. Rodgers, *A Guide for Selecting Statistical Techniques for Analyzing Social Science Data* (Ann Arbor: Institute for Social Research, University of Michigan, 1976).
7. Statistical Navigator is a product from The Idea Works, Inc.
8. Exhibit 17–7 is partially adapted from Siegel, *Nonparametric Statistics*, flyleaf.
9. See B. S. Everitt, *The Analysis of Contingency Tables* (London: Chapman and Hall, 1977).
10. The critiques are represented by W. J. Conover, "Some Reasons for Not Using the Yates' Continuity Correction on 2×2 Contingency Tables," *Journal of the American Statistical Association* 69 (1974), pp. 374–76; and N. Mantel, "Comment and a Suggestion on the Yates' Continuity Correction," *Journal of the American Statistical Association* 69 (1974), pp. 378–80.
11. This data table and the analysis of variance tables and plots in this section were prepared with SuperANOVA™.
12. See, for example, Roger E. Kirk, *Experimental Design: Procedures for the Behavioral Sciences* (Belmont, CA: Brooks/Cole, 1982), pp. 115–33. An exceptionally clear presentation for step-by-step hand computation is found in James L. Bruning and B. L. Kintz, *Computational Handbook of Statistics*, 2nd ed. (Glenview, IL: Scott, Foresman, 1977), pp. 143–68. Also, when you use a computer program, the reference manual typically provides helpful advice in addition to the set-up instructions.
13. Kirk, *Experimental Design*, pp. 90–115. Alternatively, see Bruning and Kintz, *Computational Handbook of Statistics*, pp. 113–32.
14. For a discussion and example of the Cochran Q test, see Sidney Siegel and N. J. Castellan, Jr., *Nonparametric Statistics for the Behavioral Sciences*, 2nd ed. (New York: McGraw-Hill, 1988).
15. For further details, see Siegel and Castellan, Jr., *Nonparametric Statistics*.

REFERENCES FOR SNAPSHOTS AND CAPTIONS

K2Sports

Michele Marchetti, "Extreme Selling," *Sales and Marketing Management*, October 1, 1999, p. 108.
Amy Buckalter, K2Sports.
http://www.k2sports.com

Mediascope

"New Study Looks at Drugs in Movies and Songs," America Cares, Inc., April 1999 (http://www.americacares.org/drugs_in_movies.htm).
"Substance Use in Popular Movies and Music," Office of National Drug Control Policy, April 1999 (http://www.mediacampaign.org/publications/movies/movie_partIV.html).

Star Signs

"Forget Defensive Driving, Its in the Stars," news release, Suncorp Metway, February 10 2002. For more information contact Warren Duke, national manager, personal insurance, or Catherine Hughes, public affairs (http://corporate.suncorpnetway.com.au/news/news_articles.asp).

Toothbrushes

Mark Maremont, "There's One Cavity New Toothbrushes Just Can't Handle," *The Wall Street Journal*, March 24, 2000, p. 1.

CLASSIC AND CONTEMPORARY READINGS

Aczel, Amir D., and J. Sounderpandian. *Complete Business Statistics*. 5th ed. Chicago: Irwin/McGraw-Hill, 2001. This excellent text is characterized by highly lucid explanations and numerous examples.

Cohen, Jacob. *Statistical Power Analysis for the Behavioral Sciences*. Mahwah, NJ: Lawrence Erlbaum Associates, 1990. A key reference on conducting power analysis.

DeFinetti, Bruno. *Probability, Induction, and Statistics*. New York: Wiley, 1972. A highly readable work on subjective probability and the Bayesian approach.

Kanji, Gopal K. *100 Statistical Tests*. Thousand Oaks, CA: Sage Publications, 1999. Coverage of the most commonly used statistics that students will encounter.

Kirk, Roger E. *Experimental Design: Procedures for the Behavioral Sciences*. 3rd ed. Belmont, CA: Brooks/Cole, 1995. An advanced text on the statistical aspects of experimental design.

Levine, David M., Timothy C. Krehbiel, and Mark L. Berenson. *Business Statistics: A First Course*. Upper Saddle River, NJ: Prentice-Hall, 1999. For students or managers without recent statistical coursework, this text provides an excellent review.

Siegel, Sidney, and N. J. Castellan, Jr. *Nonparametric Statistics for the Behavioral Sciences*. 2nd ed. New York: McGraw-Hill, 1988. The classic book on nonparametric statistics.

Winer, B. J. *Statistical Principles in Experimental Design*. 2nd ed. New York: McGraw-Hill, 1971. Another classic source. Thorough coverage of analysis of variance and experimental design.