Monty Rakusen /Getty Images

Statistical Testing of Differences and Relationships

LEARNING OBJECTIVES

1. Learn how to evaluate differences and changes.

2. Understand the concept of hypothesis development and how to test hypotheses.

3. 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.

4. Learn the hypotheses about one mean.

5. Learn the hypotheses about two means.

6. Learn the hypotheses about proportions.

7. Learn about analysis of variance.

8. Understand the *P* values and significance testing.

CHAPTER

This chapter addresses statistical techniques that can be used to determine whether observed differences 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:

STATISTICAL TESTING OF DIFFERENCES AND RELATIONSHIPS

- Our posttest measure of top-of-mind awareness (first brand mentioned unaided) 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 common questions in marketing and marketing research.

Although con- sidered boring by some, statistical hypothesis testing is important because it helps research- ers get closer to the answers to these questions. We say "closer" because certainty is never achieved in answering these questions in marketing research.

Statistical Significance

statistical significance

A difference that is large enough that it is not likely to have occurred because of chance or sampling error.

■ ■ ■

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 statis- tically 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 statis- tically significant but yet so small as to have little practical or managerial significance.[1]

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 sta- tistical 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:

This chapter covers different approaches to testing whether results are statistically sig- nificant. The Practicing Marketing Research feature on page 397 covers issues related to statistical significance testing.

As you review the material in this chapter, keep three things in mind:

1. **Random samples are assumed.** All of the tests we will discuss in this chapter assume the data come from random samples. Some have additional assumptions, but all assume ran- dom samples. If the data you are working with do not come from random samples, then the significance tests are not appropriate.

2. **Big data does not mean "good" data.** Big data presents some special challenges. First of all, don't be totally swayed by the sheer amount of data you have, no matter how much data, it must come from random samples. With really big data—thousands, tens of thousands or hundreds of thousands of observations—if the obervations come from random samples, then very small differences will be statistically significant because sam- ple size always figures into the calculation of significance.

3. **Don't overrely on significance testing.** Placing total reliance on significance testing is not a good idea. If, on the one hand, we are testing many measures from a particular study conducted at different points in time to access the changes that have taken place, then some percentage will give false positives (indicate significant differences incorrectly). On the other hand, routinely dismissing differences that are not significant can lead us to miss important findings.

Let's Test Everything[2]

The logic of statistical (stat) testing is not complex, but it can be difficult to understand, because it is the reverse of everyday logic and what normal people expect. Basically, to determine if two numbers differ significantly, it is assumed that they are the same. The test then determines whether this notion can be rejected, and we can say that the num- bers are "statistically significantly different at the (some pre- determined) confidence level."

While it is not complex, the logic can be subtle. One subtlety leads to a common error, aided and abetted by automatic computer stat testing—overtesting. Suppose there is a group of 200 men and one of 205 women, and they respond to a new product concept on a purchase intent scale. The data might look like that shown in Table A.

Statistical logic assumes that the two percentages to be tested are from the same population—they do not dif- fer. Therefore, it is assumed that men have the same pur- chase interest as women. The rules also assume that the numbers are unrelated, in the sense that the percentages being tested are free to be whatever they might be, from 0 percent to 100 percent. Restricting them in any way changes the probabilities, and the dynamics of the statis- tical test.

Definitely would buy
Probably would buy
Def./Prob. would buy
Might or might not buy
Probably would not buy
Definitely would not buy
Total 100

21 19 S 40 35 20 5 100

PRACTICING MARKETING RESEARCH

Table A
Purchase Intent Among Men and Women

Base: total per group	Men (200) %	Women (205) %

3 10 13 40 30 17

S = the percentages differ significantly at the 95% confidence level.

The right way to test for a difference in purchase intent is to pick a key measure to summarize the responses, and test that measure. In Table A, the Top Two Box score was tested—the combined percentages from the top two points on the scale (“definitely would buy” plus “probably would buy”). Within the group of men, this number could have turned out to be anything. It just happened to be 13 per- cent. Within the group of women, it could have been any- thing, and, as it turns out, was 40 percent. Within each

group, the number was free to be anything from 0 percent to 100 percent, so picking this percentage to test follows the statistical rule. The stat test indicates that the idea that these percentages are from the same place (or are the same) can be rejected, so we can say they are “statistically significantly different at the 95 percent confidence level.”

Something different often happens in practice, though. Because the computer programs that generate survey data do not “know” what summary measure will be important, these programs test everything. When looking at computer- generated data tables, the statistical results will look some- thing like those shown in Table B.

If the Top Two Box score is selected ahead of time, and that is all that is examined (as in Table A), then this auto- matic testing is very helpful. It does the work, and shows that 13 percent differs from 40 percent. The other stat test results are ignored. However, if the data are reported as shown in Table B, there is a problem.

The percentages for the men add to 100 percent. If one percentage is picked for testing, it is “taken out” of the scale, in a sense. The other percentages are no lon- ger free to be whatever they might be. They must add to 100 percent minus the set, fixed percent that was selected for testing. Percentages for the men can vary from 0 percent to 87 percent, but they can’t be higher, because 13 percent is “used up.” Similarly, percentages for the women can vary from 0 percent to 60 percent, but 40 percent is used already. When you look at testing in the other rows, or row by row, you are no longer using the confidence level you think you are using—it becomes something else.

Statistically, if one said of Table B that the percentages that “definitely would buy” and the percentages that “definitely/probably would buy” both differ at the 95 per- cent confidence level, it would be wrong. One of them does, but the other difference is at some unknown level of

significance, probably much less than 95 percent, given one related significant difference.

Table B
Purchase Intent Among Men and Women

Base: total per group	Men (200) %	Women (205) %

Definitely would buy
Probably would buy
Def./Prob. would buy
Might or might not buy
Probably would not buy
Definitely would not buy
Total 100

100

S = the percentages differ significantly at the 95% confidence level. D = the percentages differ directionally at the 90% confidence level.

Stat tests are very useful. Each one answers a specific question about a numerical relationship. The one most commonly asked about scale responses is whether two numbers differ significantly. If they are the right two num- bers, and the proper test is used, the question is easily answered. If they are the wrong two numbers, or the wrong test has been used, the decision maker can be misled.

Questions

1. How is it that statistical testing is the reverse of everyday logic? Explain.

2. What is the most common question we address with sta- tistical tests about scale responses?

3 S 21 10 19 13 S 40 40 35 30 20 17 D5

Hypothesis Testing

hypothesis

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

A **hypothesis** is an assumption or theory 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.

■ The results of a tracking survey show that awareness of a product is lower than it was in a similar survey conducted six 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 different enough 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 cal- culate the probability of observing a particular result if the stated hypothesis is actually true.[3]

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 dif- ference 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 seem- ingly 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 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 (some- times 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*). 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 ≠ 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 Statistical Test 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 statisti- cal tests, along with the situations where they are appropriate, are discussed in this chapter. Exhibit 16.1 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. The following Practicing Market- ing Research feature further addresses this issue.

null hypothesis

The hypothesis of status quo, no difference, no effect.

		Statistical Tests and Their Uses			
Area of Application	**Subgroups or Samples**	**Level Scaling**	**Test**	**Special Requirements**	**Example**

Hypotheses about frequency distribution

Hypotheses about means

One

Two or more

One (large sample)

One (small sample) Two (large sample)

Three or more

One (large sample)

Two (large sample)

Nominal

Nominal

Metric (interval or ratio)

Metric (interval or ratio) Metric (interval or ratio)

Metric (interval or ratio)

Metric (interval or ratio)

Metric (interval or ratio)

χ^2 χ^2

Z test for one mean

t test for one mean *Z* test for one mean

One-way ANOVA

Z test for one proportion

Z test for two proportions

Random sample

Random sample, independent samples

Random sample, $n \geq 30$

Random sample, $n < 30$ Random sample, $n \geq 30$

Random sample

Random sample, $n \geq 30$

Random sample, $n \geq 30$

Are observed differences in the number
of people responding to three different promotions likely/not likely due to chance?

Are differences in the number of men and women responding to a promotion likely/ not likely due to chance?

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?

Same as for small sample above
Is the observed difference between the means for two subgroups (mean income for men and women) likely/ not likely due to chance?

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?

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?

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?

Hypotheses about proportions

Choosing the Right Test for the Right Situation[4]

How's a researcher to know which kind of statistical testing procedure software to use to generate marketing research data tables? Three basic rules can help marketers deter- mine which testing procedure should be employed in a given situation:

- There is a difference between "two groups" and "three or more groups."

- There is a difference between percentages and means.

- There is a difference between matched samples and

When testing percentages with dependent groups, chi- squares should be used for three or more groups, and *Z* tests should be used for two groups. When testing means, ANOVAs (Analysis of Variance) are used in the case of three or more group, and *t* tests are used for the two-group case. (See the accompanying table.)

PRACTICING MARKETING RESEARCH

Which Test Should I Use?		
When measuring. . .	**2 Groups**	**3+ Groups**

independent samples.

Percentages Means

Z test *t* test

chi-square ANOVA

Hypothesis Testing 401

Although analytic software can be very useful for quickly processing survey data, researchers must be careful how they use it. Default settings in many analytic software packages will run the same test automatically, frequently violating the first rule above when testing three or more groups. Thus, in the case of percentages, many software tools will often apply multiple *Z* tests instead of a chi-square and, in the case of means, multiple *t* tests instead of an ANOVA.

Statistically, using the wrong test in these instances will result in a confidence level below what you think you are using. For example, when comparing the percentages in three different groups, the chi-square method will determine whether they vary statistically by running a single test comparing percentages for all three groups simultaneously. In contrast, the *Z* test will run three separate tests, comparing group A to group B, B to C, and C to A individually. Ultimately, it is possible that the tests could produce different outcomes, where one test would suggest a statistically significant difference exists across the three groups while others would suggest no statistically significant differences or only find differences between one or two groups.

Essentially, the chi-square test is taking all the information into account at once, whereas the *Z* test is running multiple tests on selected portions of the data. In most cases, the results of using the wrong test will probably not lead marketing managers to any critical mistakes based on the research; however, using the wrong testing method can seriously undermine credibility. As a matter of quality control, this error can be easily avoided if researchers are careful to use the right tests and settings when performing their analysis.

Questions

1. Aside from the automatic settings in analytic software mentioned above, can you think of any other procedural factors that might cause a researcher to misapply certain tests?

2. If you have access to analytic software, run a *Z* test and a chi-square test on the same set of data (you might try pulling one from earlier in the chapter) and compare the results. Were they significantly different? If they did not produce the same results, can you see why?

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 (e.g., 5 times out of 100) 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 deci- sion 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 prob- ability 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 (e.g., the differ- ence 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 alter- native hypothesis, but statistically, we can only state that the null hypothesis is not true.

StepFour:CalculatingtheValueoftheTestStatistic Inthisstep,theresearcher 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 hypo- thesis H_0.

decision rule

Rule or standard used to determine whether to reject or fail to reject the null hypothesis.

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.

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.

Types of Errors in Hypothesis Testing

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 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 mak- ing a correct decision in rejecting the null hypothesis when, in fact, it is false. The four pos- sibilities are summarized in Exhibit 16.2.

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 com- plicated 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 α = .05. As a result, the pro- cedure 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

H_0 is true Correct ($1 - \alpha$) Type I error (α) H_0 is false Type II error (β) Correct ($1 - \beta$)

EXHIBIT 16.2	Type I and Type II Errors	
Actual State of the Null Hypothesis	Fail to Reject H_0	Reject H_0

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 con- dition 100 percent of the time. Here are the hypotheses to be tested:

Thus,

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

$\alpha = P$ (rejecting H_0 when it is true)
= (test indicates that you have the disease when

you do not have it)

$\beta = P$ (FTR H_0 when in fact it is false)
= P (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 for a given sample size. 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. You may also increase the sample size in order to reduce β for a given level of α.

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, sup- pose you are testing commercials that were very expensive to produce, and you are con- cerned 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 as 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 speci- fications 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. On the one hand, a fuse must trip, or break contact, when it reaches a preset tem- perature, 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.

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 house- holds 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 m: 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.

More Tips on Significance Testing

Paul Schmiege, Analytical Science, DSS Research

On the quantitative side of the marketing research industry, you spend a great deal of time measuring and comparing. But unlike grabbing a yardstick and measuring a physical quantity, your measurements have sampling error. (*Sampling error* is a technical term and does not connote a mistake.) Then, when you compare two

measurements, both of which have sampling error, you can- not be 100 percent confident that a difference even exists. That's where statistics comes in and says, "Though you can't be 100 percent confident, you can test to see if we are 95 percent or 90 percent confident that a difference exists."

The point of making comparisons is to help lead you as interpreters of data to evaluations. "Is this difference impor- tant, is it something that we should act on, should we con- tinue with the same advertising strategy to increase unaided awareness?" Unfortunately, statistics cannot answer those questions for you.

The distinction between testing a difference with a statisti- cal test and evaluating the meaning or relevance of said dif- ference is important to remember. The singlemost common test in marketing research is the two-sample *t* test. You use it to answer questions 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 is 47.5 percent. Is this difference significant?" This type of question is so common 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-tail test. It asks, “Does a significant difference exist?” It does not ask, “Is the first significantly greater than the second?” or “Is the first sig- nificantly less than the second?” Consequently, if a sig- nificant 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

PRACTICING MARKETING RESEARCH

© Roger Gates

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, always remember that the technical term *sta- tistical significance* is not the same as the more intuitive terms like *practical significance* or *importance*. At the heart of its technical meaning, significance in the field of statistics means the difference is likely greater than we would expect due to sampling error. Don’t mistake a statistical test, a use-fultoolintheevaluation,fortheevaluationitself.

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

The following hypotheses are produced:

The decision makers at LNA are willing to use a test that will reject H_0 when it is cor- rect only 5 percent of the time (α = .05). This is the significance level of the test. LNA will

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, the managers of the meat company that supplied the chicken breasts would probably do a two-tailed test.

$X = \$54,323\ S = \$4,323\ n=250$

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

© Jack Puccio/iStockphoto

reject H_0 if the difference between the sample mean and the Census estimate is larger than can be explained by sampling error at α = .05

Standardizing the data so that the result can be directly related to *Z* values in Exhibit 2 in Appendix 3, we have the following criterion:

Reject H_0 if is larger than can be explained by sampling error at $\alpha = .05$. This expres- sion can be rewritten as

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 a mean equal to 0 and a standard deviation equal to 1.

That is, if H_0 is true, $(X - \$55{,}347/SI\ n)$ 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.

WewillrejectH$_0$ if|Z|>*k*.When|Z|>*k*,either*Z*>*k*or*Z*<–*k*,asshownin Exhibit 16.3. Given that

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 3, we find that *k* = 1.96. Therefore, the test is

and FTR H_0 otherwise. In other words,

$\sqrt{\ }$

$X - \$55{,}347 \ {}_{>k}\ S/n$

$Mean = \mu = \$55{,}347\ Standarddeviation = \frac{S}{n}$

$P(Z > k) = .05$

Reject H_0 if $X - \$55{,}347 > 1.96\ S/n$

Area= α – .05
= .5 – .025 = .475

.025 .025

Exhibit 16.3

Shaded Area Is Significance Level α

$-kkZ\ Z>k$

$\sqrt{\ }$

Reject H_0 if $X - \$55{,}347 > 1.96$ or if $X - \$55{,}347 < 1.96\ S/n\ S/n$

The question is whether $54,323 is far enough away from $55,347 for LNA to reject H_0? The results show that

Because –3.75 < –1.96, we reject H_0. On the basis of the sample results and α = .05, the conclusion is that the average household income in the market is not equal to $55,347. If H_0 is true (μ = $55,347), then the value of X obtained from the sample ($54,323) is 3.75 standard deviations to the left of the mean on the normal curve for X. A value of 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.

Having said all this, we caution against an overreliance on statistical tests and signifi- cance/nonsignificance. The Practicing Marketing Research feature below covers this point in greater detail.

$Z = X - \$55{,}347\ S/n$

$= \$54{,}323 - \$55{,}347 = -3.75\ \$4{,}322\ /\ 250$

Does Statistical Precision Validate Results?[5]

The presence of statistical significance in marketing research analysis can be deceptive. *Statistical* significance does not necessarily mean that the difference has any *practical* significance. In addition to large sample sizes, there are many poten- tial sources of error that can create problems for researchers when identifying statistically significant differences.

Generally, two kinds of error affect the validity of statisti- cal measurements. *Random error* introduces error variance, but as it occurs randomly across respondents, it does not add statistical bias to the data. *Systematic error* is consistent across respondents, creating a bias within the data that may or may not be known. Typically, the causes for these kinds of error fall into two categories: sampling error, arising in the process of building a respondent pool; and measurement error, arising from the way the questionnaire is constructed.

Sampling Error

Three major sources of sampling error include under cover- age, nonresponse, and self-selection.

1. *Under coverage*—Under coverage occurs when a certain segment of the population is not adequately represented.

2. *Nonresponse*—Nonresponse error is a result of portions of the population being unwilling to participate in research projects.

3. *Self-selection*—Self-selection can result from respond- ents having control over survey completion. For instance, participants in an online survey panel might get bored and opt out before the survey is over.

Measurement Error

The following six types of measurement error can result in random or systematic error.

1. *Question interpretation*—Respondents may interpret vague or ambiguously worded questions differently.

2. *Respondent assumptions*—Regardless of the way a question is worded, respondents still bring personal assumptions to the table, including any number of varying external factors influencing their understanding of the question.

3. *Question order*—Respondents might answer a question differently depending on where it falls in the survey, as their opinions might be influenced by their thoughts on surrounding questions.

PRACTICING MARKETING RESEARCH

4. *Method variance*—Researchers must be aware of potential errors introduced by the method used to deliver the survey.

5. *Attribute wording*—The way in which survey attributes are described may elicit different answers from respondents.

6. *Omitting important questions*—Systematic error most commonly results from inadequate coverage of critical variables within the question battery. Absent variables can significantly affect the results of data analysis.

Managerial Significance

Rather than focusing on statistical significance in and of itself, researchers need to identify results that have managerial significance—results that are relevant to the decision-making process. Given a large enough sample, any null hypothesis can be discounted, and any two unequal means can be shown to be statistically different. An absence of statistical significance between two supposedly "different" populations may be just as relevant as any demonstrated statistical significance. As such, statistical testing should be used as a tool to discover practical insights, not to define them.

Questions

1. Of the potential causes for error described above, which do you think would be easiest to identify? Hardest? Explain your reasoning.

2. Can you think of any ways that could help researchers determine whether occurrences of statistical significance in their results have managerial significance?

Researchers must be careful to differentiate between resul- tant random and systematic errors. Furthermore, they must also realize that statistical precision does not necessarily indicate that the difference is actionable or meaningful.

Commonly Used Statistical Hypothesis Tests

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.

A number of commonly used statistical hypothesis tests of differences are presented in the following sections. Many other statistical tests have been developed, 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 3.

Independent versus Related Samples

In some cases, one needs to test the hypothesis that the value of a variable in one popu- lation 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.

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. By contrast, 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 statis- tic. The number of **degrees of freedom** is the number of observations in a statistical prob- lem 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 free- dom. 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.[6] The purpose of a cross tabulation is to study rela- tionships 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.[7] 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. A case where chi-square is used is provided in the Practicing Marketing Research feature below.

degrees of freedom

Number of observations in a statistical problem that are free to vary.

chi-square test

Test of the goodness of fit between the observed distribution and the expected distribution of a variable.

Study Results Using Chi-Square Guide Improvements-Myrtle Beach Golf Passport Program[8]

Since the opening of America's first golf course in Charles- ton in 1786, golf has played a big role in the economy of South Carolina. Economic activity from visiting golfers on

and off golf courses in South Carolina created a $2.72 bil- lion economic impact in 2007.

The unique importance of golf and tourism to the area created an opportunity to study how a local affinity market- ing program, Myrtle Beach Golf Passport, affects the large number of visitors to the area, as well as those golfers who live in Myrtle Beach.

PRACTICING MARKETING RESEARCH

Myrtle Beach Golf Passport was created in 1993, enabling eligible residents, as well as those who own sec- ond homes in the area, the opportunity to enjoy reduced golf fees all year round. This program has been favorably received by over 10,000 members and has enjoyed a 75 percent annual renewal.

Build on that success

Looking to build on that success, a marketing research study was undertaken to determine if Passport should be expanded from simply reducing greens fees to include other areas of golf vacation activities such as attractions, restaurants and retail locations.

The Passport group agreed to cooperate in the market- ing research effort and helped to generate lists of attrac- tions, restaurants, and retail shopping locations that might become part of the Passport affinity marketing program, as detailed below.

Attractions. Ten participating attractions cover events for adults and children and represent the main attractions in the Myrtle Beach area.

Restaurant types. Participating restaurants represent a cross-section of restaurants available in the Myrtle Beach area.

Retail shopping locations. The 10 retail shopping loca- tions represent both golf specialty retail outlets and gen- eral-merchandise retail locations.

The survey questionnaire was distributed e-mail using several different lists. In addition to a variety of demo- graphic items, the survey participants were asked if they were an occasional visitor, seasonal visitor, part-time resi- dent, or full-time resident in the area. They were then grouped into visitor and resident segments. For each of the attractions, restaurant types and retail shopping locations, the participants indicated whether they never, rarely, some- times or always visited each of the attractions, restaurant types, and retail shopping locations.

The survey yielded responses from 529 residents and 199 visitors, for a total sample size of 728. These data were then analyzed for differences between the visitor and resi- dent segments.

Low level of willingness. Overall, the attractions showed fewer participants willing to always visit them, ranging from 1.2 percent for Myrtle Waves to 8.5 percent for Carolina Opry. In contrast, the restaurants showed a low of 2.3 per- cent for theme restaurants and a high of 33.7 percent for steakhouses. Retailer scores ranged from a low of 1 percent for Old Golf Shop and a high of 57.2 percent for Martin's PGA Superstore.

Attractions. Significant chi squares were found for the six attractions listed below along with their p values: Alabama Theater ($p < .0001$), Carolina Opry ($p < .0001$), Dixie Stampede ($p < .002$), Legends in Concert

($p < .0001$), Medieval Times ($p < .0001$), and Ripley's Aquarium ($p < .003$).

($p < .0001$), Medieval Times ($p < .0001$), and Ripley's Aquarium ($p < .003$).

In all cases, residents were significantly more willing to visit these attractions compared to visitors. However, the percentages of Passport members who always or some- times visit any of the attractions were low, averaging only 19.8 percent and ranging from a low of 10.7 percent to 39.0 percent. The unwillingness of the majority of visitor Pass- port members to visit attractions sometimes or always makes attractions a low priority for inclusion in a discount program for Passport members who are visitors. The results for residents weren't much better, showing an average of 26.1 percent and ranging from a low of 9.9 percent to 43.9 percent of residents who are sometimes or always visiting attractions.

Restaurant types. Significant chi squares were found for two restaurant types (listed along with their *p* values): Italian restaurants ($p < .002$) and seafood restaurants ($p < .008$).

A majority of visitors who are Passport members either sometimes or always visit restaurants in high percen-tages for the following: steakhouses (79.2), seafood (77.8 percent), Italian (68.0 percent), and sports bars (51.1 percent).

Although residents and visitors showed no significant dif- ference for steakhouses, the combined percentage for all Passport members who say they either sometimes or always visit steakhouse restaurants was 83.6 percent, the highest for restaurants as a group.

Retail shopping locations. Significant chi squares were found for the five retail shopping locations, listed here along with their *p* values: Coastal Grand Mall ($p < .0001$), Colonial Mall ($p < .04$), Golf Dimensions Superstore ($p < .044$), Inlet Square Mall ($p < .0001$), and MacFrugal's Golf (Murrells Inlet) ($p < .034$).

A majority of visitors who are Passport members either sometimes or always visit retail shopping locations in high percentages.

The most frequently visited retail shopping locations are either golf specialty stores or diversified retail centers.

Recommendations

The study results were presented by the research team to the Myrtle Beach Area Golf Course Owners Association at a golf owners' conference. The following recommendations were made:

- The opportunities for offering discounts in the Passport program are, in descending order of potential value: shopping, restaurants, and attractions.

- Since a majority of Passport owners who were either residents or visitors did not indicate they either some- times or always go to any of the attractions, this

category was not recommended by the research team for discount offers. The lack of a broad appeal indicates no interest.

- Discounts for steakhouses and seafood restaurants were highly recommended by the research team for both visi- tor and resident Passport members. This recommenda- tion was based on high percentages of both segments saying they would either sometimes or always visit a steakhouse or seafood restaurant.

Utilization has been high

All of these special discounts have been made available and marketed on the enhanced owners' website, myrtle- beachgolfpassport.com. Utilization has been high on the 81 courses represented on the website, and the program may be expanded to include more restaurants and golf retail outlets if partners can be found.

Questions

1. What did the chi-square statistic tell the researchers in this case?

2. What did they find regarding the potential value of shop- ping, restaurants, and attractions?

Discounts at retail shopping outlets were also recom- mended by the research team, but were confined to golf shops. Merchandise discounts were made available through Passport for members and guests at most pro shops (10 percent), Golf Dimensions (10 percent), and Callaway Per- formance Center (10 percent).

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 2 May 3 June

Total

11,700 12,100 11,780 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 ques- tion. 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 custom-

ers 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 dif- ference 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 differ- ence) hypothesis, the expected number of customers visiting the store in each deal period would be computed as follows:

$$E_i = {}^{\mathrm{TV}} N$$

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

Thus,

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

For this example,

$\chi^2 = {}^{35,580} = 11,860\ 3$

$\chi^{2} = \sum^{k} (O_{i} - E_{i})^{2}{}_{i=1} E_i$

O_i =observednumberin*i*thcategory where E_i = expected number in *i*th category

k = number of categories

For this example,

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

= 7.6

4. **Select the level of significance α. If the .05 level of significance (α) is selected, the tab- ular χ^2 value with 2 degrees of freedom (k – 1) is 5.99. (See Exhibit 4 in Appendix 3 for k – 1 = 2 d.f., α = .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 differ- ent from the others.**

Chi-SquareTestofTwoIndependentSamples Marketingresearchersoftenneed to determine whether there is any association between two or more variables. Before formu- lation 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 categories? Are pur- chasers 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.

. 2 2

. 3 5

. 5 7

. 6 2

. 7 1

. 8 2

. 9 1

. 10 7

12 3 15 5 20 6 23 1 25 1 30 1 40 1

4.4 11.1 15.6 4.4 2.2 4.4 2.2 15.6 6.7 11.1 13.3 2.2 2.2 2.2 2.2

4.4 15.6 31.1 35.6 37.8 42.2 44.4 60.0 66.7 77.8 91.1 93.3 95.6 97.8 100.0

2 5 7.0 7.0 3 4 5.6 12.7 4 7 9.9 22.5 5 10 14.1 36.6 6 6 8.5 45.1 7 3 4.2 49.3 8 6 8.5 57.7 9 2 2.8 60.6

10 13 18.3 78.9 12 4 5.6 84.5 15 3 4.2 88.7 16 2 2.8 91.5 20 4 5.6 97.2 21 1 1.4 98.6 25 1 1.4 100.0

EXHIBIT 16.4 Data for 32 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

n_m=45

Mean number of visits by males, $X_m = \sum X_m f_m$ =11.5

45

Mean number of visits by

n_f= 71

females, $X_f = \sum X_f f_f$ =8.5 71

The technique will be illustrated using the data from Exhibit 16.4. 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 16 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 fre-

quency 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			
6–14 visits			

15 and above visits Totals

14 26 40 16 34 50 15 11 26 45 71 116

surveysolutions XP

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 pro- vided below along with the output produced. Use the data set **Chisqex**, which you can down- load from the website for the text.

Steps in SPSS

1. **Select Analyze → Descriptive Statistics → Crosstabs.**

2. **Move bin to Rows.**

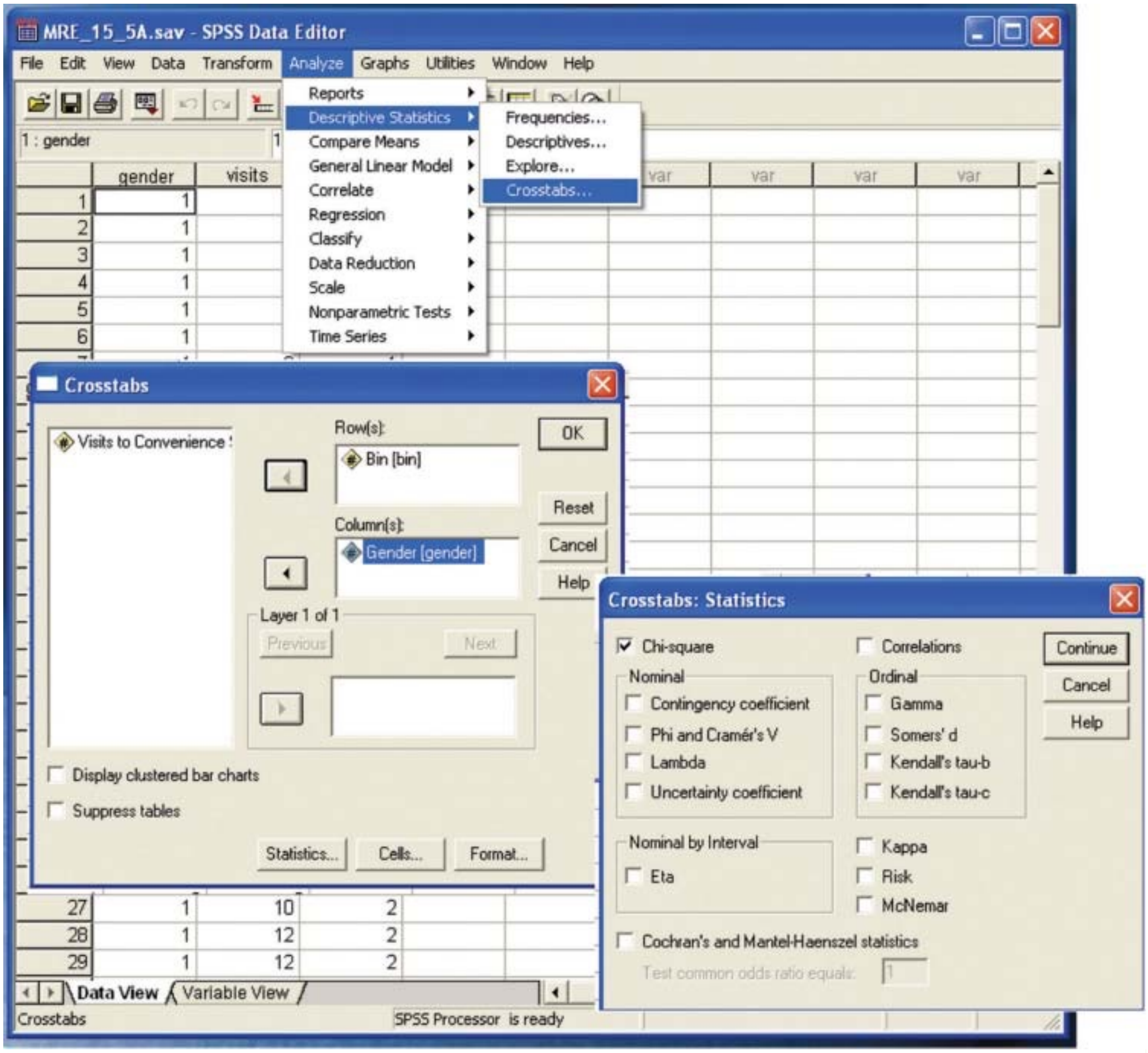
3. **Move gender to Columns.**

4. **Click Statistics.**

5. **Check box for Chi-square.**

6. **Click Continue.**

7. Click OK.



Row(s): Bin [bin]

Column(s): Gender [gender]

□ Display clustered bar charts

□ Suppress tables

☑ Chi-square

□ Correlations

Nominal

□ Contingency coefficient

□ Phi and Cramér's V

□ Lambda

□ Uncertainty coefficient

Ordinal

□ Gamma

□ Somers' d

□ Kendall's tau-b

□ Kendall's tau-c

Nominal by Interval

□ Eta

□ Kappa

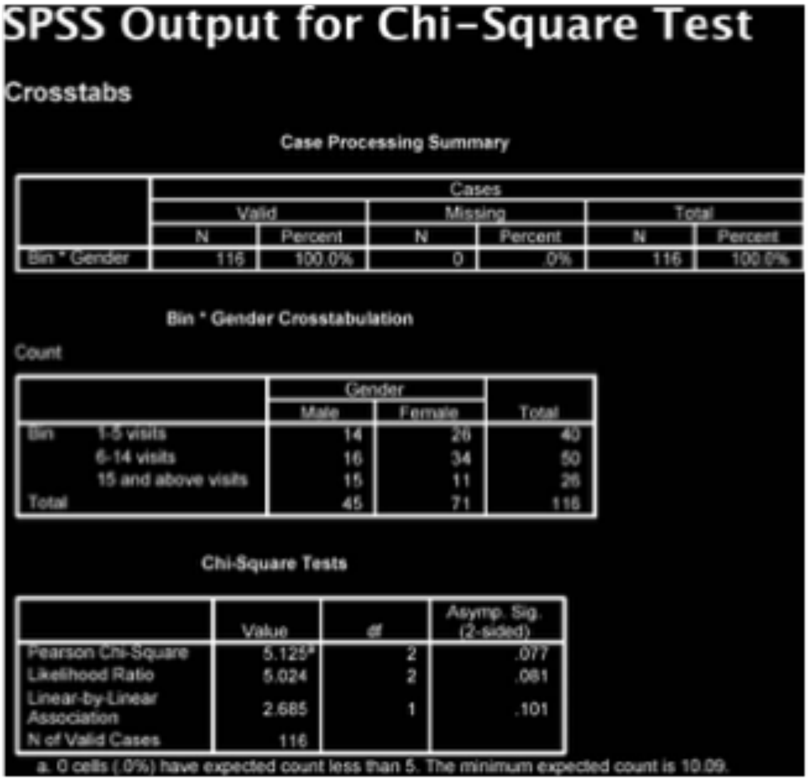
□ Risk

□ McNemar

□ Cochran's and Mantel-Haenszel statistics

	gender	visits	
1	1		
2	1		
3	1		
4	1		
5	1		
6	1		
27	1	10	2
28	1	12	2
29	1	12	2

Spss Jump Start For Chi-Square Test 415



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 6-14visits
15 and above visits

45×40 = 15.5 71×40 = 24. 116 116

45×50 =19.4 71×50 =30.

116

45×26 = 10.1 116

116

71×26 = 15. 116

The χ^2 value will be distorted if more than 20 percent of the cells have an expected fre- quency 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

$^{r\,k}(O-E)^2\,\chi^2=\sum\sum\quad ^{ij\;ij}$

$_{i=1\;j=1}\,E_{ij}$

where O$_{ij}$ 5 observed number in the *i*th row of the jth column

E_{ij} 5expectednumberinthe*i*throwofthejthcolumn

For this example,

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 Exhibit 4 of Appendix 3). 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.

Hypotheses about One Mean

$_2\,(14-15.52)^2$

$(26-24.48)^2\,(16-19.4)^2\;^{+}\;24.28\;^{+}\;19.4$

$(15-10.09)^2\,(11-15.91)^2\;^{+}\;10.09\;^{+}\;15.91$

$\chi=+$

$15.52\;(34-30.6)^2\;30.6$

$=5.1$

Z test

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

Z Test

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 ($n \geq 30$) the t test with $n - 1$ degrees of freedom (where n = sample size) should be used.

Mobile Connection, a Dallas mobile phone and accessories chain, recently completed a survey of 200 consumers in its market area. One of the questions was “Compared to other mobile phone stores in the area, would you say Mobile 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:

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

The mean rating of Mobile Connection is 3.4. The sample standard deviation is 1.9. How can the management of Mobile Connection be confident that its 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) ≤ 3 ■ Alternative hypothesis H_a : $M \leq 3$

2. Specify the level of sampling error (α) allowed. For α = .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

Response Code

Hypotheses about One Mean 417 to mean that the chance of being wrong because of sampling error should be no more than .05 (an α = .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**
   In this case,

5. **Calculate the test statistic:**

$$S_X = \frac{S}{n}$$

where S_X = estimated standard error of the mean

$$S_X = \frac{S}{n}$$

where S_X = estimated standard error of the mean

$S_X =$ $^{1.9}$ =0.13 200

(Population mean specified\ (Sample mean)–| |

Z=

= $^{3.4 - 3}$ = 3.07 0.13

| under the null hypothesis \/

Estimated standard error of the mean

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 cor- rect 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 gen- erally used for large samples, nearly all statistical packages use the t test for all sample sizes.

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 in the accompanying table.

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

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

1. Specify the null and alternative hypotheses.

- Null hypothesis $H_a : X \geq 1,000$ cases per store per week
- (X = average sales per store per week).
- Alternative hypothesis H_a: $X \geq 1{,}000$ cases per store per week.

Store	Average Sales per Week (X_i)

1 2 3 4 5 6 7 8 9

10 11 12

870

910 1,050 1,200 860 1,400 1,305 890 1,250 1,100 950 1,260

$_n\sum X_i$

Mean sales per week, $X = {}^{i=1} = 1087.1\ n$

2. Specify the level of sampling error (α) allowed. For α = .05, the table value of *t*(critical) = 1.796. (See Exhibit 3 in Appendix 3 for 12 – 1 = 11 d.f., α = .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:

where

X_i

$X\ n$

= observed sales per week in *i* th store = average sales per week
= number of stores

$S=$

n

$\sum (X_i - X)^2$

$i=1$
$n-1$

For the sample data,

Further discussion of the *t* test is provided in the SPSS feature on page 419.

$S = {}^{403,822.9} = 191.6 \; {}^{(\,1\;2\,-\,1)}$

4. Calculate the estimated standard error of the mean (SX), using the following formula:

$S_X = S$

n

$= {}^{191.6} = 55.3\ 12$

SPSS JUMP START FOR t TEST

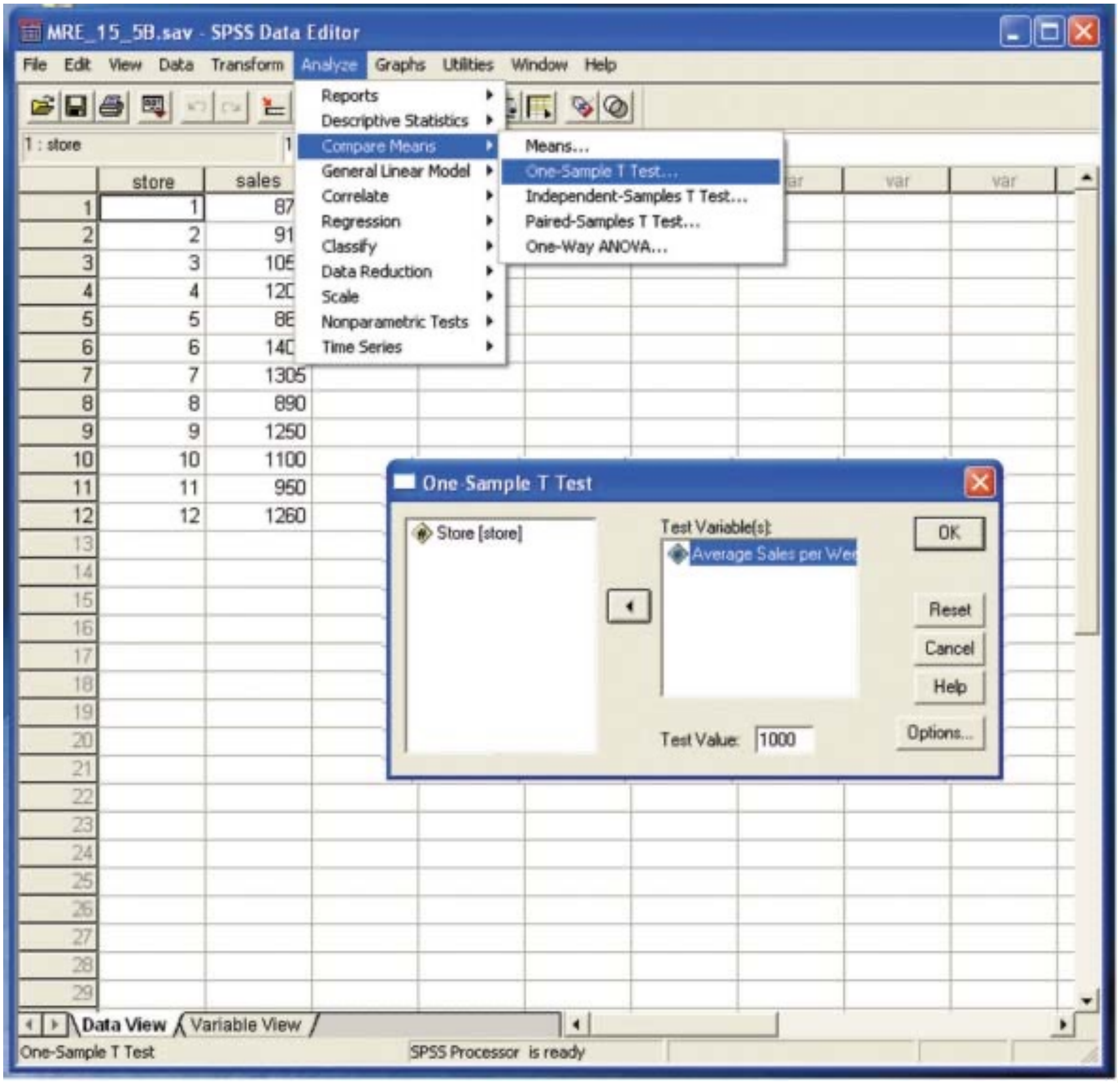
Steps that you need to go through to do the *t*-test problem shown in the book are provided below along with the output produced. Use the data set ***TTestex***, which you can download from the website for the text.

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.



1 : store

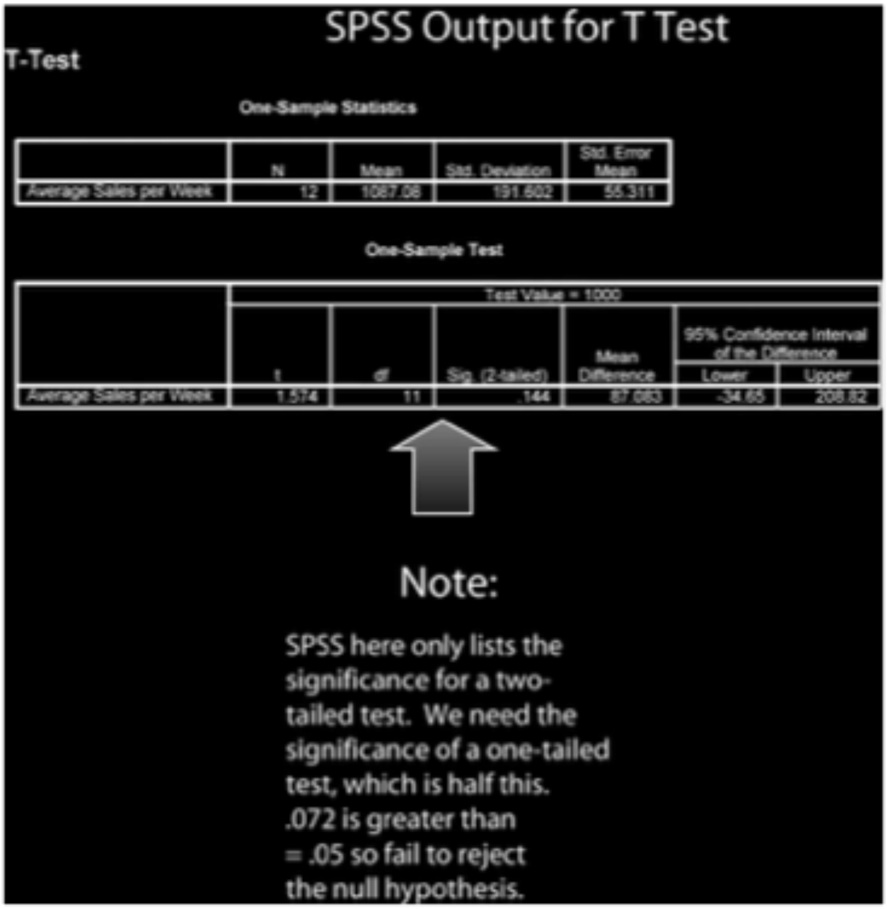
	store	sales
1	1	87
2	2	91
3	3	105
4	4	120
5	5	88
6	6	140
7	7	1305
8	8	890
9	9	1250
10	10	1100
11	11	950
12	12	1260

One-Sample T Test

Store [store]

Test Variable(s): Average Sales per We

Test Value: 1000



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 is greater than = .05 so fail to reject the null hypothesis.

5. Calculate the *t*-test statistic:

Z=

|under the null hypothesis \\/

/Population mean \\ (Sample mean)–| |

Estimated standard error of the mean $_=$ 1,087.1−1000 $_=$ 1.6

55.3

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 (*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 16.4) 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 3 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 of the differences between the two means as follows:**

Therefore,

Note that this formula is for those cases in which the two samples have unequal vari- ances. 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:

S^2 S^2 $S_{X_{m-f}} = {}^{m} + {}^{f}$

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*

$n_m\, n_f$

$(8.16)^2\,(5.23)^2$

$S_{X_{m-f}} = {}_{45} + {}_{71} = 1.37$

⎛Difference between means⎞ ⎛Difference between means⎞

| | − | |

| | under the null hypothesis

$Z =$ ⎝of first and second sample ⎠ ⎝ ⎠

Standard error of the differences between the two means = (11.49−8.51)−0

=2.18

1.37

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

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

In many situations, researchers are concerned with phenomena that are expressed in terms of percentages.[9] 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 $70,000 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 custom- ers surveyed reported family incomes of $70,000 or more per year. The procedure for the **hypothesis test of proportions** follows:

Comstock Images /Getty Images

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 $70,000 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 3 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:
Therefore,

4. **Calculate the test statistic as follows:**
   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 $70,000 or more. Management can intro- duce 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 character- istic. 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.

$S_P = P(1-P)\ n-1$

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

$S_P = .6(1-.6) = .022\ 35-1$

Z = (Observed proportion − Proportion under null hypothesis) Estimated standard error (S_P)

$= 0.743-0.60 = 6.5\ .022$

The sample proportions and the difference can be calculated from Exhibit 16.4 as follows:

2. **Set the level of sampling error α at .10 (management decision). For α = .10, the table value of Z(critical) = 1.28. (See Exhibit 3 in Appendix 3 for d.f. = ∞,.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:**

$P_m = 26 = .58\ 45$

P =30

f

= .42
71

$P - P = .58 - .42 = .16$ $_{mf}$

$S_{P_{m-f}} = P\,(\,1 - P\,)\,|\,|\,|\, + \backslash n_m$

$n\ P + nP$ *where* $P = ^{m\,m\,f\,f}$

P = proportion in sample f (women) $_f$

n_m = size of sample m n_f = size of sample f

$(11\backslash$

$|\,|\,|\,|\,n_f)$

$n_m + n_f$

P_m = proportion in sample m (men)

Therefore,

4. **Calculate the test statistic.**

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 α = .10). Management can conclude with 90 percent confidence (1 – α = .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).

$P = 45(.58)+71(.41)_{=.42}$ 45+71

and $S_{P_{m-f}} = (\ 1\ 1 \backslash\ .48(1-.48)| + | = .1$

$\backslash 45\ 71)$

(Difference between \ (Difference between proportions\

| |_| |

|observed proportions |under the null hypothesis $_{Z=}\backslash\ )\backslash\ )$

Estimated standard error of the differences between the two means $_=$

$(.58-.42)-0$ $_{=1.60}$

.10

Analysis of Variance (ANOVA)

When the goal is to test the differences among the means of two or more independent sam- ples, **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 $C > 3$). 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 situa- tions 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 associ- ated 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 I error. Because these tests must be per- formed 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 sam- pling error. At an α of .05, this could be expected to occur in 1 of 20 tests on average.

One-way ANOVA is often used to analyze experimental results. Suppose the marketing manager for a chain of brake shops was considering three dif- ferent 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 as follows. The question is, are the differences among the means larger than would be expected due to chance?

analysis of variance (ANOVA)

Test for the differences among the means of two or more independent samples.

310 318 315 322 305 333 310 315 315 385 345 310 340 312 330 308 320 312 315 340

314
315
350
305
299
309
299
312
331
335
321 337 340 325 318 330 315 345 322 320 295 325 302 328 316 330 294 342 308 330

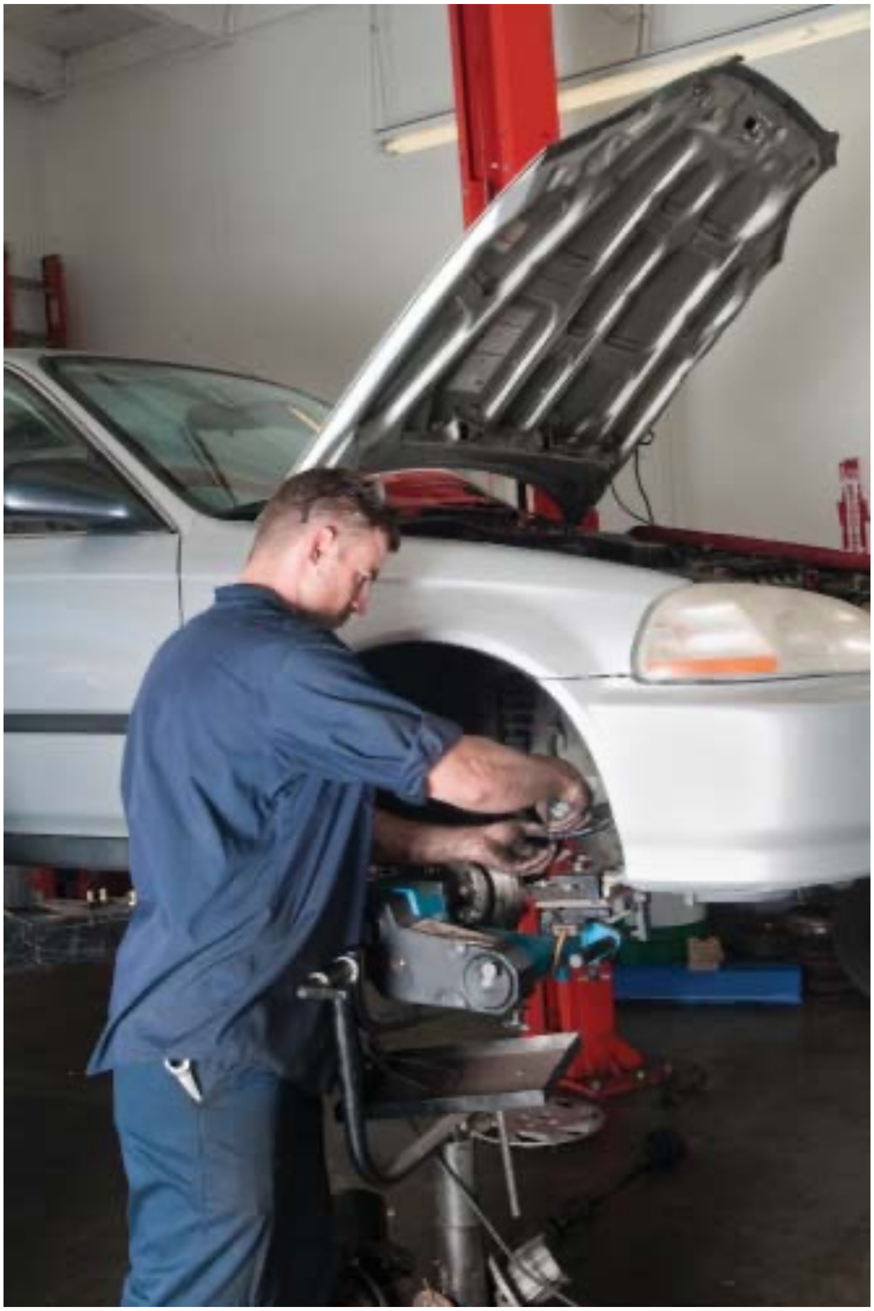
310
312
340
318
322
335
341
340
320
310

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.

$X = 323$

$X = 315$

$X = 328$



© Eliza Snow/iStockphoto

Chicago (Wheel Alignment)	Cleveland (Oil Change)	Detroit (Tune-Up)

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 and the overall sample mean 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:

In this example, the overall sample mean is

Thus,

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:

In this example,

Thus,

4. Sum the squared differences between each observation (*Xij*) and its associated sample mean 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:

$$SSA=\sum_{j=1}^{c} n_j(X_j - X_t)^2$$

$X_t = 20(323)+20(315)+20(328) = 322$ 60

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

MSA = Sum of squares among groups (SSA) Degrees of freedom (d.f.)

where Degrees of freedom = number of groups (C) − 1

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

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

$_{C}\,_{n_j}\text{SSE} = \sum \sum (X_{ij} - X_j)^2$

$_{j=1}\ _{i=1}$

$= (6644) + (4318) + (2270) = 13{,}232$

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:

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

Thus,

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 calcu- lated 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.

6. **Calculate the *F* statistic as follows:**
   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 Table 5 in Ap- pendix 3 for d.f. for denominator = 57, 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.**

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

MSE = Sum of squares within groups (SSE) Degrees of freedom (d.f.)

$(K\backslash$

$| \; | \text{d.f.} = | \sum n_j | - C$

$\backslash_{j=1}) = (20+20+20) - 3 = 57$

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

$F = {}^{\text{MSA MSE}}$

$= {}^{860} = 3.70\ 232.14$

The results of an ANOVA generally are displayed as follows:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F* Statistic

Treatments Error
Total

1,720 (SSA) 13,232 (SSE) 14,592 (SST)

2 (C – 1) 860 (MSA) 3.70 calculated 57 (n – C) 232.14 (MSE)
59 (n – 1)

An example of an ANOVA application is provided in the Practicing Marketing Research feature below.

12 Coins, 3 Shops, 2 Microbiologists, and 1 Test—ANOVA[10]

ANOVA, or the analysis of variance, is a statistical test devised in the 1920s by the English statistician, Ronald A. Fisher. British microbiologists Richard Armstrong and Anthony Hilton, both at Aston University in Birmingham, England, have found it to be the "most appropriate method" for statistical analysis of complex data sets—in this case coins collected from a butcher's shop, news agent's office, and a sandwich shop.

The researchers took four coins randomly from each premise and analyzed them for bacterial populations. Arm- strong and Hilton described their procedure as a one-way ANOVA with four replications performed in a randomized test design. The procedure took into account the variation between the various observations and partitioned these variations into portions correlated with the differences they found between the three shops and even the variation among the four coins collected in each shop. Even so, this was a single-factor ANOVA experiment, the only variable being the shop.

They next performed a factorial experiment, studying several factors at once, namely, the influence of the type of dishcloth (cloth or sponge) and the prior rinsing of the material on the quantity of bacteria transferred to a food preparation surface as well as interactions between the two

variables. Then they performed an even more factorially complex ANOVA study to determine how well two strains of bacteria survived on inoculated pound notes measured at 10 time intervals. The ANOVA method enabled them to discern a subtle pattern of three-factor interactions among the variables (such as the slight interaction between surface type and bacterial strain), the decline of bacterial numbers over time in variance with the type of surface, and the decline of one bacterial strain in numbers faster than the other in the same circumstance.

Each of these investigations yielded a data-rich ANOVA table, information that health inspectors surely would find practical and immediately useful. The researchers had high praise for the ANOVA technique, calling it a "powerful method of investigation" for applied microbiology because it could highlight the effect of single factors as well as their interaction. Even better, combining different factors in one study is efficient and often reduces the number of replications needed.

Questions

1. Can you devise a four-factor ANOVA test for the microbiologists?

2. The microbiologists applied their different ANOVA tests to food-service-related shops. How might it be applied to a dentist's office?

PRACTICING MARKETING RESEARCH

P Values and Significance Testing

For the various tests discussed in this chapter, a standard—a level of significance and associ- ated critical value of the statistics—is established, and then the value of the statistic is calcu- lated 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.

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

Mean Mean
Variable G_1:1 G_2:0 t value

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, com- monly 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) sig- nificance 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
- ≤PROB ■ PROB =

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

An example of computer output showing a *p* value calculation appears in Exhibit 16.5. This analysis shows the results of a *t* test of the differences between means for two inde- pendent 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. Sub- jects 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 aver- age 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.

```
df P
ADDED PAY 16.82292 20.04717 -1.32878 200 .185434 96

Valid N G_1:1

Valid N G_2:0
```

EXHIBIT 16.5 **Sample *t*-Test Output**

***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.

SUMMARY

The purpose of making statistical inferences is to general- ize 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 popula- tion 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 par- ticular 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 hypoth- esis 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 tabula- tions, whose purpose usually is to uncover interrelation- ships 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 distri- bution 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

statistical significance **396** hypothesis **398**
null hypothesis **399** decision rule **401**

type I error (α error) **402** type II error (β error) **402**

independent samples **408** related samples **408** degrees of freedom **409** chi-square test **409**

Z test **416** *t* test **417**

hypothesis test of proportions **422** analysis of variance (ANOVA) **425** *F* test **427**
p value **429**

QUESTIONS FOR REVIEW & CRITICAL THINKING

1. **Explain the notions of mathematical differences, manageri- ally 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 per- centages 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.

. 1 Four-foot metal stepladder 425

. 2 $50 savings bond 610

9.

American Airlines is trying to determine which bag- gage 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 approxi- mately the same quantity of baggage each month. American has decided to select the system that provides the minimum number of instances in which passen- gers 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?

4 Six pink flamingos plus an outdoor 705 thermometer

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

7. A market researcher has completed a study of pain reliev- ers. 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.

10–11 12–13 14–15 16–17 18–19 20–21 22–23 24–25 26–27 28–29 30–31 32–33 34–35 36 or more

4 10 10 8 14 14

4 20 2 12 4 6 2 12

14 4 6 13

10 8 12 6 2 8 2 8 2 2

Key Terms 431

Week	Premium	Total Given Out

3 Dinner for four at a local steakhouse

510

Minutes of Waiting	Jano Systems (Frequency)	Dynamic Enterprises (Frequency)

Pain Relievers	Men	Women

Anacin Bayer Bufferin Cope Empirin Excedrin Excedrin PM Vanquish

40 55 60 28 70 97 14 21 82 107 72 84 15 11 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 *t* and recommend to manage- ment whether the kit should have a male or a female orien- tation.

31 26 67 9 12 38 67 9 41 20 25 16 34 32 73 26

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 two weeks; the units of each dressing sold are shown in the following table. As a researcher, you want to know if the differences among the average daily sales of the dress- ings 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?

Boys	Girls	Boys	Girls

63 16 36
7 45 41 20

81 15 5

Day	Caesar	Ranch-Style	Green Goddess	Russian

1 155 2 157 3 151 4 146 5 181 6 160 7 168 8 157 9 139

10 144 11 158 12 172 13 184 14 161

WORKING THE NET

143
146
141
136
180
152
157

167
159
154
169
183
195
177
149 135 152 136 146 131 141 126 173 115 170 150 174 147 141 130 129 119 167 134 145 144 190 161 178
177 201 151

For a *t* test online calculator, available also with no charge, visit: ***www.graphpad.com/ quickcalcs/ttest1.cfm.*** Educators at Tufts University offer a helpful tutorial on reading the output from a one-way analysis of variance, an ANOVA table. Study this at: ***http:// www. JerryDallal. com/LHSP/aov1out.htm.***

1. Calculating the *p* value and performing a *Z* or *t* test are much easier when done by computers. For a *p* value calcu- lator, usable without fee, visit: ***www.graphpad.com/ quick- calcs/PValue1.cfm.***

For a *Z* test calculator, also free to use, visit: ***www.chang- bioscience.com/stat/ ztest.html***

Analyzing Global Bazaar

Segmentation Results

Nala Chan is an advertising executive with Stewart Bakin Advertising. She is responsible for the Global Bazaar account and has just finished reviewing the results of a recent customer study in the top 40 U.S. markets. The study was conducted by Internet panel in 2011 and included people who shopped at

2.

Global Bazaar in the 30 days prior to the date of the survey. The accompanying table shows selected survey results broken down by market segments (columns) identified in research previously conducted by Global. The first two rows are based on actual sales data; the rest of the table shows results from the most recent survey with statistical testing of differences. The first six rows show key metrics used by Global to guide its marketing strategy. Some of the key metrics come from actual sales data, whereas others come from the recent survey. All results, except those for the first two rows, are for the segments or based on segment column totals.

	Market Segments				
	Single	Single	Married	Married	Married
Variable	18–25	26–40	18–25	26–40	over 40

Percent of current customer base[a] 15% 20% 27% 29% 9% Percent of sales[a] 10% 13% 29% 34% 14%

	Market Segments				
	Single	Single	Married	Married	Married
Variable	18–25	26–40	18–25	26–40	over 40

Top of mind awareness of Global Bazaar[a] Image index (100-point scale)[b]
Likely to shop at Global Bazaar
in next 30 days

Likely to shop at a competitive store in next 30 days
Number of children under 23 years Average number

Income
Median income
% HH income over $75,000 Education
% College degree or more Ethnic makeup
% white
% black
% Hispanic
% Other

34%* 29%* 69%* 70%*

21%* 19%*

38%** 40%**

2.38 2.10

$28,000* $39,500 29% 28%

9%* 29%**

94% 92% 3% 4% 2% 2% 4% 4%

45% 51%** 53%** 85% 92%** 93%**

33% 39%** 42%** 28% 23%* 25%*

0

$44,430 15.3%*

26%**

91% 6% 2% 3%

0 0.29

$56.580** $69,170** 28% 36%

21% 21%

95% 95% 2% 4% 2% 1% 4% 2%

[a] Based on actual customer data. Significance testing not appropriate.
[b] Index developed by Global Bazaar based on multiple measures from survey—higher is better.
*Significantly lower than average for all customers surveyed.
**Significantly higher than average for all customers surveyed.

Questions

1. Which segment provides the largest percentage of sales?
2. In which segment does Global have the highest top of

mind awareness?
3. Which two segments account for over 60 percent of sales?

AT&T Wireless

Marc Mulwray is the new marketing research director for AT&T Wireless. Marc was hired to help AT&T address new challenges from Verizon, Sprint, and T-Mobile in the highly competitive environment for wireless customers. New deals and pricing plans, new claims about network speed and new device offerings seem to emerge daily.

One of the crucial challenges for AT&T is that of cur- rent customer retention, as other players try to pull customers

4. **In what segment does Global perform most poorly? Explain all the dimensions of their poor performance in this segment.**

5. **Based on these results, what advice would you give to Global?**

away with deals, including paying their cost to break their contracts with AT&T. Current customer retention is espe- cially important as growth in the total number of wireless customers has waned given that nearly everyone has a wireless phone or tablet.

Marc has been charged with determining how many cur- rent AT&T customers will switch to another wireless pro- vider over the next six months given the current deals offered by competitors. To address this question, Marc and his team designed and fielded a national survey among its current cus- tomers. The survey covered a number of areas, including cus- tomer demographics, psychographics, wireless usage, and so

on; but the key questions relate to their likelihood to switch to various other carriers in response to the deals, pricing and phones they are offering. Surveys were administered online based on e-mail invitations with links to the survey. The sur- vey was completed by 1,200 customers who gave complete answers to all questions.

Initial results indicate that 14 percent of customers are likely to switch. The margin of error is 2.8 percent, which means that (at the 95 percent confidence level) the actual per- centage of customers switching could be as low as 11.2 percent or as high as 16.8 percent. Given that AT&T has more than 100 million customers, this is a difference in customers lost of 16.8 million minus 11.2 million or 5.6 million. AT&T senior management is concerned about this error range of ±2.8 per- cent, which means that error spans a total of 5.6 percentage points. Further customer retention efforts must be budgeted now, and AT&T senior managers want 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 or why not?

2. Do you think AT&T senior management would find this approach to reducing the error range satisfactory?

3. If 1,000 more respondents were surveyed and 20 percent of them indicated that they would switch, what would the error range become?

surveysolutions XP

SPSS EXERCISES FOR CHAPTER 16

Exercise 1: Analyzing Data Using Cross-Tabulation Analysis

Note: Go to the Wiley website 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 addi-

tion, 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 sta- tistic to assess whether or not the results of the crosstab are statistically significant.

In this exercise, we are assessing whether 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 using only the sample data. Do not consider

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 cate- gory? ________%**

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 inde- pendent samples*.

7. Evaluate the chi-square statistic in each of your crosstab tables. Construct a table to sum- marize 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 vir- tually the same; hence SPSS refers only to a t test*.

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 deter- mine whether or not the sample results can be generalized to the population from which the sample data was drawn. If the *t* test is significant, then we can use the group statistics to deter- mine 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 freshmen, sophomores, juniors, seniors, and graduate students to test for significant differences in the importance placed on sev- eral 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 *fac- tor variable*. Be sure to click the *options* icon and check *descriptives* so that the output will pro- duce the mean responses by student classification for the sample data. As with the *t* test, the ANOVA test produces a table of *descriptives* based on sample data. If our ANOVA test is signifi- cant, 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 classification of students by the importance they place on the following movie theater 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.