CHAPTER 15

STATISTICAL TESTING OF DIFFERENCES AND RELATIONSHIPS

LEARNING OBJECTIVES

1. To become aware of the nature of statistical significance.
2. To understand the concept of hypothesis development and how to test hypotheses.
3. To understand the difference between type I and type II errors.
4. To be familiar with several of the more common statistical tests of goodness of fit, hypotheses about one mean, hypotheses about two means, and hypotheses about proportions.
5. To learn about analysis of variance.

Pat Casey, of Weed Waster, is reviewing the results of a product concept test just completed with Marketing Data Visions (MDV). The test involved evaluation of three alternative enhancements to the company's best selling edger/trimmer, the V900. This product currently has 26.4 percent of the market.

The test was done by mall-intercept interviewing in three cities—Los Angeles, Denver, and Chicago. A total of 900 qualified consumers were surveyed—300 in each market. To qualify, consumers had to be homeowners, take care of their yards themselves, and own a power lawn edger/trimmer.

The survey covered current brand of lawn edger/trimmer used, ratings of competitive brands, demographic and psychographic characteristics, and reactions to the new product concepts. The new concepts offered three different approaches to improving the plastic line that does the trimming. The same people were asked about all three new versions of the system. The order of presentation of the new concepts was randomly rotated from survey to survey to avoid any order bias.

As noted above, the survey covered many issues, but the key questions for Casey were the purchase intent questions. Specifically, the approach used for all the purchase intent questions on the survey was, "If an edger with the new system was available at stores where you normally shop and sold for $49.95, how likely would you be to buy it?" The response options were "very likely," "somewhat likely," "undecided," "somewhat unlikely," and "very unlikely." By Casey's calculations, the new product would need a Top 2 box score (sum of the "very likely" and "somewhat likely" responses) of 36 percent to reach the needed sales volume. The Top 2 box scores for the three concepts were as follows: Concept A 38.3 percent, Concept B 35.3 percent, and Concept C 42.4 percent. Casey has framed the issues that must be addressed as follows:

- Concept C scored the best, but is it possible that the true result for Concept C could be less than the 38 percent target level when sampling error is taken into account?
- Concept A scored over the 36 percent target level, but is it possible that the true result could be less than 36 percent if sampling error is taken into account?

This chapter shows you how to address these issues, and when you have completed it you should be able to answer Pat's questions.

This chapter addresses statistical techniques that can be used to determine whether the observed differences, noted above, are likely to be real differences or whether they are likely attributable to sampling error.

Evaluating Differences and Changes

The issue of whether certain measurements are different from one another is central to many questions of critical interest to marketing managers. Some specific examples of managers' questions follow:

- Our posttest measure of top-of-mind awareness is slightly higher than the level recorded in the pretest. Did top-of-mind awareness really increase, or is there some other explanation for the increase? Should we fire or commend our agency?
- Our overall customer satisfaction score increased from 92 percent 3 months ago to 93.5 percent today. Did customer satisfaction really increase? Should we celebrate?
- Satisfaction with the customer service provided by our cable TV system in Dallas is, on average, 1.2 points higher on a 10-point scale than is satisfaction with the customer service provided by our cable TV system in Cincinnati. Are customers in Dallas really more satisfied? Should the customer service manager in Cincinnati be replaced? Should the Dallas manager be rewarded?
- In a recent product concept test, 19.8 percent of those surveyed said they were very likely to buy the new product they evaluated. Is this good? Is it better than the results we got last year for a similar product? What do these results suggest in terms of whether to introduce the new product?
- A segmentation study shows that those with incomes of more than $30,000 per year frequent fast-food restaurants 6.2 times per month on average. Those with incomes of $30,000 or less go an average of 6.7 times. Is this difference real—is it meaningful?
- In an awareness test, 28.3 percent of those surveyed have heard of our product on an unaided basis. Is this a good result?

These are the eternal questions in marketing and marketing research. Although considered boring by some, statistical hypothesis testing is important because it helps researchers get closer to the ultimate answers to these questions. We say "closer" because certainty is never achieved in answering these questions in marketing research.

Statistical Significance

The basic motive for making statistical inferences is to generalize from sample results to population characteristics. A fundamental tenet of statistical inference is that it is possible for numbers to be different in a mathematical sense but not significantly different in a statistical sense. For example, suppose cola drinkers are asked to try two cola drinks in a blind taste test and indicate which they prefer; the results show that 51 percent prefer one test product and 49 percent prefer the other. There is a mathematical difference in the results, but the difference would appear to be minor and unimportant. The difference probably is well within the range of accuracy of researchers' ability to measure taste preference and thus probably is not significant in a statistical sense. Three different concepts can be applied to the notion of differences when we are talking about results from samples:

- *Mathematical differences.* By definition, if numbers are not exactly the same, they are different. This does not, however, mean that the difference is either important or statistically significant.

- *Statistical significance.* If a particular difference is large enough to be unlikely to have occurred because of chance or sampling error, then the difference is statistically significant.
- *Managerially important differences.* One can argue that a difference is important from a managerial perspective only if results or numbers are sufficiently different. For example, the difference in consumer responses to two different packages in a test market might be statistically significant but yet so small as to have little practical or managerial significance.[1] This issue is discussed in greater detail in the Practicing Marketing Research feature on page 480.

This chapter covers different approaches to testing whether results are statistically significant.

Hypothesis Testing

A **hypothesis** is an assumption or guess that a researcher or manager makes about some characteristic of the population being investigated. The marketing researcher is often faced with the question of whether research results are different enough from the norm that some element of the firm's marketing strategy should be changed. Consider the following situations.

hypothesis
Assumption or theory that a researcher or manager makes about some characteristic of the population under study.

- The results of a tracking survey show that awareness of a product is lower than it was in a similar survey conducted 6 months ago. Are the results significantly lower? Are the results sufficiently lower to call for a change in advertising strategy?
- A product manager believes that the average purchaser of his product is 35 years of age. A survey is conducted to test this hypothesis, and the survey shows that the average purchaser of the product is 38.5 years of age. Is the survey result different enough from the product manager's belief to cause him to conclude that his belief is incorrect?
- The marketing director of a fast-food chain believes that 60 percent of her customers are female and 40 percent are male. She does a survey to test this hypothesis and finds that, according to the survey, 55 percent are female and 45 percent are male. Is this result sufficiently different from her original theory to permit her to conclude that her original theory was incorrect?

All of these questions can be evaluated with some kind of statistical test. In hypothesis testing, the researcher determines whether a hypothesis concerning some characteristic of the population is likely to be true, given the evidence. A statistical hypothesis test allows us to calculate the probability of observing a particular result if the stated hypothesis is actually true.[2]

There are two basic explanations for an observed difference between a hypothesized value and a particular research result. Either the hypothesis is true and the observed difference is likely due to sampling error, or the hypothesis is false and the true value is some other value.

Steps in Hypothesis Testing

Five steps are involved in testing a hypothesis. First, the hypothesis is specified. Second, an appropriate statistical technique is selected to test the hypothesis. Third, a decision rule is specified as the basis for determining whether to reject or fail to reject (FTR) the null hypothesis H_0. Please note that we did not say "reject H_0 or accept H_0." Although a seemingly small distinction, it is an important one. The distinction will be discussed in greater detail later on. Fourth, the value of the test statistic is calculated and the test is

PRACTICING MARKETING RESEARCH

Why We Need Statistical Tests of Differences

A 13-year-old spent $4\frac{1}{2}$ hours searching through the Internet for a workable definition of statistically significant. He needed the definition for his science class. Finally, he came upon Ask A Scientist©, a feature sponsored by the Argonne National Laboratory, Division of Educational Programs for grades K–12. Dr. Ali Khounsary of the laboratory volunteered to answer the student's question.

Say you have a new drug to treat cancer and you try it on sick rats. You apply the drug to 100 sick rats and notice that 20 get well. Does this mean the drug works? Although it seemingly cured 20 rats, perhaps they could have recovered on their own, without the drug. Now you test the drug on 200 sick rats, randomly selected. Divide these 200 rats into two groups of 100 each. Treat one group with the actual drug and the other with a fake or placebo drug. The person administering the real and fake drugs will not know which pills are which.

After a while, see how many rats have recovered. If 20 taking the real drug and only 5 from the placebo group are cured, then you would comfortably feel that the drug really works. But if 18 rats taking the placebo actually survived, then you would not be sure that the new drug had any effect because the difference between the two survival rates, 20 on the real drug, 18 on the placebo, is not very big.

Statistical tests provide a basis for determining whether differences are greater or not greater than we would expect due to chance.[3]

performed. Fifth, the conclusion is stated from the perspective of the original research problem or question.

Step One: Stating the Hypothesis Hypotheses are stated using two basic forms: the null hypothesis H_0 and the alternative hypothesis H_a. The null hypothesis H_0 (sometimes called the *hypothesis of the status quo*) is the hypothesis that is tested against its complement, the alternative hypothesis H_a (sometimes called the *research hypothesis of interest*). For more discussion of the null hypothesis see the Practicing Marketing Research feature on page 481. Suppose the manager of Burger City believes that his operational procedures will guarantee that the average customer will wait 2 minutes in the drive-in window line. He conducts research, based on the observation of 1,000 customers at randomly selected stores at randomly selected times. The average customer observed in this study spends 2.4 minutes in the drive-in window line. The null hypothesis and the alternative hypothesis might be stated as follows:

- Null hypothesis H_0: Mean waiting time $= 2$ minutes
- Alternative hypothesis H_a: Mean waiting time $\neq 2$ minutes

It should be noted that the null hypothesis and the alternative hypothesis must be stated in such a way that both cannot be true. The idea is to use the available evidence to ascertain which hypothesis is more likely to be true.

Step Two: Choosing the Appropriate Test Statistic As you will see in the following sections of this chapter, the analyst must choose the appropriate statistical test, given the characteristics of the situation under investigation. A number of different statistical tests, along with the situations where they are appropriate, are discussed in this

PRACTICING MARKETING RESEARCH

The Null Hypothesis—Is It a Joking Matter?

Even though the null hypothesis fulfills a legitimate function within the model of scientific discovery, and especially in market research, ever since Karl Popper developed it as a statistical tool of empirical research it has had mixed reviews. Clearly, a bias against the null hypothesis has been observed in the social sciences because measuring no-effects has little practical use, it's believed.

Perhaps that is what led to the following joke about the null hypothesis. A person tumbles into a deep hole. He tries to escape many times but fails. He is exhausted, on the verge of collapse, and finally he mutters, "It must be impossible to get out of here." Then he hears a voice not far away, also muttering in the darkness of the hole. "You are so right. I tried every method you did to get out of this hole. None of them works." The first person, hearing this, is startled. Then he answers with bitterness, "Then why didn't you tell me before now?" The other voice answers, "So who publishes null results?"[4]

chapter. Exhibit 15.1 provides a guide to selecting the appropriate test for various situations. All the tests in this table are covered in detail later in this chapter.

Step Three: Developing a Decision Rule Based on our previous discussions of distributions of sample means, you may recognize that one is very unlikely to get a sample result that is exactly equal to the value of the population parameter. The problem is determining whether the difference, or deviation, between the actual value of the sample mean and its expected value based on the hypothesis could have occurred by chance (5 times out of 100, for example) if the statistical hypothesis is true. A decision rule, or standard, is needed to determine whether to reject or fail to reject the null hypothesis. Statisticians state such decision rules in terms of significance levels.

The significance level (α) is critical in the process of choosing between the null and alternative hypotheses. The level of significance—.10, .05, or .01, for example—is the probability that is considered too low to justify acceptance of the null hypothesis.

Consider a situation in which the researcher has decided that she wants to test a hypothesis at the .05 level of significance. This means that she will reject the null hypothesis if the test indicates that the probability of occurrence of the observed result (for example, the difference between the sample mean and its expected value) because of chance or sampling error is less than 5 percent. Rejection of the null hypothesis is equivalent to supporting the alternative hypothesis.

Step Four: Calculating the Value of the Test Statistic In this step, the researcher does the following:

- Uses the appropriate formula to calculate the value of the statistic for the test chosen.
- Compares the value just calculated to the critical value of the statistic (from the appropriate table), based on the decision rule chosen.
- Based on the comparison, determines to either reject or fail to reject the null hypothesis H_0.

Step Five: Stating the Conclusion The conclusion summarizes the results of the test. It should be stated from the perspective of the original research question.

EXHIBIT 15.1 Statistical Tests and Their Uses

Area of Application	Subgroups or Samples	Level Scaling	Test	Special Requirements	Example
Hypotheses about frequency distribution	One	Nominal	χ^2	Random sample	Are observed differences in the numbers of people responding to three different promotions likely/not likely due to chance?
	Two or more	Nominal	χ^2	Random sample, independent samples	Are differences in the numbers of men and women responding to a promotion likely/not likely due to chance?
Hypotheses about means	One (large sample)	Metric (interval or ratio)	Z test for one mean	Random sample, $n \geq 30$	Is the observed difference between a sample estimate of the mean and some set standard or expected value of the mean likely/not likely due to chance?
	One (small sample)	Metric (interval or ratio)	t test for one mean	Random sample, $n < 30$	Same as for small sample above
	Two (large sample)	Metric (interval or ratio)	Z test for one mean	Random sample, $n \geq 30$	Is the observed difference between the means for two subgroups (mean income for men and women) likely/not likely due to chance?
	Three or more	Metric (interval or ratio)	One-way ANOVA	Random sample	Is the observed variation between means for three or more subgroups (mean expenditures on entertainment for high-, moderate-, and low-income people) likely/not likely due to chance?
Hypotheses about proportions	One (large sample)	Metric (interval or ratio)	Z test for one proportion	Random sample, $n \geq 30$	Is the observed difference between a sample estimate of proportion (percentage who say they will buy) and some set standard or expected value likely/not likely due to chance?
	Two (large sample)	Metric (interval or ratio)	Z test for two proportions	Random sample, $n \geq 30$	Is the observed difference between estimated percentages for two subgroups (percentage of men and women who have college degrees) likely/not likely due to chance?

Types of Errors in Hypothesis Testing

type I error (α error) Rejection of the null hypothesis when, in fact, it is true.

Hypothesis tests are subject to two general types of errors, typically referred to as type I error and type II error. A **type I error** involves rejecting the null hypothesis when it is, in fact, true. The researcher may reach this incorrect conclusion because the observed difference between the sample and population values is due to sampling error. The researcher must

FROM THE FRONT LINE

Tips on Significance Testing

Paul Schmiege, Marketing Science, DSS Research

A typical question in the marketing research industry goes something like this: "Last year, the percentage of respondents aware of our brand on an unaided basis was 43.2 percent. This year the corresponding percentage was 47.5 percent. Is this difference significant?" This is not the same as asking, "Is this difference important, is it something that we should act on, should we continue with the same advertising strategy to increase unaided awareness?"

However, you must be careful to distinguish the technical term *statistical significance* from more intuitive terms such as *practical significance* or *importance*. At the heart of its technical meaning, significance in the field of statistics means that the difference is likely greater than we would expect due to sampling error.

In the above example, you only have statistics about samples and not population parameters. Therefore, you can never be entirely certain that a difference between two sample results is a real difference. The difference may only reflect sampling error.

The single most common test in marketing research is the two-sample *t* test. You might not ever see any other test in your entire career. Because the two-sample *t* test is so important, it is good to keep at least two points in mind about it:

- The two-sample *t* test is a two-tailed test. It asks, "Does a significant difference exist?" It does not ask, "Is the first significantly greater than the second?" or "Is the first significantly less than the second?" Consequently, if a significant difference exists you should say, "A statistically significant difference exists, and that observed difference is higher (or lower)."
- The two-sample *t* test is run with the assumption of equal variances. The true standard deviation for the combined populations is unknown, so you "pool" the two-sample standard deviations together to calculate something similar to a weighted average. In academic research, you would first test whether to assume equal or unequal variance, but you will probably never have to do that in the business world.

For any given observed difference, there are sample sizes large enough that the difference will be significant in a two-sample *t* test (sample sizes go in the denominator of the equation). Or to think of it from another perspective, when you start testing with larger and larger sample sizes, smaller and smaller differences become statistically significant, but practical significance remains the same. Is a 0.5 percent increase worth telling management about even if it should happen to be statistically significant? Probably not.

In the end, you have to rely on your expertise and judgment regarding what's important and what's not important for a particular measure in a particular industry. Significance, or nonsignificance, is just another piece of information you use to decide what to recommend to clients.

decide how willing she or he is to commit a type I error. The probability of committing a type I error is referred to as the *alpha* (α) *level*. Conversely, $1 - \alpha$ is the probability of making a correct decision by not rejecting the null hypothesis when, in fact, it is true.

A **type II error** involves failing to reject the null hypothesis when it actually is false. A type II error is referred to as a *beta* (β) *error*. The value $1 - \beta$ reflects the probability of making a correct decision in rejecting the null hypothesis when, in fact, it is false. The four possibilities are summarized in Exhibit 15.2.

type II error (β error)
Failure to reject the null hypothesis when, in fact, it is false.

EXHIBIT 15.2 Type I and Type II Errors

Actual State of the Null Hypothesis	Fail to Reject H_0	Reject H_0
H_0 is true	Correct $(1 - \alpha)$	Type I error (α)
H_0 is false	Type II error (β)	Correct $(1 - \beta)$

As we consider the various types of hypothesis tests, keep in mind that when a researcher rejects or fails to reject the null hypothesis, this decision is never made with 100 percent certainty. There is a probability that the decision is correct and there is a probability that the decision is not correct. The level of α is set by the researcher, after consulting with his or her client, considering the resources available for the project, and considering the implications of making type I and type II errors. However, the estimation of β is more complicated and is beyond the scope of our discussion. Note that type I and type II errors are not complementary; that is, $\alpha + \beta \neq 1$.

It would be ideal to have control over n (the sample size), α (the probability of a type I error), and β (the probability of a type II error) for any hypothesis test. Unfortunately, only two of the three can be controlled. For a given problem with a fixed sample size, n is fixed, or controlled. Therefore, only one of α and β can be controlled.

Assume that for a given problem you have decided to set $\alpha = .05$. As a result, the procedure you use to test H_0 versus H_a will reject H_0 when it is true (type I error) 5 percent of the time. You could set $\alpha = 0$ so that you would never have a type I error. The idea of never rejecting a correct H_0 sounds good. However, the downside is that β (the probability of a type II error) is equal to 1 in this situation. As a result, you will always fail to reject H_0 when it is false. For example, if $\alpha = 0$ in the fast-food service time example, where H_0 is mean waiting time = 2 minutes, then the resulting test of H_0 versus H_a will automatically fail to reject H_0 (mean waiting time = 2 minutes) whenever the estimated waiting time is any value other than 2 minutes. If, for example, we did a survey and determined that the mean waiting time for the people surveyed was 8.5 minutes, we would still fail to reject (FTR) H_0. As you can see, this is not a good compromise. We need a value of α that offers a more reasonable compromise between the probabilities of the two types of errors. Note that in the situation in which $\alpha = 0$ and $\beta = 1$, $\alpha + \beta = 1$. As you will see later on, this is not true as a general rule.

The value of α selected should be a function of the relative importance of the two types of errors. Suppose you have just had a diagnostic test. The purpose of the test is to determine whether you have a particular medical condition that is fatal in most cases. If you have the disease, a treatment that is painless, inexpensive, and totally without risk will cure the condition 100 percent of the time. Here are the hypotheses to be tested:

H_0: Test indicates that you do not have the disease.
H_a: Test indicates that you do have the disease.

Thus,

$$\begin{aligned}\alpha &= P(\text{rejecting } H_0 \text{ when it is true})\\ &= (\text{test indicates that you have the disease when}\\ &\quad \text{you do not have it})\end{aligned}$$

$$\beta = P(\text{FTR H}_0 \text{ when in fact it is false})$$
$$= P(\text{test indicates that you do not have the disease when you do have it})$$

Clearly, a type I error (measured by α) is not nearly as serious as a type II error (measured by β). A type I error is not serious because the test will not harm you if you are well. However, a type II error means that you will not receive the treatment you need even though you are ill.

The value of β is never set in advance. When α is made smaller, β becomes larger. If you want to minimize type II error, then you choose a larger value for α in order to make β smaller. In most situations, the range of acceptable values for α is .01 to .1.

In the case of the diagnostic test situation, you might choose a value of α at or near .1 because of the seriousness of a type II error. Conversely, if you are more concerned about type I errors in a given situation, then a small value of α is appropriate. For example, suppose you are testing commercials that were very expensive to produce, and you are concerned about the possibility of rejecting a commercial that is really effective. If there is no real difference between the effects of type I and type II errors, as is often the case, an α value of .05 is commonly used.

Accepting H_0 versus Failing to Reject (FTR) H_0

Researchers often fail to make a distinction between accepting H_0 and failing to reject H_0. However, as noted earlier, there is an important distinction between these two decisions. When a hypothesis is tested, H_0 is presumed to be true until it is demonstrated to be likely to be false. In any hypothesis testing situation, the only other hypothesis that can be accepted is the alternative hypothesis H_a. Either there is sufficient evidence to support H_a (reject H_0) or there is not (fail to reject H_0). The real question is whether there is enough evidence in the data to conclude that H_a is correct. If we fail to reject H_0, we are saying that the data do not provide sufficient support of the claim made in H_a—not that we accept the statement made in H_0.

One-Tailed versus Two-Tailed Test

Tests are either one-tailed or two-tailed. The decision as to which to use depends on the nature of the situation and what the researcher is trying to demonstrate. For example, when the quality control department of a fast-food organization receives a shipment of chicken breasts from one of its vendors and needs to determine whether the product meets specifications in regard to fat content, a one-tailed test is appropriate. The shipment will be rejected if it does not meet minimum specifications. On the other hand, the managers of the meat company that supplies the product should run two-tailed tests to determine two factors. First, they must make sure that the product meets the minimum specifications of their customer before they ship it. Second, they want to determine whether the product exceeds specifications because this can be costly to them. If they are consistently providing a product that exceeds the level of quality they have contracted to provide, their costs may be unnecessarily high.

The classic example of a situation requiring a two-tailed test is the testing of electric fuses. A fuse must trip, or break contact, when it reaches a preset temperature or a fire may result. On the other hand, you do not want the fuse to break contact before it reaches the specified temperature or it will shut off the electricity unnecessarily. The test used in the quality control process for testing fuses must, therefore, be two-tailed.

PRACTICING MARKETING RESEARCH

When Statistical Significance Expands into Practical Significance

Statistical significance refers to results that are true and do not represent random sampling fluctuations or chance in a marketing survey. But managerially important differences, also called practical significance, refer to differences that are judged sufficient or impressive enough to warrant changes in your marketing approach. Here's an example from Honda and its new hybrid Civic that combines a gasoline engine and an electric motor. The question pertains to gas mileage.

Honda focused its advertising approach on gas savings. The hybrid Civic uses its electric motor at slow speeds, then switches to a combination of gas and electric motors for higher speeds, such as when you are on the freeways. Consider that the Civic maintains 50 mpg on average with a deviation of $\pm$10 mpg. Then researchers, mulling over their data on all the vehicles used in testing, discover the actual average mpg is 51. Is this statistically significant? Is the 1 mpg difference real? It must be, for it's based on a count and quantification of mileage of all Civic vehicles. Random sampling was not involved in generating this result; it is a population value.

Although this number, then, is statistically significant, the next level question is: does it have practical significance, big enough to motivate the marketing department to change commercials already produced based on a 50 mpg claim?

The quality control department of a fast-food organization would probably do a one-tailed test to determine whether a shipment of chicken breasts met product specifications. However, a two-tailed test would probably be done by the managers of the meat company that supplied the chicken breasts.

Example of Performing a Statistical Test

Income is an important determinant of the sales of luxury cars. Lexus North America (LNA) is in the process of developing sales estimates for the Southern California market, one of its key markets. According to the U.S. Census, the average annual family income in the market is \$55,347. LNA has just completed a survey of 250 randomly selected households in the market to obtain other measures needed for its sales forecasting model. The recently completed survey indicates that the average annual family income in the market is \$54,323. The actual value of the population mean (μ) is unknown. This gives us two estimates of μ: the census result and the survey result. The difference between these two estimates could make a substantial difference in the estimates of Lexus sales produced by LNA's forecasting model. In the calculations, the U.S. Census Bureau estimate is treated as the best estimate of μ.

LNA decides to statistically compare the census and survey estimates. The statistics for the sample are

$$\bar{X} = \$54{,}323$$
$$S = \$4{,}323$$
$$n = 250$$

The following hypotheses are produced:

$$H_0\colon \mu = \$55{,}347$$
$$H_a\colon \mu \neq \$55{,}347$$

The decision makers at LNA are willing to use a test that will reject H_0 when it is correct only 5 percent of the time ($\alpha = .05$). This is the significance level of the test. LNA will reject H_0 if $|\bar{X} - \$55{,}347|$ is larger than can be explained by sampling error at $\alpha = .05$.

Standardizing the data so that the result can be directly related to Z values in Exhibit 2 in Appendix 2, we have the following criterion:

Reject H_0 if $\left|\dfrac{\bar{X} - \$55{,}347}{S/\sqrt{n}}\right|$ is larger than can be explained by sampling error at $\alpha = .05$. This expression can be rewritten as

$$\left|\frac{\bar{X} - \$55{,}347}{S/\sqrt{n}}\right| > k$$

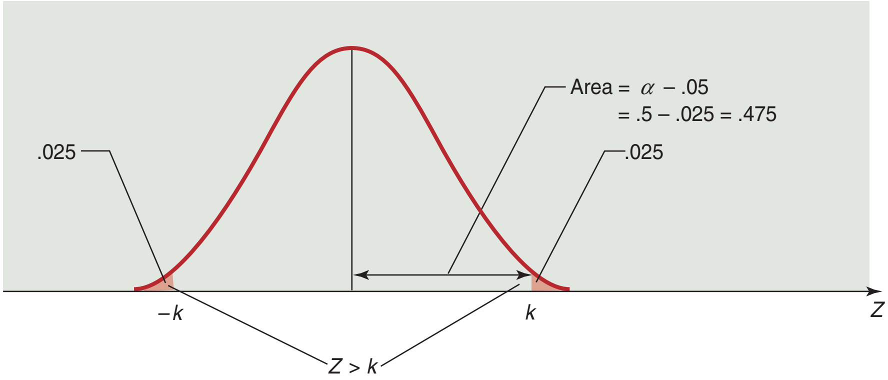
What is the value of k? If H_0 is true and the sample size is large (≥ 30), then (based on the central limit theorem) X approximates a normal random variable with

$$\text{Mean} = \mu = \$55{,}347$$
$$\text{Standard deviation} = \frac{S}{\sqrt{n}}$$

That is, if H_0 is true, $(\bar{X} - \$55{,}347)/(S/\sqrt{n})$ approximates a standard normal variable Z for samples of 30 or larger with a mean equal to 0 and a standard deviation equal to 1.

Exhibit 15.3

Shaded Area Is Significance Level α



We will reject H_0 if $|Z| > k$. When $|Z| > k$, either $Z > k$ or $Z < -k$, as shown in Exhibit 15.3. Given that

$$P(|Z| > k) = .05$$

the total shaded area is .05, with .025 in each tail (two-tailed test). The area between 0 and k is .475. Referring to Exhibit 2 in Appendix 2, we find that $k = 1.96$. Therefore, the test is

$$\text{Reject } H_0 \text{ if } \left|\frac{\overline{X} - \$55{,}347}{S/\sqrt{n}}\right| > 1.96$$

and FTR H_0 otherwise. In other words,

$$\text{Reject } H_0 \text{ if } \left|\frac{\overline{X} - \$55{,}347}{S/\sqrt{n}}\right| > 1.96 \text{ or if } \left|\frac{\overline{X} - \$55{,}347}{S/\sqrt{n}}\right| < 1.96$$

The question is, is $\overline{X} = \$54{,}323$ far enough away from \$55,347 for LNA to reject H_0? The results show that

$$Z = \frac{\overline{X} - \$55{,}347}{S/\sqrt{n}}$$
$$= \frac{\$54{,}323 - \$55{,}347}{\$4{,}322/\sqrt{250}} = -3.75$$

Because $-3.75 < -1.96$, we reject H_0. On the basis of the sample results and $\alpha = .05$, the conclusion is that the average household income in the market is not equal to \$55,347. If H_0 is true ($\mu = \$55{,}347$), then the value of $\overline{X}$ obtained from the sample (\$54,323) is 3.75 standard deviations to the left of the mean on the normal curve for $\overline{X}$. A value of $\overline{X}$ this far away from the mean is very unlikely (probability is less than .05). As a result, we conclude that H_0 is not likely to be true, and we reject it.

Commonly Used Statistical Hypothesis Tests

A number of commonly used statistical hypothesis tests of differences are presented in the following sections. Many other statistical tests have been developed and are used, but a full discussion of all of them is beyond the scope of this text.

The distributions used in the following sections for comparing the computed and tabular values of the statistics are the Z distribution, the t distribution, the F distribution, and the chi-square (χ^2) distribution. The tabular values for these distributions appear in Exhibits 2, 3, 4, and 5 of Appendix 2.

Independent versus Related Samples

In some cases, one needs to test the hypothesis that the value of a variable in one population is equal to the value of that same variable in another population. Selection of the appropriate test statistic requires the researcher to consider whether the samples are independent or related. **Independent samples** are those in which measurement of the variable of interest in one sample has no effect on measurement of the variable in the other sample. It is not necessary that there be two different surveys, only that the measurement of the variable in one population has no effect on the measurement of the variable in the other population. In the case of **related samples**, measurement of the variable of interest in one sample may influence measurement of the variable in another sample.

independent samples Samples in which measurement of a variable in one population has no effect on measurement of the variable in the other.

related samples Samples in which measurement of a variable in one population may influence measurement of the variable in the other.

If, for example, men and women were interviewed in a particular survey regarding their frequency of eating out, there is no way that a man's response could affect or change the way a woman would respond to a question in the survey. Thus, this would be an example of independent samples. On the other hand, consider a situation in which the researcher needed to determine the effect of a new advertising campaign on consumer awareness of a particular brand. To do this, the researcher might survey a random sample of consumers before introducing the new campaign and then survey the same sample of consumers 90 days after the new campaign was introduced. These samples are not independent. The measurement of awareness 90 days after the start of the campaign may be affected by the first measurement.

Degrees of Freedom

Many of the statistical tests discussed in this chapter require the researcher to specify degrees of freedom in order to find the critical value of the test statistic from the table for that statistic. The number of degrees of freedom is the number of observations in a statistical problem that are not restricted or are free to vary.

The number of degrees of freedom (d.f.) is equal to the number of observations minus the number of assumptions or constraints necessary to calculate a statistic. Consider the problem of adding five numbers when the mean of the five numbers is known to be 20. In this situation, only four of the five numbers are free to vary. Once four of the numbers are known, the last value is also known (can be calculated) because the mean value must be 20. If four of the five numbers were 14, 23, 24, and 18, then the fifth number would have to be 21 to produce a mean of 20. We would say that the sample has $n - 1$ or 4 degrees of freedom. It is as if the sample had one less observation—the inclusion of degrees of freedom in the calculation adjusts for this fact.

Goodness of Fit

Chi-Square Test

As noted earlier in the text, data collected in surveys are often analyzed by means of one-way frequency counts and cross tabulations.[5] The purpose of a cross tabulation is to study relationships among variables. The question is, do the numbers of responses that fall into the various categories differ from what one would expect? For example, a study might involve partitioning users into groups by gender (male, female), age (under 18, 18 to 35, over 35), or income level (low, middle, high) and cross tabulating on the basis of answers to questions about preferred brand or level of use. The **chi-square** (χ^2) **test** enables the research analyst to determine whether an observed pattern of frequencies corresponds to, or fits, an "expected" pattern.[6] It tests the "goodness of fit" of the observed distribution to an expected distribution. We will look at the application of this technique to test distributions of cross-tabulated categorical data for a single sample and for two independent samples.

chi-square test
Test of the goodness of fit between the observed distribution and the expected distribution of a variable.

Chi-Square Test of a Single Sample

Suppose the marketing manager of a retail electronics chain needs to test the effectiveness of three special deals (deal 1, deal 2, and deal 3). Each deal will be offered for a month. The manager wants to measure the effect of each deal on the number of customers visiting a test store during the time the deal is on. The number of customers visiting the store under each deal is as follows:

Deal	Month	Customers per Month
1	April	11,700
2	May	12,100
3	June	11,780
Total		35,580

The marketing manager needs to know whether there is a significant difference between the numbers of customers visiting the store during the time periods covered by the three deals. The chi-square (χ^2) one-sample test is the appropriate test to use to answer this question. This test is applied as follows:

1. Specify the null and alternative hypotheses.
   - Null hypothesis H_0: The numbers of customers visiting the store under the various deals are equal.
   - Alternative hypothesis H_a: There is a significant difference in the numbers of customers visiting the store under the various deals.

2. Determine the number of visitors who would be expected in each category if the null hypothesis were correct (E_i). In this example, the null hypothesis is that there is no difference in the numbers of customers attracted by the different deals. Therefore, an equal number of customers would be expected under each deal. Of course, this assumes that no other factors influenced the number of visits to the store. Under

the null (no difference) hypothesis, the expected number of customers visiting the store in each deal period would be computed as follows:

$$E_i = \frac{\text{TV}}{N}$$

where TV = total number of visits
N = number of months

Thus,

$$E_i = \frac{35{,}580}{3} = 11{,}860$$

The researcher should always check for cells in which small expected frequencies occur because they can distort χ^2 results. No more than 20 percent of the categories should have an expected frequency of less than 5, and none should have an expected frequency of less than 1. This is not a problem in this case.

3. Calculate the χ^2 value, using the formula

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

where O_i = observed number in ith category
E_i = expected number in ith category
k = number of categories

For this example,

$$\chi^2 = \frac{(11{,}700 - 11{,}860)^2}{11{,}860} + \frac{(12{,}100 - 11.860)^2}{11{,}860} + \frac{(11{,}780 - 11{,}860)^2}{11{,}860} = 7.6$$

4. Select the level of significance α. If the .05 level of significance (α) is selected, the tabular χ^2 value with 2 degrees of freedom ($k - 1$) is 5.99. (See Exhibit 4 in Appendix 2 for $k - 1 = 2$ d.f., $\alpha = .05$.)
5. State the result. Because the calculated χ^2 value (7.6) is higher than the table value (5.99), we *reject the null hypothesis*. Therefore, we conclude with 95 percent confidence that customer response to the deals was significantly different. Unfortunately, this test tells us only that the overall variation among the cell frequencies is greater than would be expected by chance. It does not tell us whether any individual cell is significantly different from the others.

PRACTICING MARKETING RESEARCH

A Simple Field Application of Chi-Square Goodness of Fit

Say you are a marketing manager and you want to know if an observed pattern of frequencies differs from an expected pattern. The best test for addressing this question is the chi-square test of goodness of fit in which only one categorical variable is involved. Here's how to apply it in a case of packaging color design.

As a marketing manager you have five colors to choose from for your packaging design, but you can only use one. Which one does the market prefer? Obviously, you want to know this before you debut the product. You do a random sampling of 400 consumers and ask them, getting these results:

Package Color	Consumer Preference
Red	70
Blue	106
Green	80
Pink	70
Orange	74
TOTAL	400

The results suggest that people prefer blue, but you need to be sure that this seeming preference for blue is not a chance result. Your null hypothesis posits that all colors are preferred equally, but your alternative hypothesis says they are not equally preferred. In the calculations, note that in terms of the null hypothesis for equal preference for all colors, the expected frequencies for all colors will equal 80. The chi-square value is 11.40 calculated per the standard formula. The results are below.

Because the critical value of chi-square at the 0.5 level of significance (5 percent probability of incorrectly rejecting the null hypothesis) for 4 degrees of freedom is 9.488, you can eliminate the null hypothesis. You can conclude that consumers do not equally prefer all colors but in fact like blue the best. As marketing manager, you can introduce your product in a blue package and do so confidently that it is the best option of those available to you.[7]

Package Color	Observed Frequencies (O)	Expected Frequencies (E)	(O − E)	$\chi^2 = \sum\left(\frac{(O-E)^2}{E}\right)$
Red	70	80	100	1.25
Blue	106	80	676	8.45
Green	80	80	0	0.00
Pink	70	80	100	1.25
Orange	74	80	36	0.45
TOTAL	400	400	—	11.40

Chi-Square Test of Two Independent Samples Marketing researchers often need to determine whether there is any association between two or more variables. Before formulation of a marketing strategy, questions such as the following may need to be answered: Are men and women equally divided into heavy-, medium-, and light-user

EXHIBIT 15.4 **Data for χ^2 Test of Two Independent Samples**

Visits to Convenience Store by Males				Visits to Convenience Stores by Females			
Number X_m	Frequency f_m	Percent	Cumulative Percent	Number X_f	Frequency f_f	Percent	Cumulative Percent
2	2	4.4	4.4	2	5	7.0	7.0
3	5	11.1	15.6	3	4	5.6	12.7
5	7	15.6	31.1	4	7	9.9	22.5
6	2	4.4	35.6	5	10	14.1	36.6
7	1	2.2	37.8	6	6	8.5	45.1
8	2	4.4	42.2	7	3	4.2	49.3
9	1	2.2	44.4	8	6	8.5	57.7
10	7	15.6	60.0	9	2	2.8	60.6
12	3	6.7	66.7	10	13	18.3	78.9
15	5	11.1	77.8	12	4	5.6	84.5
20	6	13.3	91.1	15	3	4.2	88.7
23	1	2.2	93.3	16	2	2.8	91.5
25	1	2.2	95.6	20	4	5.6	97.2
30	1	2.2	97.8	21	1	1.4	98.6
40	1	2.2	100.0	25	1	1.4	100.0

$n_m = 45$

Mean number of visits by males, $\bar{X}_m = \dfrac{\sum X_m f_m}{45} = 11.5$

$n_f = 71$

Mean number of visits by females, $\bar{X}_f = \dfrac{\sum X_f f_f}{71} = 8.5$

categories? Are purchasers and nonpurchasers equally divided into low-, middle-, and high-income groups? The chi-square (χ^2) test for two independent samples is the appropriate test in such situations.

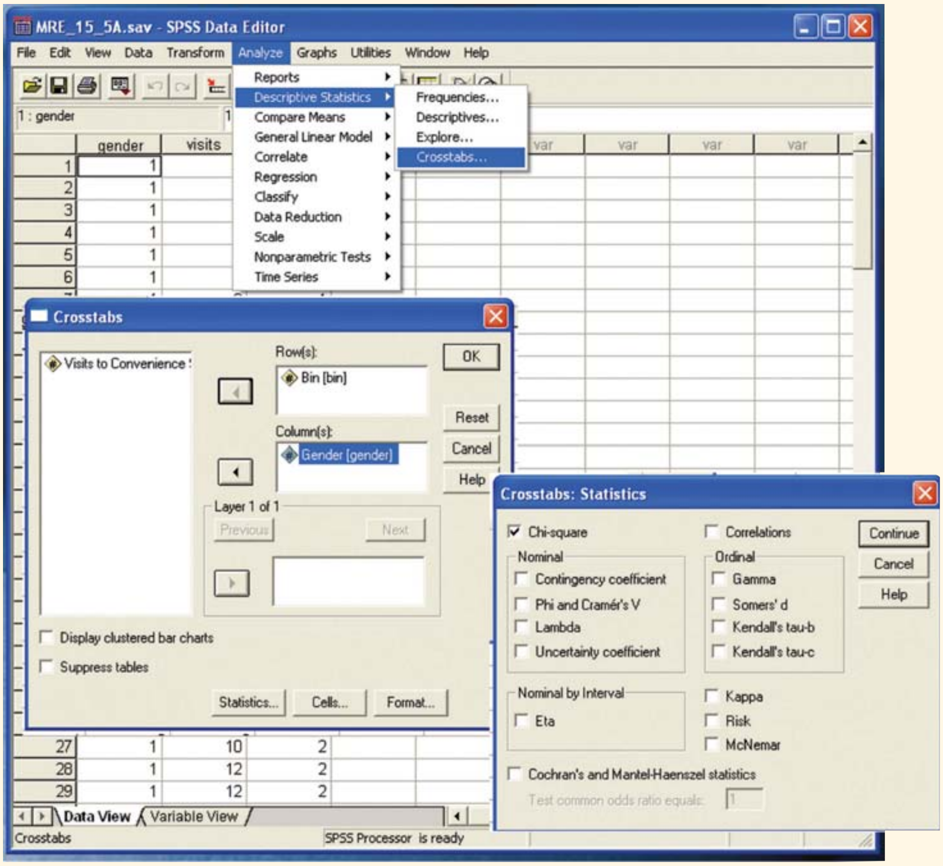
The technique will be illustrated using the data from Exhibit 15.4. A convenience store chain wants to determine the nature of the relationship, if any, between gender of customer and frequency of visits to stores in the chain. Frequency of visits has been divided into three categories: 1 to 5 visits per month (light user), 6 to 14 visits per month (medium user), and 15 and above visits per month (heavy user). The steps in conducting this test follow.

1. State the null and alternative hypotheses.
   - Null hypothesis H_0: There is no relationship between gender and frequency of visits.
   - Alternative hypothesis H_a: There is a significant relationship between gender and frequency of visits.
2. Place the observed (sample) frequencies in a $k \times r$ table (cross-tabulation or contingency table), using the k columns for the sample groups and the r rows for the conditions or treatments. Calculate the sum of each row and each column. Record those totals at the margins of the table (they are called *marginal totals*). Also, calculate the total for the entire table (N).

	Male	Female	Totals
1–5 visits	14	26	40
6–14 visits	16	34	50
15 and above visits	15	11	26
Totals	45	71	116

SPSS JUMP START FOR CHI-SQUARE TEST

Steps that you need to go through to do the chi-square test problem shown in the book are provided below along with the output produced. Use the data set **Chisqex**, which you can download from the Web site for the text.



Steps in SPSS

1. Select *Analyze* → *Descriptive Statistics* → *Crosstabs.*
2. Move **bin** to Rows.
3. Move **gender** to Columns.

4. Click Statistics.
5. Check box for Chi-square
6. Click Continue.
7. Click OK.

SPSS Output for Chi-Square Test

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Bin * Gender	116	100.0%	0	.0%	116	100.0%

Bin * Gender Crosstabulation

Count

		Gender		Total
		Male	Female	
Bin	1-5 visits	14	26	40
	6-14 visits	16	34	50
	15 and above visits	15	11	26
Total		45	71	116

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	5.125[a]	2	.077
Likelihood Ratio	5.024	2	.081
Linear-by-Linear Association	2.685	1	.101
N of Valid Cases	116		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 10.09.

3. Determine the expected frequency for each cell in the contingency table by calculating the product of the two marginal totals common to that cell and dividing that value by *N*.

	Male	Female
1–5 visits	$\frac{45 \times 40}{116} = 15.5$	$\frac{71 \times 40}{116} = 24.5$
6–14 visits	$\frac{45 \times 50}{116} = 19.4$	$\frac{71 \times 50}{116} = 30.6$
15 and above visits	$\frac{45 \times 26}{116} = 10.1$	$\frac{71 \times 26}{116} = 15.9$

The χ^2 value will be distorted if more than 20 percent of the cells have an expected frequency of less than 5 or if any cell has an expected frequency of less than 1. The test should not be used under these conditions.

4. Calculate the value of χ^2 using

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} = observed number in the *i*th row of the *j*th column
E_{ij} = expected number in the *i*th row of the *j*th column

For this example,

$$\chi^2 = \frac{(14 - 15.52)^2}{15.52} + \frac{(26 - 24.48)^2}{24.48} + \frac{(16 - 19.4)^2}{19.4} + \frac{(34 - 30.6)^2}{30.6} + \frac{(15 - 10.09)^2}{10.09} + \frac{(11 - 15.91)^2}{15.91} = 5.1$$

5. State the result. The tabular χ^2 value at a .05 level of significance and $(r - 1) \times (k - 1) = 2$ degrees of freedom is 5.99 (see Table 4 of Appendix 2). Because the calculated $\chi^2 = 5.1$ is less than the tabular value, we *fail to reject (FTR) the null hypothesis* and conclude that there is no significant difference between males and females in terms of frequency of visits.

Another chi-square example is provided in the Practicing Marketing Research feature on page 492.

Hypotheses about One Mean

Z Test

Z test
Hypothesis test used for a single mean if the sample is large enough and drawn at random.

One of the most common goals of marketing research studies is to make some inference about the population mean. If the sample size is large enough ($n \geq 30$), the appropriate test statistic for testing a hypothesis about a single mean is the ***Z* test**. For small samples ($n < 30$), the t test with $n - 1$ degrees of freedom (where n = sample size) should be used.

Video Connection, a Dallas video store chain, recently completed a survey of 200 consumers in its market area. One of the questions was "Compared to other video stores in the area, would you say Video Connection is much better than average, somewhat better than average, average, somewhat worse than average, or much worse than average?" Responses were coded as follows:

Response	Code
Much better	5
Somewhat better	4
Average	3
Somewhat worse	2
Much worse	1

The mean rating of Video Connection is 3.4. The sample standard deviation is 1.9. How can the management of Video Connection be confident that its video stores' mean rating is significantly higher than 3 (average in the rating scale)? The Z test for hypotheses about one mean is the appropriate test in this situation. The steps in the procedure follow.

1. Specify the null and alternative hypotheses.
   - Null hypothesis H_0: $M \leq 3$ (M = response on rating scale).
   - Alternative hypothesis H_a: $M > 3$.
2. Specify the level of sampling error (α) allowed. For $\alpha = .05$, the table value of Z(critical) = 1.64. (See Exhibit 3 in Appendix 2 for d.f. = ∞, .05 significance, one-tail. The table for t is used because $t = Z$ for samples greater than 30.) Management's need to be very confident that the mean rating is significantly higher than 3 is interpreted to mean that the chance of being wrong because of sampling error should be no more than .05 (an α of .05).
3. Determine the sample standard deviation (S), which is given as $S = 1.90$.
4. Calculate the estimated standard error of the mean, using the formula

$$S_{\overline{X}} = \frac{S}{\sqrt{n}}$$

where $S_{\overline{X}}$ = estimated standard error of the mean

In this case,

$$S_{\overline{X}} = \frac{1.9}{\sqrt{200}} = 0.13$$

5. Calculate the test statistic:

$$Z = \frac{(\text{Sample mean}) - \begin{pmatrix}\text{Population mean specified} \\ \text{under the null hypothesis}\end{pmatrix}}{\text{Estimated standard error of the mean}}$$

$$= \frac{3.4 - 3}{0.13} = 3.07$$

6. State the result. *The null hypothesis can be rejected* because the calculated Z value (3.07) is larger than the critical Z value (1.64). Management of Video Connection can infer with 95 percent confidence that its video stores' mean rating is significantly higher than 3.

t Test

As noted earlier, for small samples ($n < 30$), the ***t* test** with $n - 1$ degrees of freedom is the appropriate test for making statistical inferences. The t distribution also is theoretically correct for large samples ($n \geq 30$). However, it approaches and becomes indistinguishable from the normal distribution for samples of 30 or more observations. Although the Z test is generally used for large samples, nearly all statistical packages use the t test for all sample sizes.

***t* test**
Hypothesis test used for a single mean if the sample is too small to use the Z test.

To see the application of the t test, consider a soft drink manufacturer that test markets a new soft drink in Denver. Twelve supermarkets in that city are selected at random, and the new soft drink is offered for sale in these stores for a limited period. The company estimates that it must sell more than 1,000 cases per week in each store for the brand to be profitable enough to warrant large-scale introduction. Actual average sales per store per week for the test are shown below.

Here is the procedure for testing whether sales per store per week are more than 1,000 cases:

1. Specify the null and alternative hypotheses.
   - Null hypothesis H_0: $\overline{X} \leq 1{,}000$ cases per store per week ($\overline{X}$ = average sales per store per week)
   - Alternative hypothesis H_a: $X > 1{,}000$ cases per store per week

Store	Average Sales per Week (X_i)
1	870
2	910
3	1,050
4	1,200
5	860
6	1,400
7	1,305
8	890
9	1,250
10	1,100
11	950
12	1,260

$$\text{Mean sales per week, } \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = 1087.1$$

2. Specify the level of sampling error (α) allowed. For $\alpha = .05$, the table value of t(critical) = 1.796. (See Exhibit 3 in Appendix 2 for $12 - 1 = 11$ d.f., $\alpha = .05$, one-tail test. A one-tailed t test is appropriate because the new soft drink will be introduced on a large scale only if sales per week are more than 1,000 cases.)
3. Determine the sample standard deviation (S) as follows:

$$S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n - 1}}$$

where X_i = observed sales per week in ith store

$\overline{X}$ = average sales per week

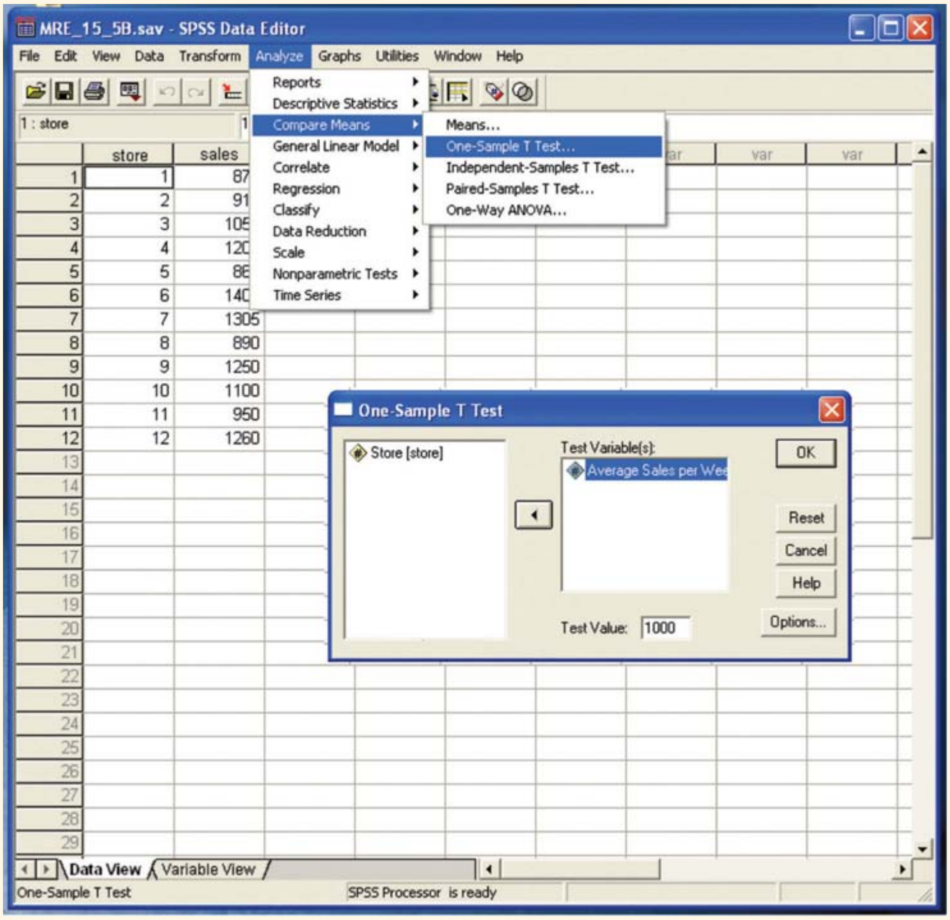
n = number of stores

For the sample data,

$$S = \sqrt{\frac{403{,}822.9}{(12 - 1)}} = 191.6$$

SPSS JUMP START FOR *T* TEST

Steps that you need to go through to do the *T* test problem shown in the book are provided below along with the output produced. Use the data set **TTestex**, which you can download from the Web site for the text.



1. Select *Analyze* → *Compare Means* → *One-Sample T Test.*
2. Move **sales** to Test Variable(s).
3. Input 1000 after Test Value.
4. Click OK.

surveysolutionsXP

SPSS Output for T Test

T-Test

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Average Sales per Week	12	1087.08	191.602	55.311

One-Sample Test

	Test Value = 1000					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Average Sales per Week	1.574	11	.144	87.083	-34.65	208.82

Note:

SPSS here only lists the significance for a two-tailed test. We need the significance of a one-tailed test, which is half this. 072. .072 is greater than the _ = .05 so fail to reject the null hypothesis.

4. Calculate the estimated standard error of the mean ($S_{\overline{X}}$), using the following formula:

$$S_{\overline{X}} = \frac{S}{\sqrt{n}}$$

$$= \frac{191.6}{\sqrt{12}} = 55.3$$

5. Calculate the t-test statistic:

$$t = \frac{(\text{Sample mean}) - \left(\begin{array}{c}\text{Population mean}\\ \text{under the null hypothesis}\end{array}\right)}{\text{Estimated standard error of the mean}}$$

$$= \frac{1{,}087.1 - 1000}{55.31} = 1.6$$

6. State the result. *The null hypothesis cannot be rejected* because the calculated value of t is less than the critical value of t. Although mean sales per store per week ($\overline{X} = 1087.1$) are higher than 1,000 units, the difference is not statistically significant, based on the 12 stores sampled. On the basis of this test and the decision criterion specified, the large-scale introduction of the new soft drink is not warranted.

Hypotheses about Two Means

Marketers are frequently interested in testing differences between groups. In the following example of testing the differences between two means, the samples are independent.

The management of a convenience store chain is interested in differences between the store visit rates of men and women. Believing that men visit convenience stores more often than women, management collected data on convenience store visits from 1,000 randomly selected consumers. Testing this hypothesis involves the following steps:

1. Specify the null and alternative hypotheses.

- Null hypothesis H_0: $M_m - M_f \leq 0$; the mean visit rate of men (M_m) is the same as or less than the mean visit rate of women (M_f).
- Alternative hypothesis H_a: $M_m - M_f > 0$; the mean visit rate of men (M_m) is higher than the mean visit rate of women (M_f).

The observed difference in the two means (Exhibit 15.4) is $11.49 - 8.51 = 2.98$.

2. Set the level of sampling error (α). The managers decided that the acceptable level of sampling error for this test is $\alpha = .05$. For $\alpha = .05$, the table value of Z(critical) $= 1.64$. (See Exhibit 3 in Appendix 2 for d.f. $= \infty$, .05 significance, one-tail. The table for t is used because $t = Z$ for samples greater than 30.)

3. Calculate the estimated standard error of the differences between the two means as follows:

$$S_{X_{m-f}} = \sqrt{\frac{S_m^2}{n_m} + \frac{S_f^2}{n_f}}$$

where

S_m = estimated standard deviation of population m (men)

S_f = estimated standard deviation of population f (women)

n_m = sample size for sample m

n_f = sample size for sample f

Therefore,

$$S_{X_{m-f}} = \sqrt{\frac{(8.16)^2}{45} + \frac{(5.23)^2}{71}} = 1.37$$

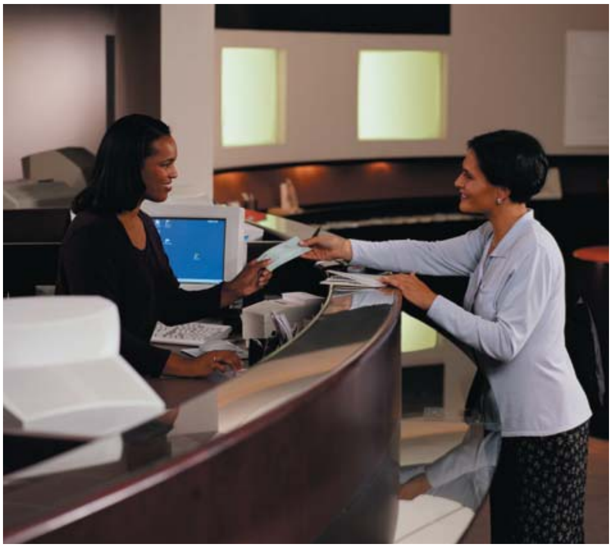
Note that this formula is for those cases in which the two samples have unequal variances. A separate formula is used when the two samples have equal variances. When this test is run in SAS and many other statistical packages, two t values are provided—one for each variance assumption.

4. Calculate the test statistic Z as follows:

$$Z = \frac{\begin{pmatrix}\text{Difference between means}\\ \text{of first and second sample}\end{pmatrix} - \begin{pmatrix}\text{Difference between means}\\ \text{under the null hypothesis}\end{pmatrix}}{\text{Standard error of the differences between the two means}}$$
$$= \frac{(11.49 - 8.51) - 0}{1.37} = 2.18$$

Before launching new services designed for families with an annual income of more than $50,000, the bank needs to be certain about the percentage of its customers who meet or exceed this threshold income.

5. State the result. The calculated value of Z (2.18) is larger than the critical value (1.64), so *the null hypothesis is rejected*. Management can conclude with 95 percent confidence ($1 - \alpha = .95$) that, on average, men visit convenience stores more often than do women.



Hypotheses about Proportions

In many situations, researchers are concerned with phenomena that are expressed in terms of percentages.[8] For example, marketers might be interested in testing for the proportion of respondents who prefer brand A versus those who prefer brand B or those who are brand loyal versus those who are not.

Proportion in One Sample

A survey of 500 customers conducted by a major bank indicated that slightly more than 74 percent had family incomes of more than $50,000 per year. If this is true, the bank will develop a special package of services for this group. Before developing and introducing the new package of services, management wants to determine whether the true percentage is greater than 60 percent. The survey results show that 74.3 percent of the bank's customers surveyed reported family incomes of $50,000 or more per year. The procedure for the **hypothesis test of proportions** follows:

hypothesis test of proportions Test to determine whether the difference between proportions is greater than would be expected because of sampling error.

1. Specify the null and alternative hypotheses.
   - Null hypothesis H_0: $P \leq .60$.
   - Alternative hypothesis H_a: $P > .60$ (P = proportion of customers with family incomes of $50,000 or more per year).
2. Specify the level of sampling error (α) allowed. For $\alpha = .05$, the table value of Z(critical) = 1.64. (See Exhibit 3 in Appendix 2 for d.f. = ∞, .05 significance, one-tail. The table for t is used because $t = Z$ for samples greater than 30.)
3. Calculate the estimated standard error, using the value of P specified in the null hypothesis:

$$S_p = \sqrt{\frac{P(1-P)}{n-1}}$$

where P = proportion specified in the null hypothesis
n = sample size

Therefore,

$$S_p = \sqrt{\frac{.6(1 - .6)}{35 - 1}} = .084$$

4. Calculate the test statistic as follows:

$$Z = \frac{(\text{Observed proportion} - \text{Proportion under null hypothesis})}{\text{Estimated standard error } (S_p)}$$

$$= \frac{0.7429 - 0.60}{.084} = 1.7$$

The *null hypothesis is rejected* because the calculated Z value is larger than the critical Z value. The bank can conclude with 95 percent confidence ($1 - \alpha = .95$) that more than 60 percent of its customers have family incomes of $50,000 or more. Management can introduce the new package of services targeted at this group.

Two Proportions in Independent Samples

In many instances, management is interested in the difference between the proportions of people in two different groups who engage in a certain activity or have a certain characteristic. For example, management of a convenience store chain had reason to believe, on the basis of a research study, that the percentage of men who visit convenience stores nine or more times per month (heavy users) is larger than the percentage of women who do so. The specifications required and the procedure for testing this hypothesis are as follows.

1. Specify the null and alternative hypotheses:
   - Null hypothesis H_0: $P_m - P_f \leq 0$; the proportion of men (P_m) reporting nine or more visits per month is the same as or less than the proportion of women (P_f) reporting nine or more visits per month.
   - Alternative hypothesis H_a: $P_m - P_f > 0$; the proportion of men (P_m) reporting nine or more visits per month is greater than the proportion of women (P_f) reporting nine or more visits per month.

   The sample proportions and the difference can be calculated from Exhibit 15.4 as follows:

$$P_m = \frac{26}{45} = .58$$

$$P_f = \frac{30}{71} = .42$$

$$P_m - P_f = .58 - .42 = .16$$

2. Set the level of sampling error α at .10 (management decision). For $\alpha = .10$, the table value of Z(critical) $= 1.28$. (See Exhibit 3 in Appendix 2 for d.f. $= \infty$, .10 significance, one-tail. The table for t is used because $t = Z$ for samples greater than 30.)

3. Calculate the estimated standard error of the differences between the two proportions as follows:

$$S_{P_{m-f}} = \sqrt{P(1 - P)\left(\frac{1}{n_m} + \frac{1}{n_f}\right)}$$

$$\text{where} \qquad P = \frac{n_m P_m + n_f P_f}{n_m + n_f}$$

P_m = proportion in sample m (men)
P_f = proportion in sample f (women)
n_m = size of sample m
n_f = size of sample f

Therefore,

$$P = \frac{45(.58) + 71(.41)}{45 + 71} = .48$$

$$\text{and} \qquad S_{P_{m-f}} = \sqrt{.48(1 - .48)\left(\frac{1}{45} + \frac{1}{71}\right)} = .10$$

4. Calculate the test statistic.

$$Z = \frac{\begin{pmatrix}\text{Difference between}\\ \text{observed proportions}\end{pmatrix} - \begin{pmatrix}\text{Difference between proportions}\\ \text{under the null hypothesis}\end{pmatrix}}{\begin{array}{c}\text{Estimated standard error of the differences}\\ \text{between the two means}\end{array}}$$

$$= \frac{(.58 - .42) - 0}{.10} = 1.60$$

5. State the result. *The null hypothesis is rejected* because the calculated Z value (1.60) is larger than the critical Z value (1.28 for $\alpha = .10$). Management can conclude with 90 percent confidence ($1 - \alpha = .90$) that the proportion of men who visit convenience stores nine or more times per month is larger than the proportion of women who do so.

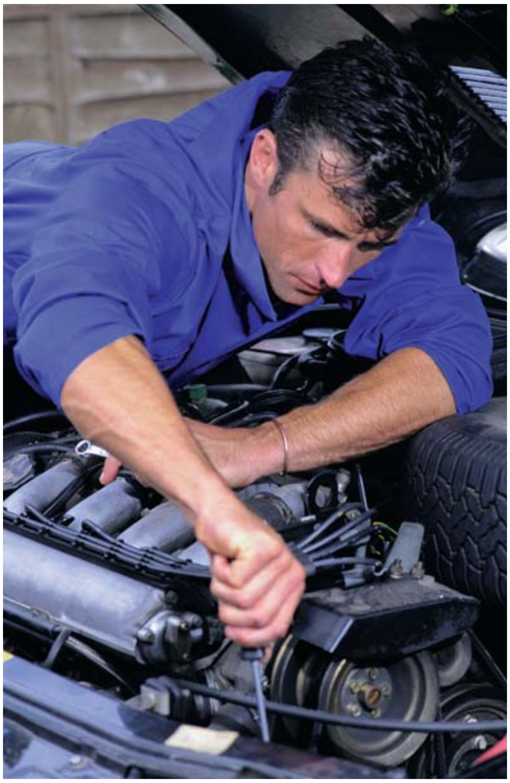
It should be noted that if the level of sampling error α had been set at .05, the critical Z value would equal 1.64. In this case, we would fail to reject (FTR) the null hypothesis because Z(calculated) would be smaller than Z(critical).

Analysis of Variance (ANOVA)

analysis of variance (ANOVA) Test for the differences among the means of two or more independent samples.

When the goal is to test the differences among the means of two or more independent samples, **analysis of variance (ANOVA)** is an appropriate statistical tool. Although it can be used to test differences between two means, ANOVA is more commonly used for hypothesis tests regarding the differences among the means of several (C) independent groups (where $C \geq 3$). It is a statistical technique that permits the researcher to determine whether the variability among or across the C sample means is greater than expected because of sampling error.

The Z and t tests described earlier normally are used to test the null hypothesis when only two sample means are involved. However, in situations in which there are three or more samples, it would be inefficient to test differences between the means two at a time. With five samples and associated means, 10 t tests would be required to test all pairs of means. More important, the use of Z or t tests in situations involving three or more means increases the probability of a type 1 error. Because these tests must be performed for all possible pairs of means, the more pairs, the more tests that must be performed. And the more tests performed, the more likely it is that one or more tests will show significant differences that are really due to sampling error. At an α of .05, this could be expected to occur in 1 of 20 tests on average.



A brake shop might use analysis of variance to analyze experimental results with respect to several new services before deciding on a particular new service to offer.

One-way ANOVA is often used to analyze experimental results. Suppose the marketing manager for a chain of brake shops was considering three different services for a possible in-store promotion: wheel alignment, oil change, and tune-up. She was interested in knowing whether there were significant differences in potential sales of the three services.

Sixty similar stores (20 in each of three cities) were selected at random from among those operated by the chain. One of the services was introduced in each of three cities. Other variables under the firm's direct control, such as price and advertising, were kept at the same level during the course of the experiment. The experiment was conducted for a 30-day period, and sales of the new services were recorded for the period.

Average sales for each shop are shown below. The question is, are the differences among the means larger than would be expected due to chance?

Chicago (Wheel Alignment)		Cleveland (Oil Change)		Detroit (Tune-Up)	
310	318	314	321	337	310
315	322	315	340	325	312
305	333	350	318	330	340
310	315	305	315	345	318
315	385	299	322	320	322
345	310	309	295	325	335
340	312	299	302	328	341
330	308	312	316	330	340
320	312	331	294	342	320
315	340	335	308	330	310
$\overline{X} = 323$		$\overline{X} = 315$		$\overline{X} = 328$	

1. Specify the null and alternative hypotheses.

- ☐ Null hypothesis H_0: $M_1 = M_2 = M_3$; mean sales of the three items are equal.
- ☐ Alternative hypothesis H_a: The variability in group means is greater than would be expected because of sampling error.

2. Sum the squared differences between each subsample mean ($\bar{X}_j$) and the overall sample mean ($\bar{X}_t$), weighted by sample size (n_j). This is called the *sum of squares among groups* or among group variation (SSA). SSA is calculated as follows:

$$\text{SSA} = \sum_{j=1}^{C} n_j(\bar{X}_j - \bar{X}_t)^2$$

In this example, the overall sample mean is

$$\bar{X}_t = \frac{20(323) + 20(315) + 20(328)}{60} = 322$$

Thus,

$$\begin{aligned}\text{SSA} &= 20(323 - 322)^2 + 20(315 - 322)^2 + 20(328 - 322)^2 \\ &= 1720\end{aligned}$$

The greater the differences among the sample means, the larger the SSA will be.

3. Calculate the variation among group means as measured by the *mean sum of squares among groups* (MSA). The MSA is calculated as follows:

$$\text{MSA} = \frac{\text{Sum of squares among groups (SSA)}}{\text{Degrees of freedom (d.f.)}}$$

$$\text{where} \qquad \text{Degrees of freedom} = \text{number of groups } (C) - 1$$

In this example,

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

Thus,

$$\text{MSA} = \frac{1720}{2} = 860$$

4. Sum the squared differences between each observation (X_{ij}) and its associated sample mean ($\bar{X}_j$), accumulated over all C levels (groups). Also called the *sum of squares within groups* or *within group variation*, it is generally referred to as the *sum of squared error* (SSE). For this example, the SSE is calculated as follows:

$$\begin{aligned}\text{SSE} &= \sum_{j=1}^{C} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2 \\ &= (6644) + (4318) + (2270) = 13{,}232\end{aligned}$$

5. Calculate the variation within the sample groups as measured by the mean sum of squares within groups. Referred to as *mean square error* (MSE), it represents an estimate of the random error in the data. The MSE is calculated as follows:

$$\text{MSE} = \frac{\text{Sum of squares within groups (SSE)}}{\text{Degrees of freedom (d.f.)}}$$

The number of degrees of freedom is equal to the sum of the sample sizes for all groups minus the number of groups (C):

$$\begin{aligned}\text{d.f.} &= \left(\sum_{j=1}^{K} n_j\right) - C \\ &= (20 + 20 + 20) - 3 = 57\end{aligned}$$

Thus,

$$\text{MSE} = \frac{13{,}232}{57} = 232.14$$

As with the Z distribution and t distribution, a sampling distribution known as the *F distribution* permits the researcher to determine the probability that a particular calculated value of F could have occurred by chance rather than as a result of the treatment effect. The F distribution, like the t distribution, is really a set of distributions whose shape changes slightly depending on the number and size of the samples involved. To use the ***F* test**, it is necessary to calculate the degrees of freedom for the numerator and the denominator.

***F* test**
Test of the probability that a particular calculated value could have been due to chance.

6. Calculate the F statistic as follows:

$$\begin{aligned}F &= \frac{\text{MSA}}{\text{MSE}} \\ &= \frac{860}{232.14} = 3.70\end{aligned}$$

The numerator is the MSA, and the number of degrees of freedom associated with it is 2 (step 3). The denominator is the MSE, and the number of degrees of freedom associated with it is 57 (step 5).

7. State the results. For an alpha of .05, the table value of F(critical) with 2 (numerator) and 57 (denominator) degrees of freedom is approximately 3.15. (See Exhibit 5 in Appendix 2 for d.f. for denominator = 5, d.f. for numerator = 2, .05 significance.) The calculated F value (3.70) is greater than the table value (3.15), and so *the null hypothesis is rejected*. By rejecting the null hypothesis, we conclude that the variability observed in the three means is greater than expected due to chance.

The results of an ANOVA generally are displayed as follows:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F* Statistic
Treatments	1,720 (SSA)	2 ($C - 1$)	860 (MSA)	3.70 calculated
Error	13,232 (SSE)	57 ($n - C$)	232.14 (MSE)	
Total	14,592 (SST)	59 ($n - 1$)		

p Values and Significance Testing

For the various tests discussed in this chapter, a standard—a level of significance and associated critical value of the statistics—is established, and then the value of the statistic is calculated to see whether it beats that standard. If the calculated value of the statistic exceeds the critical value, then the result being tested is said to be statistically significant at that level.

However, this approach does not give the exact probability of getting a computed test statistic that is largely due to chance. The calculations to compute this probability, commonly referred to as the ***p* value**, are tedious to perform by hand. Fortunately, they are easy for computers. The *p* value is the most demanding level of statistical (not managerial) significance that can be met, based on the calculated value of the statistic. Computer statistical packages usually use one of the following labels to identify the probability that the distance between the hypothesized population parameter and the observed test statistic could have occurred due to chance:

***p* value**
Exact probability of getting a computed test statistic that is due to chance. The smaller the *p* value, the smaller the probability that the observed result occurred by chance.

- *p* value
- ≤ PROB
- PROB =

The smaller the *p* value, the smaller is the probability that the observed result occurred by chance (sampling error).

An example of computer output showing a *p* value calculation appears in Exhibit 15.6. This analysis shows the results of a *t* test of the differences between means for two independent samples. In this case, the null hypothesis H_0 is that there is no difference between what men and women would be willing to pay for a new communications service. (The variable name is GENDER, with the numeric codes of 0 for males and 1 for females. Subjects were asked how much they would be willing to pay per month for a new wireless communications service that was described to them via a videotape. Variable ADDEDPAY is their response to the question.) The results show that women are willing to pay an average of $16.82 for the new service and men are willing to pay $20.04. Is this a significant difference? The calculated value for *t* of −1.328 indicates, via the associated *p* value of .185, that there is an 18.5 percent chance that the difference is due to sampling error. If, for example, the standard for the test were set at .10 (willing to accept a 10 percent chance of incorrectly rejecting H_0), then the analyst would *fail to reject* H_0 in this case.

EXHIBIT 15.6 Sample t-Test Output

```
Stat.       Grouping: GENDER (pcs. sta)
Basic       Group 1: G_1:1
Stats       Group 2: G_2:0
```

Variable	Mean G_1:1	Mean G_2:0	*t*-value	df	P	Valid N G_1:1	Valid N G_2:0
ADDED PAY	16.82292	20.04717	−1.32878	200	.185434	96	106

SUMMARY

The purpose of making statistical inferences is to generalize from sample results to population characteristics. Three important concepts applied to the notion of differences are mathematical differences, managerially important differences, and statistical significance.

A hypothesis is an assumption or theory that a researcher or manager makes about some characteristic of the population being investigated. By testing, the researcher determines whether a hypothesis concerning some characteristic of the population is valid. A statistical hypothesis test permits the researcher to calculate the probability of observing the particular result if the stated hypothesis actually were true. In hypothesis testing, the first step is to specify the hypothesis. Next, an appropriate statistical technique should be selected to test the hypothesis. Then, a decision rule must be specified as the basis for determining whether to reject or fail to reject the hypothesis. Hypothesis tests are subject to two types of errors called type I (α error) and type II (β error). A type I error involves rejecting the null hypothesis when it is, in fact, true. A type II error involves failing to reject the null hypothesis when the alternative hypothesis actually is true. Finally, the value of the test statistic is calculated, and a conclusion is stated that summarizes the results of the test.

Marketing researchers often develop cross tabulations, whose purpose usually is to uncover interrelationships among the variables. Usually the researcher needs to determine whether the numbers of subjects, objects, or responses that fall into some set of categories differ from those expected by chance. Thus, a test of goodness of fit of the observed distribution in relation to an expected distribution is appropriate. One common test of goodness of fit is chi square.

Often marketing researchers need to make inferences about a population mean. If the sample size is equal to or greater than 30 and the sample comes from a normal population, the appropriate test statistic for testing hypotheses about means is the Z test. For small samples, researchers use the t test with $n - 1$ degrees of freedom when making inferences (n is the size of the sample).

When researchers are interested in testing differences between responses to the same variable, such as advertising, by groups with different characteristics, they test for differences between two means. A Z value is calculated and compared to the critical value of Z. Based on the result of the comparison, they either reject or fail to reject the null hypothesis. The Z test also can be used to examine hypotheses about proportions from one sample or independent samples.

When researchers need to test for differences among the means of three or more independent samples, analysis of variance is an appropriate statistical tool. It is often used for hypothesis tests regarding the differences among the means of several independent groups. It permits the researcher to test the null hypothesis that there are no significant differences among the population group means.

KEY TERMS & DEFINITIONS

hypothesis Assumption or theory that a researcher or manager makes about some characteristic of the population under study.

type I error (α error) Rejection of the null hypothesis when, in fact, it is true.

type II error (β error) Failure to reject the null hypothesis when, in fact, it is false.

independent samples Samples in which measurement of a variable in one population has no effect on measurement of the variable in the other.

related samples Samples in which measurement of a variable in one population may influence measurement of the variable in the other.

chi-square test Test of the goodness of fit between the observed distribution and the expected distribution of a variable.

***Z* test** Hypothesis test used for a single mean if the sample is large enough and drawn at random.

***t* test** Hypothesis test used for a single mean if the sample is too small to use the *Z* test.

hypothesis test of proportions Test to determine whether the difference between proportions is greater than would be expected because of sampling error.

analysis of variance (ANOVA) Test for the differences among the means of two or more independent samples.

***F* test** Test of the probability that a particular calculated value could have been due to chance.

***p* value** Exact probability of getting a computed test statistic that is due to chance. The smaller the *p* value, the smaller the probability that the observed result occurred by chance.

QUESTIONS FOR REVIEW & CRITICAL THINKING

1. Explain the notions of mathematical differences, managerially important differences, and statistical significance. Can results be statistically significant and yet lack managerial importance? Explain your answer.
2. Describe the steps in the procedure for testing hypotheses. Discuss the difference between a null hypothesis and an alternative hypothesis.
3. Distinguish between a type I error and a type II error. What is the relationship between the two?
4. What is meant by the terms *independent samples* and *related samples*? Why is it important for a researcher to determine whether a sample is independent?
5. Your university library is concerned about student desires for library hours on Sunday morning (9:00 A.M.–12:00 P.M.). It has undertaken to survey a random sample of 1,600 undergraduate students (one-half men, one-half women) in each of four status levels (i.e., 400 freshmen, 400 sophomores, 400 juniors, 400 seniors). If the percentages of students preferring Sunday morning hours are those shown below, what conclusions can the library reach?

	Seniors	Juniors	Sophomores	Freshmen
Women	70	53	39	26
Men	30	48	31	27

6. A local car dealer was attempting to determine which premium would draw the most visitors to its showroom. An individual who visits the showroom and takes a test drive is given a premium with no obligation. The dealer chose four premiums and offered each for one week. The results are as follows.

Week	Premium	Total Given Out
1	Four-foot metal stepladder	425
2	$50 savings bond	610
3	Dinner for four at a local steak house	510
4	Six pink flamingos plus an outdoor thermometer	705

Using a chi-square test, what conclusions can you draw regarding the premiums?

7. A market researcher has completed a study of pain relievers. The following table depicts the brands purchased most often, broken down by men versus women. Perform a chi-square test on the data and determine what can be said regarding the cross tabulation.

Pain Relievers	Men	Women
Anacin	40	55
Bayer	60	28
Bufferin	70	97
Cope	14	21
Empirin	82	107
Excedrin	72	84
Excedrin PM	15	11
Vanquish	20	26

8. A child psychologist observed 8-year-old children behind a one-way mirror to determine how long they would play with a toy medical kit. The company that designed the toy was attempting to determine whether to give the kit a masculine or a feminine orientation. The lengths of time (in minutes) the children played with the kits are shown below. Calculate the value of Z and recommend to management whether the kit should have a male or a female orientation.

Boys	Girls	Boys	Girls
31	26	67	9
12	38	67	9
41	20	25	16
34	32	73	26
63	16	36	81
7	45	41	20
		15	5

9. American Airlines is trying to determine which baggage handling system to put in its new hub terminal in San Juan, Puerto Rico. One system is made by Jano Systems, and the second is manufactured by Dynamic Enterprises. American has installed a small Jano system and a small Dynamic Enterprises system in two of its low-volume terminals. Both terminals handle approximately the same quantity of baggage each month. American has decided to select the system that provides the minimum number of instances in which passengers disembarking must wait 20 minutes or longer for baggage. Analyze the data that follow and determine whether there is a significant difference at the .95 level of confidence between the two systems. If there is a difference, which system should American select?

Minutes of Waiting	Jano Systems (Frequency)	Dynamic Enterprises (Frequency)
10–11	4	10
12–13	10	8
14–15	14	14
16–17	4	20
18–19	2	12
20–21	4	6
22–23	2	12

24–25	14	4
26–27	6	13
28–29	10	8
30–31	12	6
32–33	2	8
34–35	2	8
36 or more	2	2

10. Menu space is always limited in fast-food restaurants. However, McDonald's has decided that it needs to add one more salad dressing to its menu for its garden salad and chef salad. It has decided to test-market four flavors: Caesar, Ranch-Style, Green Goddess, and Russian. Fifty restaurants were selected in the North-Central region to sell each new dressing. Thus, a total of 200 stores were used in the research project. The study was conducted for 2 weeks; the units of each dressing sold are shown below. As a researcher, you want to know if the differences among the average daily sales of the dressings are larger than can be reasonably expected by chance. If so, which dressing would you recommend be added to the inventory throughout the United States?

Day	Caesar	Ranch-Style	Green Goddess	Russian
1	155	143	149	135
2	157	146	152	136
3	151	141	146	131
4	146	136	141	126
5	181	180	173	115
6	160	152	170	150
7	168	157	174	147
8	157	167	141	130
9	139	159	129	119
10	144	154	167	134
11	158	169	145	144
13	184	195	178	177
14	161	177	201	151

WORKING THE NET

1. Go to ***http://www.ifigure.com/index.html*** and click on the highlighted word "article" in the paragraph of text. This will take you to an article that provides examples of the wide range of situations for which online calculators are available. To locate a tool, use the menu or the search box at the top left of the page.

2. Go to ***http://www.ifigure.com/math/stat/testing.htm*** and review the range of statistical calculators available. Use the *Hypothesis Testing of Means* calculator to address the following problem. The Fort Worth Cats minor league baseball team tracked actual hot dog sales several years ago and found (by dividing actual sales by the total number of attendees) that the average fan consumed 1.2 hot dogs per game. They quit tracking actual sales shortly after that. Now they need to figure out what is happening to hot dog sales. To do this, they took samples of fans at three randomly selected games. Four hundred were sampled at the end of each game, and they got an average consumption figure of 1.4 hot dogs with a standard deviation of .96 hot dogs. Can they conclude that consumption has increased? In the calculator, use:

- 1.4 for the sample value of the mean
- 1.2 for the hypothesized value of the mean
- .96 for the standard deviation
- 1,200 for the sample size

What is your answer?

Experiment with other values of the inputs for the calculator and see how the changes affect the result.

REAL-LIFE RESEARCH • 15.1

I Can't Believe It's Yogurt

Phil Jackson, research manager for I Can't Believe It's Yogurt (ICBIY), is trying to develop a more rational basis for evaluating alternative store locations. ICBIY has been growing rapidly, and historically the issue of store location has not been critical. It didn't seem to matter where stores were located—all were successful. However, the yogurt craze has faded, and some of its new stores and a few of its old ones are experiencing difficulties in the form of declining sales.

ICBIY wants to continue expanding but recognizes that it must be much more careful in selecting locations than it was in the past. It has determined that the percentage of individuals in an area who have visited a frozen yogurt store in the past 30 days is the best predictor of the potential for one of its stores—the higher that percentage, the better.

ICBIY wants to locate a store in Denver and has identified two locations that, on the basis of the other criteria, look good. It has conducted a survey of households in the areas that would be served from each location. The results of that survey are shown below.

Yogurt Store Patronage	Both Areas	Area A	Area B
Have patronized in past 30 days	465	220	245
Have not patronized	535	280	255

Questions

1. Determine whether there is a significant difference at the .05 level between the two areas.
2. Based on this analysis, what would you recommend to ICBIY regarding which of the two areas it should choose for the new store? Explain your recommendation.

REAL-LIFE RESEARCH • 15.2

New Mexico Power

Marc Guerraz is the new marketing research director for New Mexico Power, an investor-owned, vertically integrated electric utility involved in the generation, transmission, distribution, and energy service domains of the electricity industry. Marc was hired to help New Mexico Power transition to a competitive environment for energy services, which will occur in one year. In this environment, customers will be allowed to select their provider of electricity, much as customers now select long-distance telecommunications service from providers such as Sprint, Verizon, and SBC.

One of the crucial questions for New Mexico Power is that of current customer retention. How many current customers will switch to another provider of electric services during the first 6 months of competition? To address this question, Marc and his team designed and fielded a survey in the current service territory of New Mexico Power. They had 500 customers give complete answers to all questions. Initial results indicate that 22 percent of customers would switch. The margin of error is 4 percent, which means that (at the 95 percent confidence level) the actual percentage of customers switching could be as low as 18 percent or as high as 26 percent. New Mexico Power senior management is concerned about this error range of 4 percent, which means that error spans a total of 8 percentage points. Further customer retention efforts must be budgeted now, and New Mexico Power senior management wants firmer numbers on which to base strategies and budgets.

Questions

1. How could the error range be reduced without collecting more data? Would you recommend taking this approach? Why/Why not?
2. Do you think New Mexico Power senior management would find this approach to reducing the error range satisfactory?
3. If 500 more respondents were surveyed and 30 percent of them indicated that they would switch, what would the error range become?

surveysolutionsXP

SPSS EXERCISES FOR CHAPTER 15

surveysolutionsXP

Exercise #1: Analyzing Data Using Cross-tabulation Analysis

Note: Go to the Wiley Web site at *www.wiley.com/college/mcdaniel* and download the *Segmenting the College Student Market for Movie Attendance* database to SPSS windows.

Use the *analyze/descriptive statistics/crosstab* sequence to obtain cross-tabulated results. In addition, click on the "cell" icon and make sure the *observed, expected, total, row, and column* boxes are checked. Then, click on the "statistics" icon and check the *chi-square* box. Once you run the analysis, on the output for the chi-square analysis, you will only need the *Pearson chi-square statistic* to assess whether or not the results of the crosstab are statistically significant.

In this exercise we are assessing whether or not persons who attend movies at movie theaters are demographically different from those who do not. Invoke the crosstab analysis for the following pairs of variables:

a. Q1 & Q11
b. Q1 & Q12
c. Q1 & Q13
d. Q1 & Q14

Answer questions 1–6 <u>using only the sample data</u>. <u>Do not consider</u> the results of the *chi-square test.*

1. What % of males do not attend movies at movie theaters? __________%
2. What % of all respondents are African American and do not attend movies at movie theaters? __________%
3. What % of respondents not attending movies at movie theaters are in the 19–20 age category? __________%
4. Which classification group is most likely to attend movies at movie theaters? __________
5. Which age category is least likely to attend movies at a movie theater? __________
6. Are Caucasians less likely to attend movie theaters than African Americans? __________

For question 7, the objective is to determine statistically whether, in the population from which the sample data was drawn, there were demographic differences in persons who attend and do not attend movies at movie theaters. We do this by using the results of the *chi-square test for independent samples.*

7. Evaluate the chi-square statistic in each of your crosstab tables. Construct a table to summarize the results. For example:

Variables	Pearson Chi-Square	Degrees of Freedom	Asymp sig.	Explanation
Q1 (attend or not attend movies at movie theaters & Q12 (gender)	2.71	1	.10	We can be 90% confident that based on our sample results, males differ significantly from females in their tendency to attend or not attend movies at movie theaters.

Exercise #2: T/Z Test for Independent Samples

Use the *analyze/compare means/independent samples t-test* sequence to complete this exercise. This exercise compares males and females regarding the information sources they utilize to search for information about movies at movie theaters. SPSS calls the variable in which the means are being computed the *test variable,* and the variable in which we are grouping responses the *grouping variable.*

Note: In statistics, if a sample has fewer than 30 observations or cases, then we invoke a *t* test. If there are 30 or more cases, we invoke a *z* test, as *the t test values and z test values are virtually the same; hence SPSS refers only to a t test.*

Answer the following questions.

The result of the *t* test generates a table of ***group statistics***, which is based only on the ***sample*** data. The other output table generated by the *t* test has statistical data from which we can determine whether or not the sample results can be generalized to the population from surveysolutionsXP

which the sample data was drawn. If the *t* test is significant, then we can use the group statistics to determine the specifics of the computed results. For example, a significant *t* test may tell us that males differ from females regarding the importance they place on the newspaper as an information source, but the group statistics tell us "who" considers it most important.

From our *sample data*, can we generalize our results to the population by saying that males differ from females regarding the importance they place on various information sources to get information about movies at movie theaters by:

1. the newspaper (Q7a)?
2. the Internet (Q7b)?
3. phoning in to the movie theater for information (Q7c)?
4. the television (Q7d)?
5. friends or family (Q7e)?

You may want to use the template below to summarize your *t* test results. For example:

Variables	Variance Prob of Sig Diff	Means Prob of Sig Diff	Interpretation of Results
Q12 (gender) & Q7a (newspaper)	.000	.035	96.5% confident that based on our sample results, males differ significantly from females concerning the importance they place on the newspaper as an information source about movies at movie theaters **(means test).** 100% confident that males and females were significantly different regarding the variance of response within each gender **(variance test)**.

Exercise #3: ANOVA Test for Independent Samples

Invoke the *analyze/compare means/One-Way ANOVA* sequence to invoke the ANOVA test to complete this exercise. This exercise compares the responses of freshman, sophomores, juniors, seniors, and graduate students to test for significant differences in the importance placed on several movie theater items. For the ANOVA test, SPSS calls the variable in which means are being computed the *independent variable* and the variable in which we are grouping responses the *factor variable.* <u>Be sure to click the *options* icon and check *descriptives* so that the output will produce the mean responses by student classification for the sample data.</u> As with the *t* test, the ANOVA test produces a table of *descriptives* based on sample data. If our ANOVA test is significant, the *descriptives* can be used to determine, for example, which student classification places the most importance on comfortable seats.

Answer the following questions.

From our sample data, can we generalize our results to the population by saying that there are significant differences across the <u>classification of students</u> by the importance they place on the following movie theatre items?

1. video arcade at the movie theater (Q5a)?
2. soft drinks and food items (Q5b)

3. plentiful restrooms (Q5c)
4. comfortable chairs (Q5d)
5. auditorium-type seating (Q5e)
6. size of the movie theater screens (Q5f)
7. quality of the sound system (Q5g)
8. number of screens at a movie theater (Q5h)
9. clean restroom (Q5i)
10. Using only the *descriptive statistics,* which classification group (Q13) places the least amount of importance on clean restrooms (Q5i)? ____________________
11. Using only the *descriptive statistics,* which classification group (Q13) places the greatest amount of importance on quality of sound system (Q5i)? ________________

Summarize the results of your ANOVA analysis using a table similar to the one below.

Variables	Degrees of Freedom	F-Value	Probability of Insignificance	Interpretation of Results
Q5a (importance of a video arcade) & Q13 (student classification)	4,461	12.43	.001	99.9% confident that based on the sample results, students differ significantly by classification concerning the importance placed on there being a video arcade at the movie theater.

surveysolutions XP