CHAPTER 15

# Frequency Distribution, Cross-Tabulation, and Hypothesis Testing

"Frequency distribution and cross-tabulations are basic techniques that provide rich insights into the data and lay the foundation for more advanced analysis."

*Laurie Harrington, marketing manager—customer retention, First Choice Power*

## Objectives

At the end of this chapter, the student should be able to:

1. Describe the significance of preliminary data analysis and the insights that can be obtained from such an analysis.
2. Discuss data analysis associated with frequencies, including measures of location, measures of variability, and measures of shape.
3. Explain data analysis associated with cross-tabulations and the associated statistics: chi-square, phi coefficient, contingency coefficient, Cramer's *V*, and lambda coefficient.
4. Describe data analysis associated with parametric hypothesis testing for one sample, two independent samples, and paired samples.
5. Understand data analysis associated with nonparametric hypothesis testing for one sample, two independent samples, and paired samples.

## Overview

Once the data have been prepared for analysis (Chapter 14), the researcher should conduct some basic analysis. This chapter describes basic data analysis, including frequency distribution, cross-tabulation, and hypothesis testing. First, we describe the frequency distribution and explain how it provides both an indication of the number of out-of-range, missing, or extreme values as well as insights into the central tendency, variability, and shape of the underlying distribution. Next, we introduce hypothesis testing by describing the general procedure. Hypothesis-testing procedures are classified as tests of associations or tests of differences. We consider the use of cross-tabulation for understanding the associations between variables taken two or three at a time. Although the nature of the association can be observed from tables, statistics are available for examining the significance and strength of the association. Finally, we present tests for examining hypotheses related to differences based on one or two samples.

Many commercial marketing research projects do not go beyond basic data analysis. These findings are often displayed using tables and graphs, as discussed further in Chapter 22. Although the findings of basic analysis are valuable in their own right, they also provide guidance for conducting multivariate analysis. The insights gained from the basic analysis are also invaluable in interpreting the results obtained from more sophisticated statistical techniques. To provide the reader with a flavor of these techniques, we illustrate the use of cross-tabulation, chi-square analysis, and hypothesis testing.

### REAL RESEARCH

### *Commercial Battle of the Sexes*

A comparison of television advertising in Australia, Mexico, and the United States focused on the analysis of sex roles in advertising. Results showed differences in the portrayal of the sexes in different countries. Australian advertisements revealed somewhat fewer, and Mexican advertisements slightly more, sex-role differences than U.S. advertisements. Cross-tabulation and chi-square analysis provided the following information for Mexico.

| | *Persons Appearing in the AD (%)* | |
|---|---|---|
| ***Product Advertised Used by*** | ***Women*** | ***Men*** |
| Females | 25.0 | 4.0 |
| Males | 6.8 | 11.8 |
| Either | 68.2 | 84.2 |

$\chi^2 = 19.73$, $p \leq 0.001$

These results indicate that in Mexican commercials, women appeared in commercials for products used by women or by either sex but rarely in commercials for men's products. Men appeared in commercials for products used by either sex. These differences were also found in the U.S. ads, although to a lesser extent. In the United States, the increasing population of Hispanic Americans has turned many advertisers' attention to Spanish-language television advertising. Sex roles in the Hispanic culture show women as traditional homemakers, conservative, and dependent upon men for support, but many Hispanic families in the United States do not fit this traditionally held view. In 2006, more than half of Hispanic women worked outside the home, which almost matched the proportion of women in the Anglo population that worked outside the home in the United States. Therefore, many U.S. consumer products companies appear to be advertising in Mexico in the same ways in which they advertise to the general U.S. market.[1] ■

Cross-tabulations have been used to analyze sex roles in advertising in Australia, Mexico, and the United States.

## REAL RESEARCH

### *Catalogs Are Risky Business*

Twelve product categories were examined to compare catalog to store shopping. The null hypothesis that there is no significant difference in the overall amount of risk perceived when buying products by catalog compared to buying the same products in a retail store was rejected. The hypothesis was tested by computing 12 (one for each product) paired-observations $t$ tests. Mean scores for overall perceived risk for some of the products in both buying situations are presented in the following table, with higher scores indicating greater risk.

Mean Scores of Overall Perceived Risk for Products by Purchase Mode

| | *Overall Perceived Risk* | |
|---|---|---|
| ***Product*** | ***Catalog*** | ***Retail Store*** |
| Stereo hi-fi | 48.89 | 41.98* |
| Record albums | 32.65 | 28.74* |
| Dress shoes | 58.60 | 50.80* |
| 13-inch color TV | 48.53 | 40.91* |
| Athletic socks | 35.22 | 30.22* |
| Pocket calculator | 49.62 | 42.00* |
| 35-mm camera | 48.13 | 39.52* |
| Perfume | 34.85 | 29.79* |

*Significant at 0.01 level

As can be seen, a significantly ($p < 0.01$) higher overall amount of perceived risk was attached to products purchased by catalog as compared to those purchased from a retail store. Although this study reveals risk associated with catalog purchasing, terrorist threats, time shortage, and increased convenience have increased the amount of products that are purchased from catalogs.[2] ■

These two examples show how basic data analysis can be useful in its own right. The cross-tabulation and chi-square analysis in the international television advertising example, and the paired $t$ tests in the catalog shopping example, enabled us to draw specific conclusions from the data. These and other concepts discussed in this chapter are illustrated in the context of explaining Internet usage for personal (nonprofessional) reasons. Table 15.1 contains data for 30 respondents giving the sex (1 = male, 2 = female), familiarity with the Internet (1 = very unfamiliar, 7 = very familiar), Internet usage in hours per week, attitude toward Internet and toward technology, both measured on a 7-point scale (1 = very

**TABLE 15.1**
Internet Usage Data

| Respondent Number | Sex | Familiarity | Internet Usage | Attitude Toward Internet | Attitude Toward Technology | Usage of Internet: Shopping | Usage of Internet: Banking |
|---|---|---|---|---|---|---|---|
| 1 | 1.00 | 7.00 | 14.00 | 7.00 | 6.00 | 1.00 | 1.00 |
| 2 | 2.00 | 2.00 | 2.00 | 3.00 | 3.00 | 2.00 | 2.00 |
| 3 | 2.00 | 3.00 | 3.00 | 4.00 | 3.00 | 1.00 | 2.00 |
| 4 | 2.00 | 3.00 | 3.00 | 7.00 | 5.00 | 1.00 | 2.00 |
| 5 | 1.00 | 7.00 | 13.00 | 7.00 | 7.00 | 1.00 | 1.00 |
| 6 | 2.00 | 4.00 | 6.00 | 5.00 | 4.00 | 1.00 | 2.00 |
| 7 | 2.00 | 2.00 | 2.00 | 4.00 | 5.00 | 2.00 | 2.00 |
| 8 | 2.00 | 3.00 | 6.00 | 5.00 | 4.00 | 2.00 | 2.00 |
| 9 | 2.00 | 3.00 | 6.00 | 6.00 | 4.00 | 1.00 | 2.00 |
| 10 | 1.00 | 9.00 | 15.00 | 7.00 | 6.00 | 1.00 | 2.00 |
| 11 | 2.00 | 4.00 | 3.00 | 4.00 | 3.00 | 2.00 | 2.00 |
| 12 | 2.00 | 5.00 | 4.00 | 6.00 | 4.00 | 2.00 | 2.00 |
| 13 | 1.00 | 6.00 | 9.00 | 6.00 | 5.00 | 2.00 | 1.00 |
| 14 | 1.00 | 6.00 | 8.00 | 3.00 | 2.00 | 2.00 | 2.00 |
| 15 | 1.00 | 6.00 | 5.00 | 5.00 | 4.00 | 1.00 | 2.00 |
| 16 | 2.00 | 4.00 | 3.00 | 4.00 | 3.00 | 2.00 | 2.00 |
| 17 | 1.00 | 6.00 | 9.00 | 5.00 | 3.00 | 1.00 | 1.00 |
| 18 | 1.00 | 4.00 | 4.00 | 5.00 | 4.00 | 1.00 | 2.00 |
| 19 | 1.00 | 7.00 | 14.00 | 6.00 | 6.00 | 1.00 | 1.00 |
| 20 | 2.00 | 6.00 | 6.00 | 6.00 | 4.00 | 2.00 | 2.00 |
| 21 | 1.00 | 6.00 | 9.00 | 4.00 | 2.00 | 2.00 | 2.00 |
| 22 | 1.00 | 5.00 | 5.00 | 5.00 | 4.00 | 2.00 | 1.00 |
| 23 | 2.00 | 3.00 | 2.00 | 4.00 | 2.00 | 2.00 | 2.00 |
| 24 | 1.00 | 7.00 | 15.00 | 6.00 | 6.00 | 1.00 | 1.00 |
| 25 | 2.00 | 6.00 | 6.00 | 5.00 | 3.00 | 1.00 | 2.00 |
| 26 | 1.00 | 6.00 | 13.00 | 6.00 | 6.00 | 1.00 | 1.00 |
| 27 | 2.00 | 5.00 | 4.00 | 5.00 | 5.00 | 1.00 | 1.00 |
| 28 | 2.00 | 4.00 | 2.00 | 3.00 | 2.00 | 2.00 | 2.00 |
| 29 | 1.00 | 4.00 | 4.00 | 5.00 | 3.00 | 1.00 | 2.00 |
| 30 | 1.00 | 3.00 | 3.00 | 7.00 | 5.00 | 1.00 | 2.00 |


**SPSS Data File**

unfavorable, 7 = very favorable), and whether the respondents have done shopping or banking on the Internet (1 = yes, 2 = no). For illustrative purposes, we consider only a small number of observations. In actual practice, the analysis is performed on a much larger sample such as that in the Dell Experiential Research considered later. As a first step in the analysis, it is often useful to examine the frequency distributions of the relevant variables.

## FREQUENCY DISTRIBUTION

Marketing researchers often need to answer questions about a single variable. For example:

- How many users of the brand may be characterized as brand loyal?
- What percentage of the market consists of heavy users, medium users, light users, and nonusers?
- How many customers are very familiar with a new product offering? How many are familiar, somewhat familiar, and unfamiliar with the brand? What is the mean familiarity rating? Is there much variance in the extent to which customers are familiar with the new product?
- What is the income distribution of brand users? Is this distribution skewed toward low-income brackets?

SPSS Output File

**TABLE 15.2**
**Frequency Distribution of Familiarity with the Internet**

| Value Label | Value | Frequency (*N*) | Percentage | Valid Percentage | Cumulative Percentage |
|---|---|---|---|---|---|
| Very unfamiliar | 1 | 0 | 0.0 | 0.0 | 0.0 |
| | 2 | 2 | 6.7 | 6.9 | 6.9 |
| | 3 | 6 | 20.0 | 20.7 | 27.6 |
| | 4 | 6 | 20.0 | 20.7 | 48.3 |
| | 5 | 3 | 10.0 | 10.3 | 58.6 |
| | 6 | 8 | 26.7 | 27.6 | 86.2 |
| Very familiar | 7 | 4 | 13.3 | 13.8 | 100.0 |
| Missing | 9 | 1 | 3.3 | | |
| | TOTAL | 30 | 100.0 | 100.0 | |

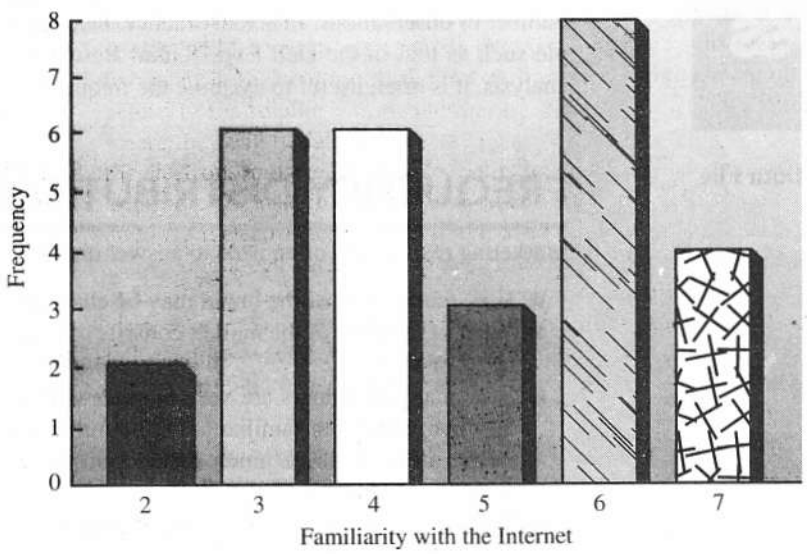
*frequency distribution*
A mathematical distribution whose objective is to obtain a count of the number of responses associated with different values of one variable and to express these counts in percentage terms.

The answers to these kinds of questions can be determined by examining frequency distributions. In a ***frequency distribution,*** one variable is considered at a time. The objective is to obtain a count of the number of responses associated with different values of the variable. The relative occurrence, or frequency, of different values of the variable is expressed in percentages. A frequency distribution for a variable produces a table of frequency counts, percentages, and cumulative percentages for all the values associated with that variable.

Table 15.2 gives the frequency distribution of familiarity with the Internet. In the table, the first column contains the labels assigned to the different categories of the variable, and the second column indicates the codes assigned to each value. Note that a code of 9 has been assigned to missing values. The third column gives the number of respondents checking each value. For example, three respondents checked value 5, indicating that they were somewhat familiar with the Internet. The fourth column displays the percentage of respondents checking each value. The next column shows percentages calculated by excluding the cases with missing values. If there are no missing values, columns 4 and 5 are identical. The last column represents cumulative percentages after adjusting for missing cases. As can be seen, of the 30 respondents who participated in the survey, 10.0 percent checked value 5. If the one respondent with a missing value is excluded, this percentage changes to 10.3. The cumulative percentage corresponding to the value of 5 is 58.6. In other words, 58.6 percent of the respondents with valid responses indicated a familiarity value of 5 or less.

***Figure 15.1***
Frequency Histogram

SPSS Output File



A frequency distribution helps determine the extent of item nonresponse (1 respondent out of 30 in Table 15.1). It also indicates the extent of illegitimate responses. Values of 0 and 8 would be illegitimate responses, or errors. The cases with these values could be identified and corrective action taken. The presence of outliers or cases with extreme values can also be detected. In the case of a frequency distribution of household size, a few isolated families with household sizes of 9 or more might be considered outliers. A frequency distribution also indicates the shape of the empirical distribution of the variable. The frequency data may be used to construct a histogram, or a vertical bar chart, in which the values of the variable are portrayed along the *X*-axis and the absolute or relative frequencies of the values are placed along the *Y*-axis. Figure 15.1 is a histogram of the frequency data in Table 15.1. From the histogram, one could examine whether the observed distribution is consistent with an expected or assumed distribution, such as the normal distribution.

**REAL RESEARCH**

## *Basic Analysis Yields Olympic Results*

For the 1996 Olympic games in Atlanta, more than 2 million unique visitors came to the games and more than 11 million tickets were sold. In Sydney, at the 2000 Olympic games, as well as in Athens at the 2004 Olympic games, 5 million tickets were sold. It is obvious that this is a potential target market that cannot be ignored. Researchers at the University of Colorado at Boulder decided to find what motivated the international and domestic travelers to come to the Olympic games in Atlanta. A survey was developed and administered to visitors via personal interviews during a nine-day period surrounding the completion of the 1996 Olympic games. Three hundred twenty surveys were completed correctly and were used in the data analysis.

The results (see the following table) showed that the top three factors that motivated people to attend the games were: a once-in-a-lifetime opportunity, availability of housing, and availability of tickets. The results of this study helped planners for the 2008 Olympic games in Beijing find what specific characteristics the city needed to improve. For instance, from this research, Beijing put funds into projects that added hotel rooms to the city. They also constructed state-of-the-art transportation and unique venues (Olympic parks, stadiums, tourist sites) so that visitors truly feel that they are getting a once-in-a-lifetime experience. As this survey continues to evolve over the years, the data received will become very valuable to the next host city.[3]

Motivational Factors that Influenced the Decision to Attend the Olympic Games

| *Motivational Factor* | *Frequency* | *Percentage* |
|---|---|---|
| Once-in-a-lifetime opportunity | 95 | 29.7 |
| Availability of housing | 36 | 11.2 |
| Availability of tickets | 27 | 8.4 |
| Distance away from home | 24 | 7.5 |
| Business/employment | 17 | 5.3 |
| Availability of money—overall expenses | 17 | 5.3 |
| Availability of time | 12 | 3.8 |
| Personal relationship with participant or official | 8 | 2.5 |
| Other motivational factor | 8 | 2.5 |
| Visit Atlanta | 4 | 1.3 |
| Security | 3 | 0.9 |
| Did not respond | 69 | 21.6 |
| Total | 320 | 100.0 ■ |

Note that the numbers and percentages in the preceding example indicate the extent to which the various motivational factors attract individuals to the Olympic games. Because numbers are involved, a frequency distribution can be used to calculate descriptive or summary statistics.

# STATISTICS ASSOCIATED WITH FREQUENCY DISTRIBUTION

As illustrated in the previous section, a frequency distribution is a convenient way of looking at different values of a variable. A frequency table is easy to read and provides basic information, but sometimes this information may be too detailed and the researcher must summarize it by the use of descriptive statistics. The most commonly used statistics associated with frequencies are measures of location (mean, mode, and median), measures of variability (range, interquartile range, standard deviation, and coefficient of variation), and measures of shape (skewness and kurtosis).[4]

## Measures of Location

***measures of location***
A statistic that describes a location within a data set. Measures of central tendency describe the center of the distribution.

The ***measures of location*** that we discuss are measures of central tendency because they tend to describe the center of the distribution. If the entire sample is changed by adding a fixed constant to each observation, then the mean, mode, and median change by the same fixed amount.

***mean***
The average; that value obtained by summing all elements in a set and dividing by the number of elements.

***Mean.*** The ***mean,*** or average value, is the most commonly used measure of central tendency. It is used to estimate the mean when the data have been collected using an interval or ratio scale. The data should display some central tendency, with most of the responses distributed around the mean.

The mean, $\overline{X}$, is given by

$$\overline{X} = \sum_{i=1}^{n} X_i / n$$

where

$X_i$ = Observed values of the variable $X$
$n$ = Number of observations (sample size)

If there are no outliers, the mean is a robust measure and does not change markedly as data values are added or deleted. For the frequencies given in Table 15.2, the mean value is calculated as follows:

$$\begin{aligned}\overline{X} &= (2 \times 2 + 6 \times 3 + 6 \times 4 + 3 \times 5 + 8 \times 6 + 4 \times 7)/29 \\ &= (4 + 18 + 24 + 15 + 48 + 28)/29 \\ &= 137/29 \\ &= 4.724\end{aligned}$$

***mode***
A measure of central tendency given as the value that occurs the most in a sample distribution.

***Mode.*** The ***mode*** is the value that occurs most frequently. It represents the highest peak of the distribution. The mode is a good measure of location when the variable is inherently categorical or has otherwise been grouped into categories. The mode in Table 15.2 is 6.000.

***median***
A measure of central tendency given as the value above which half of the values fall and below which half of the values fall.

***Median.*** The ***median*** of a sample is the middle value when the data are arranged in ascending or descending order. If the number of data points is even, the median is usually estimated as the midpoint between the two middle values—by adding the two middle values and dividing their sum by 2. The median is the 50th percentile. The median is an appropriate measure of central tendency for ordinal data. In Table 15.2, the median is 5.000.

As can be seen from Table 15.1, the three measures of central tendency for this distribution are different (mean = 4.724, mode = 6.000, median = 5.000). This is not surprising, because each measure defines central tendency in a different way. So which measure should be used? If the variable is measured on a nominal scale, the mode should

be used. If the variable is measured on an ordinal scale, the median is appropriate. If the variable is measured on an interval or ratio scale, the mode is a poor measure of central tendency. This can be seen from Table 15.2. Although the modal value of 6.000 has the highest frequency, it represents only 27.6 percent of the sample. In general, for interval or ratio data, the median is a better measure of central tendency, although it too ignores available information about the variable. The actual values of the variable above and below the median are ignored. The mean is the most appropriate measure of central tendency for interval or ratio data. The mean makes use of all the information available because all of the values are used in computing it. However, the mean is sensitive to extremely small or extremely large values (outliers). When there are outliers in the data, the mean is not a good measure of central tendency and it is useful to consider both the mean and the median.


**SPSS Output File**

In Table 15.2, since there are no extreme values and the data are interval, the mean value of 4.724 is a good measure of location or central tendency. Although, this value is greater than 4, it is still not high (i.e., it is less than 5). If this were a large and representative sample, the interpretation would be that people, on the average, are only moderately familiar with the Internet. This would call for both managerial action on the part of Internet service providers and public policy initiatives on the part of the governmental bodies to make people more familiar with the Internet and increase Internet usage.

## Measures of Variability

***measures of variability***
A statistic that indicates the distribution's dispersion.

The ***measures of variability***, which are calculated on interval or ratio data, include the range, interquartile range, variance or standard deviation, and coefficient of variation.

***range***
The difference between the largest and smallest values of a distribution.

***Range.*** The ***range*** measures the spread of the data. It is simply the difference between the largest and smallest values in the sample. As such, the range is directly affected by outliers.

$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$

If all the values in the data are multiplied by a constant, the range is multiplied by the same constant. The range in Table 15.2 is $7 - 2 = 5.000$.

***interquartile range***
The range of a distribution encompassing the middle 50 percent of the observations.

***Interquartile Range.*** The ***interquartile range*** is the difference between the 75th and 25th percentile. For a set of data points arranged in order of magnitude, the $p$th percentile is the value that has $p$ percent of the data points below it and $(100 - p)$ percent above it. If all the data points are multiplied by a constant, the interquartile range is multiplied by the same constant. The interquartile range in Table 15.2 is $6 - 3 = 3.000$.

***variance***
The mean squared deviation of all the values from the mean.

***standard deviation***
The square root of the variance.

***Variance and Standard Deviation.*** The difference between the mean and an observed value is called the *deviation from the mean*. The ***variance*** is the mean squared deviation from the mean. The variance can never be negative. When the data points are clustered around the mean, the variance is small. When the data points are scattered, the variance is large. If all the data values are multiplied by a constant, the variance is multiplied by the square of the constant. The ***standard deviation*** is the square root of the variance. Thus, the standard deviation is expressed in the same units as the data, rather than in squared units. The standard deviation of a sample, $s$, is calculated as:

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

We divide by $n - 1$ instead of $n$ because the sample is drawn from a population and we are trying to determine how much the responses vary from the mean of the entire population. However, the population mean is unknown; therefore the sample mean is used instead. The use of the sample mean makes the sample seem less variable than it

really is. By dividing by $n - 1$, instead of $n$, we compensate for the smaller variability observed in the sample. For the data given in Table 15.2, the variance is calculated as follows:

$$s^2 = \frac{\{2 \times (2 - 4.724)^2 + 6 \times (3 - 4.724)^2 + 6 \times (4 - 4.724)^2 + 3 \times (5 - 4.724)^2 + 8 \times (6 - 4.724)^2 + 4 \times (7 - 4.724)^2\}}{28}$$

$$= \frac{\{14.840 + 17.833 + 3.145 + 0.229 + 13.025 + 20.721\}}{28}$$

$$= \frac{69.793}{28}$$

$$= 2.493$$

The standard deviation, therefore, is calculated as:

$$s = \sqrt{2.493} = 1.579$$

***coefficient of variation***
A useful expression in sampling theory for the standard deviation as a percentage of the mean.

***Coefficient of Variation.*** The ***coefficient of variation*** is the ratio of the standard deviation to the mean, expressed as a percentage, and it is a unitless measure of relative variability. The coefficient of variation, *CV*, is expressed as:

$$CV = \frac{s}{\bar{X}}$$

The coefficient of variation is meaningful only if the variable is measured on a ratio scale. It remains unchanged if all the data values are multiplied by a constant. Because familiarity with the Internet is not measured on a ratio scale, it is not meaningful to calculate the coefficient of variation for the data in Table 15.2. From a managerial viewpoint, measures of variability are important because if a characteristic shows good variability, then perhaps the market could be segmented based on that characteristic.

## Measures of Shape

In addition to measures of variability, measures of shape are also useful in understanding the nature of the distribution. The shape of a distribution is assessed by examining skewness and kurtosis.

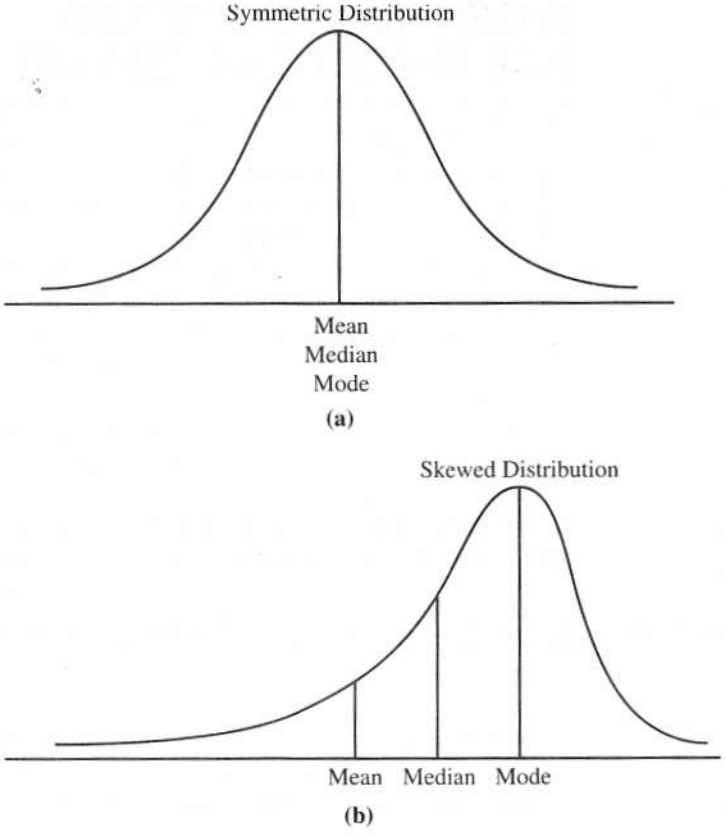
***Skewness.*** Distributions can be either symmetric or skewed. In a symmetric distribution, the values on either side of the center of the distribution are the same, and the mean, mode, and median are equal. The positive and corresponding negative deviations from the mean are also equal. In a skewed distribution, the positive and negative deviations from the mean are unequal. ***Skewness*** is the tendency of the deviations from the mean to be larger in one direction than in the other. It can be thought of as the tendency for one tail of the distribution to be heavier than the other (see Figure 15.2). The skewness value for the data of Table 15.2 is –0.094, indicating a slight negative skew.

***skewness***
A characteristic of a distribution that assesses its symmetry about the mean.

***kurtosis***
A measure of the relative peakedness or flatness of the curve defined by the frequency distribution.

***Kurtosis.*** ***Kurtosis*** is a measure of the relative peakedness or flatness of the curve defined by the frequency distribution. The kurtosis of a normal distribution is zero. If the kurtosis is positive, then the distribution is more peaked than a normal distribution. A negative value means that the distribution is flatter than a normal distribution. The value of this statistic for Table 15.2 is $-1.261$, indicating that the distribution is flatter than a normal distribution. Measures of shape are important, because if a distribution is highly skewed or markedly peaked or flat, then statistical procedures that assume normality should be used with caution.

*Figure 15.2*
Skewness of a Distribution




**ACTIVE RESEARCH**

Visit *www.wendys.com* and conduct an Internet search using a search engine and your library's online database to obtain information on the heavy users of fast-food restaurants.

As the marketing director for Wendy's, how would you target the heavy users of fast-food restaurants?

In a survey for Wendy's, information was obtained on the number of visits to Wendy's per month. How would you identify the heavy users of Wendy's and what statistics would you compute to summarize the number of visits to Wendy's per month?

## INTRODUCTION TO HYPOTHESIS TESTING

Basic analysis invariably involves some hypothesis testing. Examples of hypotheses generated in marketing research abound:

- The department store is being patronized by more than 10 percent of the households.
- The heavy and light users of a brand differ in terms of psychographic characteristics.
- One hotel has a more upscale image than its close competitor.
- Familiarity with a restaurant results in greater preference for that restaurant.

Chapter 12 covered the concepts of the sampling distribution, standard error of the mean or the proportion, and the confidence interval.[5] All these concepts are relevant to hypothesis testing and should be reviewed. Now we describe a general procedure for hypothesis testing that can be applied to test hypotheses about a wide range of parameters.

# A GENERAL PROCEDURE FOR HYPOTHESIS TESTING

The following steps are involved in hypothesis testing (Figure 15.3).

1. Formulate the null hypothesis $H_0$ and the alternative hypothesis $H_1$.
2. Select an appropriate statistical technique and the corresponding test statistic.
3. Choose the level of significance, $\alpha$.
4. Determine the sample size and collect the data. Calculate the value of the test statistic.
5. Determine the probability associated with the test statistic under the null hypothesis, using the sampling distribution of the test statistic. Alternatively, determine the critical values associated with the test statistic that divide the rejection and nonrejection regions.
6. Compare the probability associated with the test statistic with the level of significance specified. Alternatively, determine whether the test statistic has fallen into the rejection or the nonrejection region.
7. Make the statistical decision to reject or not reject the null hypothesis.
8. Express the statistical decision in terms of the marketing research problem.

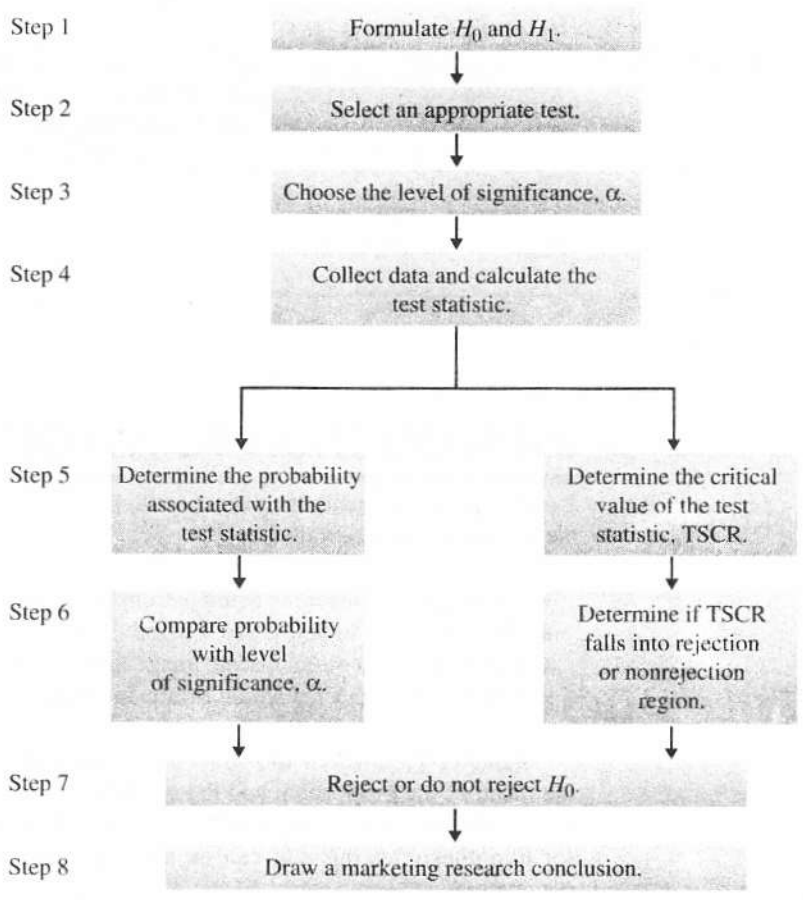
***null hypothesis***
A statement in which no difference or effect is expected. If the null hypothesis is not rejected, no changes will be made.

***alternative hypothesis***
A statement that some difference or effect is expected. Accepting the alternative hypothesis will lead to changes in opinions or actions.

## Step 1: Formulate the Hypotheses

The first step is to formulate the null and alternative hypotheses. A ***null hypothesis*** is a statement of the status quo, one of no difference or no effect. If the null hypothesis is not rejected, no changes will be made. An ***alternative hypothesis*** is one in which some difference or effect is expected. Accepting the alternative hypothesis will lead to changes in opinions or actions. Thus, the alternative hypothesis is the opposite of the null hypothesis.

*Figure 15.3*
A General Procedure for Hypothesis Testing



The null hypothesis is always the hypothesis that is tested. The null hypothesis refers to a specified value of the population parameter (e.g., $\mu$, $\sigma$, $\pi$), not a sample statistic (e.g., $\bar{X}$). A null hypothesis may be rejected, but it can never be accepted based on a single test. A statistical test can have one of two outcomes. One is that the null hypothesis is rejected and the alternative hypothesis accepted. The other outcome is that the null hypothesis is not rejected based on the evidence. However, it would be incorrect to conclude that because the null hypothesis is not rejected, it can be accepted as valid. In classical hypothesis testing, there is no way to determine whether the null hypothesis is true.

In marketing research, the null hypothesis is formulated in such a way that its rejection leads to the acceptance of the desired conclusion. The alternative hypothesis represents the conclusion for which evidence is sought. For example, a major department store is considering the introduction of an Internet shopping service. The new service will be introduced if more than 40 percent of the Internet users shop via the Internet. The appropriate way to formulate the hypotheses is:

$$H_0: \pi \leq 0.40$$
$$H_1: \pi > 0.40$$

If the null hypothesis $H_0$ is rejected, then the alternative hypothesis $H_1$ will be accepted and the new Internet shopping service will be introduced. On the other hand, if $H_0$ is not rejected, then the new service should not be introduced unless additional evidence is obtained.

***one-tailed test***
A test of the null hypothesis where the alternative hypothesis is expressed directionally.

***two-tailed test***
A test of the null hypothesis where the alternative hypothesis is not expressed directionally.

This test of the null hypothesis is a ***one-tailed test,*** because the alternative hypothesis is expressed directionally: The proportion of Internet users who use the Internet for shopping is greater than 0.40. On the other hand, suppose the researcher wanted to determine whether the proportion of Internet users who shop via the Internet is different from 40 percent. Then a ***two-tailed test*** would be required, and the hypotheses would be expressed as:

$$H_0: \pi = 0.40$$
$$H_1: \pi \neq 0.40$$

In commercial marketing research, the one-tailed test is used more often than a two-tailed test. Typically, there is some preferred direction for the conclusion for which evidence is sought. For example, the higher the profits, sales, and product quality, the better. The one-tailed test is more powerful than the two-tailed test. The power of a statistical test is discussed further in step 3.

## Step 2: Select an Appropriate Test

To test the null hypothesis, it is necessary to select an appropriate statistical technique. The researcher should take into consideration how the test statistic is computed and the sampling distribution that the sample statistic (e.g., the mean) follows. The ***test statistic*** measures how close the sample has come to the null hypothesis. The test statistic often follows a well-known distribution, such as the normal, $t$, or chi-square distribution. Guidelines for selecting an appropriate test or statistical technique are discussed later in this chapter. In our example, the $z$ statistic, which follows the standard normal distribution, would be appropriate. This statistic would be computed as follows:

***test statistic***
A measure of how close the sample has come to the null hypothesis. It often follows a well-known distribution, such as the normal, $t$, or chi-square distribution.

$$z = \frac{p - \pi}{\sigma_p}$$

where

$$\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{n}}$$

## Step 3: Choose Level of Significance, $\alpha$

Whenever we draw inferences about a population, there is a risk that an incorrect conclusion will be reached. Two types of errors can occur:

***type I error***
Also known as alpha error, it occurs when the sample results lead to the rejection of a null hypothesis that is in fact true.

***level of significance***
The probability of making a type I error.

***type II error***
Also known as beta error, it occurs when the sample results lead to the nonrejection of a null hypothesis that is in fact false.

***power of a test***
The probability of rejecting the null hypothesis when it is in fact false and should be rejected.

***Type I Error.*** ***Type I error*** occurs when the sample results lead to the rejection of the null hypothesis when it is in fact true. In our example, a Type I error would occur if we concluded, based on the sample data, that the proportion of customers preferring the new service plan was greater than 0.40, when in fact it was less than or equal to 0.40. The probability of Type I error ($\alpha$) is also called the ***level of significance.*** The Type I error is controlled by establishing the tolerable level of risk of rejecting a true null hypothesis. The selection of a particular risk level should depend on the cost of making a Type I error.

***Type II Error.*** ***Type II error*** occurs when, based on the sample results, the null hypothesis is not rejected when it is in fact false. In our example, the Type II error would occur if we concluded, based on sample data, that the proportion of customers preferring the new service plan was less than or equal to 0.40 when, in fact, it was greater than 0.40. The probability of Type II error is denoted by $\beta$. Unlike $\alpha$, which is specified by the researcher, the magnitude of $\beta$ depends on the actual value of the population parameter (proportion). The probability of Type I error ($\alpha$) and the probability of Type II error ($\beta$) are shown in Figure 15.4. The complement ($1 - \beta$) of the probability of a Type II error is called the *power of a statistical test.*

***Power of a Test.*** The ***power of a test*** is the probability ($1 - \beta$) of rejecting the null hypothesis when it is false and should be rejected. Although $\beta$ is unknown, it is related to $\alpha$. An extremely low value of $\alpha$ (e.g., = 0.001) will result in intolerably high $\beta$ errors. So it is necessary to balance the two types of errors. As a compromise, $\alpha$ is often set at 0.05; sometimes it is 0.01; other values of $\alpha$ are rare. The level of $\alpha$, along with the sample size, will determine the level of $\beta$ for a particular research design. The risk of both $\alpha$ and $\beta$ can be controlled by increasing the sample size. For a given level of $\alpha$, increasing the sample size will decrease $\beta$, thereby increasing the power of the test.

## Step 4: Collect Data and Calculate Test Statistic

Sample size is determined after taking into account the desired $\alpha$ and $\beta$ errors and other qualitative considerations, such as budget constraints. Then the required data are collected and the value of the test statistic computed. In our example, 30 users were surveyed and 17 indicated that they used the Internet for shopping. Thus the value of the sample proportion is $p = 17/30 = 0.567$.

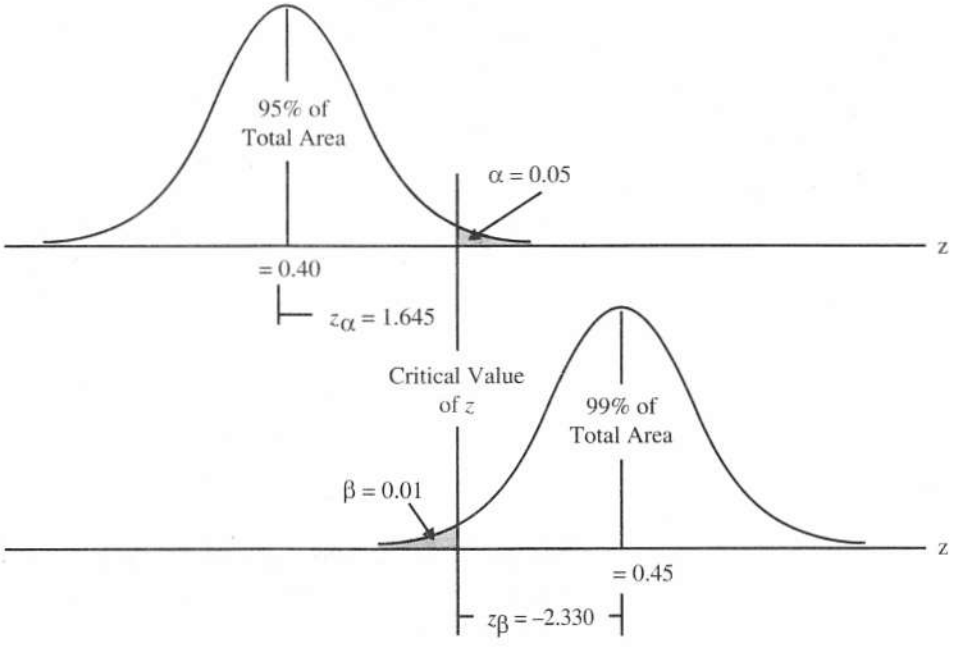



***Figure 15.4***
Type I Error ($\alpha$) and Type II Error ($\beta$)

The value of $\sigma_p$ can be determined as follows:

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$
$$= \sqrt{\frac{(0.40)(0.60)}{30}}$$
$$= 0.089$$

The test statistic $z$ can be calculated as follows:

$$z = \frac{p-\pi}{\sigma_p}$$
$$= \frac{0.567 - 0.40}{0.089}$$
$$= 1.88$$

## Step 5: Determine the Probability (Critical Value)

Using standard normal tables (Table 2 of the Statistical Appendix), the probability of obtaining a $z$ value of 1.88 can be calculated (see Figure 15.5). The shaded area between $-\infty$ and 1.88 is 0.9699. Therefore, the area to the right of $z = 1.88$ is $1.0000 - 0.9699 = 0.0301$. Alternatively, the critical value of $z$, which will give an area to the right side of the critical value of 0.05, is between 1.64 and 1.65 and equals 1.645. Note that in determining the critical value of the test statistic, the area to the right of the critical value is either $\alpha$ or $\alpha/2$. It is $\alpha$ for a one-tail test and $\alpha/2$ for a two-tail test.

## Steps 6 and 7: Compare the Probability (Critical Value) and Make the Decision

The probability associated with the calculated or observed value of the test statistic is 0.0301. This is the probability of getting a $p$ value of 0.567 when $\pi = 0.40$. This is less than the level of significance of 0.05. Hence, the null hypothesis is rejected. Alternatively, the calculated value of the test statistic $z = 1.88$ lies in the rejection region, beyond the value of 1.645. Again, the same conclusion to reject the null hypothesis is reached. Note that the two ways of testing the null hypothesis are equivalent but mathematically opposite in the direction of comparison. If the probability associated with the calculated or observed value of the test statistic ($TS_{CAL}$) is *less than* the level of significance ($\alpha$), the null hypothesis is rejected. However, if the calculated value of the test statistic is *greater than* the critical value of the test statistic ($TS_{CR}$), the null hypothesis is rejected. The reason for this sign shift is that the larger the value of $TS_{CAL}$, the smaller the probability of obtaining a more extreme value of the test statistic under the null hypothesis. This sign shift can be easily seen:

if probability of $TS_{CAL}$ < significance level ($\alpha$), then reject $H_0$,

but

if $TS_{CAL} > TS_{CR}$, then reject $H_0$.

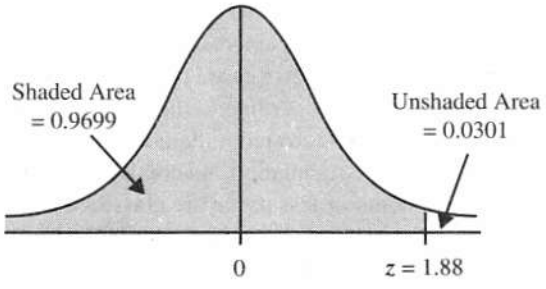
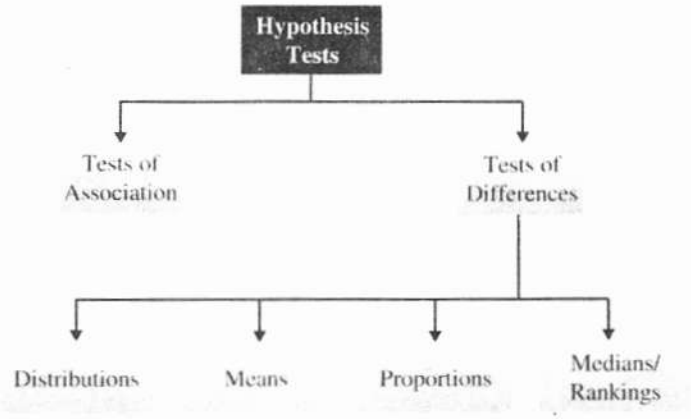



***Figure 15.5***
Probability of z with a One-Tailed Test

*Figure 15.6*
A Broad Classification of Hypothesis Tests




## Step 8: Marketing Research Conclusion

The conclusion reached by hypothesis testing must be expressed in terms of the marketing research problem. In our example, we conclude that there is evidence that the proportion of Internet users who shop via the Internet is significantly greater than 0.40. Hence, the recommendation to the department store would be to introduce the new Internet shopping service.

As can be seen from Figure 15.6, hypotheses testing can be related to either an examination of associations or an examination of differences. In tests of associations, the null hypothesis is that there is no association between the variables ($H_0$: . . . is NOT related to . . .). In tests of differences, the null hypothesis is that there is no difference ($H_0$: . . . is NOT different from . . .). Tests of differences could relate to distributions, means, proportions, medians, or rankings. First, we discuss hypotheses related to associations in the context of cross-tabulations.

# CROSS-TABULATIONS

Although answers to questions related to a single variable are interesting, they often raise additional questions about how to link that variable to other variables. To introduce the frequency distribution, we posed several representative marketing research questions. For each of these, a researcher might pose additional questions to relate these variables to other variables. For example:

- How many brand-loyal users are males?
- Is product use (measured in terms of heavy users, medium users, light users, and nonusers) related to interest in outdoor activities (high, medium, and low)?
- Is familiarity with a new product related to age and education levels?
- Is product ownership related to income (high, medium, and low)?

***cross-tabulation***
A statistical technique that describes two or more variables simultaneously and results in tables that reflect the joint distribution of two or more variables that have a limited number of categories or distinct values.

The answers to such questions can be determined by examining cross-tabulations. Whereas a frequency distribution describes one variable at a time, a ***cross-tabulation*** describes two or more variables simultaneously. A cross-tabulation is the merging of the frequency distribution of two or more variables in a single table. It helps us to understand how one variable such as brand loyalty relates to another variable such as sex. Cross-tabulation results in tables that reflect the joint distribution of two or more variables with a limited number of categories or distinct values. The categories of one variable are cross-classified with the categories of one or more other variables. Thus, the frequency distribution of one variable is subdivided according to the values or categories of the other variables.

Suppose we are interested in determining whether Internet usage is related to sex. For the purpose of cross-tabulation, respondents are classified as light or heavy users. Those reporting five hours or less usage are classified as light users, and the remaining are heavy users. The cross-tabulation is shown in Table 15.3. A cross-tabulation includes a cell for every combination of the categories of the two variables. The number in each cell shows

**SPSS Output File**

**TABLE 15.3**

Sex and Internet Usage

| | Sex | | |
|---|---|---|---|
| Internet Usage | Male | Female | Row Total |
| Light (1) | 5 | 10 | 15 |
| Heavy (2) | 10 | 5 | 15 |
| Column total | 15 | 15 | |

how many respondents gave that combination of responses. In Table 15.3, 10 respondents were females who reported light Internet usage. The marginal totals in this table indicate that of the 30 respondents with valid responses on both the variables, 15 reported light usage and 15 were heavy users. In terms of sex, 15 respondents were females and 15 were males. Note that this information could have been obtained from a separate frequency distribution for each variable. In general, the margins of a cross-tabulation show the same information as the frequency tables for each of the variables. Cross-tabulation tables are also called ***contingency tables.*** The data are considered to be qualitative or categorical data, because each variable is assumed to have only a nominal scale.[6]

***contingency table***
A cross-tabulation table. It contains a cell for every combination of categories of the two variables.

Cross-tabulation is widely used in commercial marketing research, because (1) cross-tabulation analysis and results can be easily interpreted and understood by managers who are not statistically oriented; (2) the clarity of interpretation provides a stronger link between research results and managerial action; (3) a series of cross-tabulations may provide greater insights into a complex phenomenon than a single multivariate analysis; (4) cross-tabulation may alleviate the problem of sparse cells, which could be serious in discrete multivariate analysis; and (5) cross-tabulation analysis is simple to conduct and appealing to less sophisticated researchers.[7]

## Two Variables

Cross-tabulation with two variables is also known as bivariate cross-tabulation. Consider again the cross-classification of Internet usage with sex given in Table 15.3. Is usage related to sex? It appears to be from Table 15.3. We see that disproportionately more of the respondents who are males are heavy Internet users as compared to females. Computation of percentages can provide more insights.

Because two variables have been cross-classified, percentages could be computed either columnwise, based on column totals (Table 15.4), or rowwise, based on row totals (Table 15.5). Which of these tables is more useful? The answer depends on which variable

**SPSS Output File**

**TABLE 15.4**

Sex by Internet Usage

| | Sex | |
|---|---|---|
| Internet Usage | Male | Female |
| Light | 33.3% | 66.7% |
| Heavy | 66.7% | 33.3% |
| Column total | 100.0% | 100.0% |

**TABLE 15.5**

Internet Usage by Sex

| | Internet Usage | | |
|---|---|---|---|
| Sex | Light | Heavy | Total |
| Male | 33.3% | 66.7% | 100.0% |
| Female | 66.7% | 33.3% | 100.0% |

will be considered as the independent variable and which as the dependent variable. The general rule is to compute the percentages in the direction of the independent variable, across the dependent variable. In our analysis, sex may be considered as the independent variable and Internet usage as the dependent variable, and the correct way of calculating percentages is as shown in Table 15.4. Note that whereas 66.7 percent of the males are heavy users, only 33.3 percent of females fall into this category. This seems to indicate that males are more likely to be heavy users of the Internet as compared to females.

Note that computing percentages in the direction of the dependent variable across the independent variable, as shown in Table 15.5, is not meaningful in this case. Table 15.5 implies that heavy Internet usage causes people to be males. This latter finding is implausible. It is possible, however, that the association between Internet usage and sex is mediated by a third variable, such as age or income. This kind of possibility points to the need to examine the effect of a third variable.

## Three Variables

Often the introduction of a third variable clarifies the initial association (or lack of it) observed between two variables. As shown in Figure 15.7, the introduction of a third variable can result in four possibilities.

1. It can refine the association observed between the two original variables.
2. It can indicate no association between the two variables, although an association was initially observed. In other words, the third variable indicates that the initial association between the two variables was spurious.
3. It can reveal some association between the two variables, although no association was initially observed. In this case, the third variable reveals a suppressed association between the first two variables: a suppressor effect.
4. It can indicate no change in the initial association.[8]

These cases are explained with examples based on a sample of 1,000 respondents. Although these examples are contrived to illustrate specific cases, such cases are not uncommon in commercial marketing research.

***Refine an Initial Relationship.*** An examination of the relationship between the purchase of fashion clothing and marital status resulted in the data reported in Table 15.6. The respondents were classified into either high or low categories based on their purchase of fashion clothing. Marital status was also measured in terms of two categories: currently married or unmarried. As can be seen from Table 15.6, 52 percent of unmarried respondents fell in the high-purchase category, as opposed to 31 percent of the married respondents.

*Figure 15.7*
The Introduction of a Third Variable in Cross-Tabulation

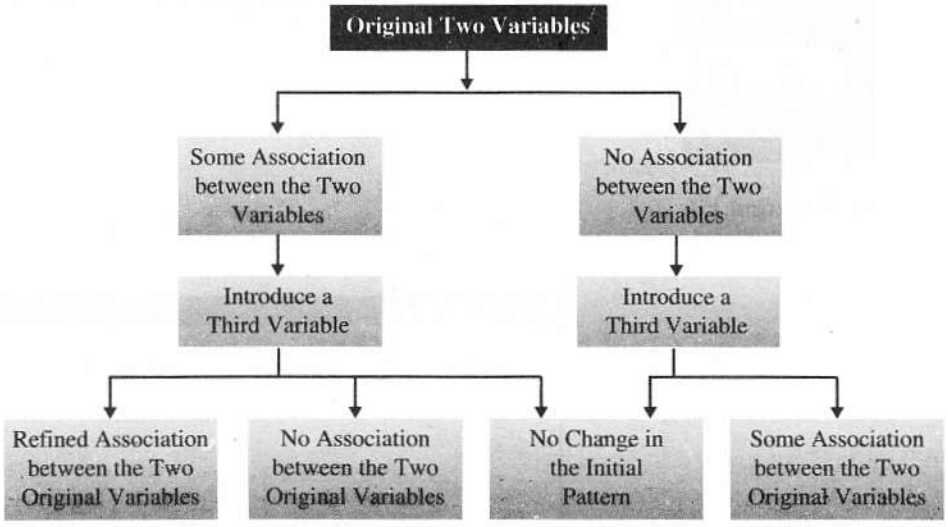


**TABLE 15.6**

**Purchase of Fashion Clothing by Marital Status**

| Purchase of Fashion Clothing | Current Marital Status | |
|---|---|---|
| | Married | Unmarried |
| High | 31% | 52% |
| Low | 69% | 48% |
| Column | 100% | 100% |
| Number of respondents | 700 | 300 |

**TABLE 15.7**

**Purchase of Fashion Clothing by Marital Status and Sex**

| Purchase of Fashion Clothing | Sex | | | |
|---|---|---|---|---|
| | Male Marital Status | | Female Marital Status | |
| | Married | Unmarried | Married | Unmarried |
| High | 35% | 40% | 25% | 60% |
| Low | 65% | 60% | 75% | 40% |
| Column totals | 100% | 100% | 100% | 100% |
| Number of cases | 400 | 120 | 300 | 180 |

Before concluding that unmarried respondents purchase more fashion clothing than those who are married, a third variable, the buyer's sex, was introduced into the analysis.

The buyer's sex was selected as the third variable based on past research. The relationship between purchase of fashion clothing and marital status was reexamined in light of the third variable, as shown in Table 15.7. In the case of females, 60 percent of the unmarried fall in the high-purchase category, as compared to 25 percent of those who are married. On the other hand, the percentages are much closer for males, with 40 percent of the unmarried and 35 percent of the married falling in the high-purchase category. Hence, the introduction of sex (third variable) has refined the relationship between marital status and purchase of fashion clothing (original variables). Unmarried respondents are more likely to fall in the high-purchase category than married ones, and this effect is much more pronounced for females than for males.

***Initial Relationship Was Spurious.*** A researcher working for an advertising agency promoting a line of automobiles costing more than $30,000 was attempting to explain the ownership of expensive automobiles (see Table 15.8). The table shows that 32 percent of those with college degrees own an expensive automobile, as compared to 21 percent of those without college degrees. The researcher was tempted to conclude that education influenced ownership of expensive automobiles. Realizing that income may also be a factor, the researcher decided to reexamine the relationship between education and ownership of expensive automobiles in light of income level. This resulted in Table 15.9. Note that the percentages of those with and without college degrees who own expensive

**TABLE 15.8**

**Ownership of Expensive Automobiles by Education Level**

| Own Expensive Automobile | Education | |
|---|---|---|
| | College Degree | No College Degree |
| Yes | 32% | 21% |
| No | 68% | 79% |
| Column total | 100% | 100% |
| Number of cases | 250 | 750 |

**TABLE 15.9**

**Ownership of Expensive Automobiles by Education and Income Levels**

| | Income | | | |
|---|---|---|---|---|
| | Low Income Education | | High Income Education | |
| Own Expensive Automobile | College Degree | No College Degree | College Degree | No College Degree |
| Yes | 20% | 20% | 40% | 40% |
| No | 80% | 80% | 60% | 60% |
| Column totals | 100% | 100% | 100% | 100% |
| Number of respondents | 100 | 700 | 150 | 50 |

automobiles are the same for each of the income groups. When the data for the high-income and low-income groups are examined separately, the association between education and ownership of expensive automobiles disappears, indicating that the initial relationship observed between these two variables was spurious.

***Reveal Suppressed Association.*** A researcher suspected desire to travel abroad may be influenced by age. However, a cross-tabulation of the two variables produced the results in Table 15.10, indicating no association. When sex was introduced as the third variable, Table 15.11 was obtained. Among men, 60 percent of those under 45 indicated a desire to travel abroad, as compared to 40 percent of those 45 or older. The pattern was reversed for women, where 35 percent of those under 45 indicated a desire to travel abroad, as opposed to 65 percent of those 45 or older. Because the association between desire to travel abroad and age runs in the opposite direction for males and females, the relationship between these two variables is masked when the data are aggregated across sex, as in Table 15.10. But when the effect of sex is controlled, as in Table 15.11, the suppressed association between desire to travel abroad and age is revealed for the separate categories of males and females.

***No Change in Initial Relationship.*** In some cases, the introduction of the third variable does not change the initial relationship observed, regardless of whether the original variables were associated. This suggests that the third variable does not influence the

**TABLE 15.10**

**Desire to Travel Abroad by Age**

| | Age | |
|---|---|---|
| Desire to Travel Abroad | Less Than 45 | 45 or More |
| Yes | 50% | 50% |
| No | 50% | 50% |
| Column total | 100% | 100% |
| Number of respondents | 500 | 500 |

**TABLE 15.11**

**Desire to Travel Abroad by Age and Sex**

| | Sex | | | |
|---|---|---|---|---|
| Desire to Travel Abroad | Male Age | | Female Age | |
| | < 45 | ≥ 45 | < 45 | ≥ 45 |
| Yes | 60% | 40% | 35% | 65% |
| No | 40% | 60% | 65% | 35% |
| Column total | 100% | 100% | 100% | 100% |
| Number of cases | 300 | 300 | 200 | 200 |

**TABLE 15.12**
**Eating Frequently in Fast-Food Restaurants by Family Size**

| Eat Frequently in Fast-Food Restaurants | Family Size | |
|---|---|---|
| | Small | Large |
| Yes | 65% | 65% |
| No | 35% | 35% |
| Column total | 100% | 100% |
| Number of cases | 500 | 500 |

**TABLE 15.13**
**Eating Frequently in Fast-Food Restaurants by Family Size and Income**

| Eat Frequently in Fast-Food Restaurants | Income | | | |
|---|---|---|---|---|
| | Low Income | | High Income | |
| | Family size | | Family size | |
| | Small | Large | Small | Large |
| Yes | 65% | 65% | 65% | 65% |
| No | 35% | 35% | 35% | 35% |
| Column total | 100% | 100% | 100% | 100% |
| Number of respondents | 250 | 250 | 250 | 250 |

relationship between the first two. Consider the cross-tabulation of family size and the tendency to eat out frequently in fast-food restaurants, as shown in Table 15.12. The respondents were classified into small and large family size categories based on a median split of the distribution, with 500 respondents in each category. No association is observed. The respondents were further classified into high- or low-income groups based on a median split. When income was introduced as a third variable in the analysis, Table 15.13 was obtained. Again, no association was observed.

## General Comments on Cross-Tabulation

More than three variables can be cross-tabulated, but the interpretation is quite complex. Also, because the number of cells increases multiplicatively, maintaining an adequate number of respondents or cases in each cell can be problematic. As a general rule, there should be at least five expected observations in each cell for the statistics computed to be reliable. Thus, cross-tabulation is an inefficient way of examining relationships when there are several variables. Note that cross-tabulation examines association between variables, not causation. To examine causation, the causal research design framework should be adopted (see Chapter 7).

# STATISTICS ASSOCIATED WITH CROSS-TABULATION

We will discuss the statistics commonly used for assessing the statistical significance and strength of association of cross-tabulated variables. The statistical significance of the observed association is commonly measured by the chi-square statistic. The strength of association, or degree of association, is important from a practical or substantive perspective. Generally, the strength of association is of interest only if the association is statistically significant. The strength of the association can be measured by the phi correlation coefficient, the contingency coefficient, Cramer's $V$, and the lambda coefficient.

## Chi-Square

***chi-square statistic***
The statistic used to test the statistical significance of the observed association in a cross-tabulation. It assists us in determining whether a systematic association exists between the two variables.

The ***chi-square statistic*** ($\chi^2$) is used to test the statistical significance of the observed association in a cross-tabulation. It assists us in determining whether a systematic association exists between the two variables. The null hypothesis, $H_0$, is that there is no association between the variables. The test is conducted by computing the cell frequencies that would be expected if no association were present between the variables, given the existing row and column totals. These expected cell frequencies, denoted $f_e$, are then compared to the actual observed frequencies, $f_o$, found in the cross-tabulation to calculate the chi-square statistic. The greater the discrepancies between the expected and actual frequencies, the larger the value of the statistic. Assume that a cross-tabulation has $r$ rows and $c$ columns and a random sample of $n$ observations. Then the expected frequency for each cell can be calculated by using a simple formula:

$$f_e = \frac{n_r n_c}{n}$$

where

$n_r$ = total number in the row
$n_c$ = total number in the column
$n$ = total sample size


**SPSS Output File**

For the data in Table 15.3, the expected frequencies for the cells, going from left to right and from top to bottom, are:

$$\frac{15 \times 15}{30} = 7.50 \qquad \frac{15 \times 15}{30} = 7.50$$
$$\frac{15 \times 15}{30} = 7.50 \qquad \frac{15 \times 15}{30} = 7.50$$

Then the value of $\chi^2$ is calculated as follows:

$$\chi^2 = \sum_{\substack{\text{all} \\ \text{cells}}} \frac{(f_o - f_e)^2}{f_e}$$

For the data in Table 15.3, the value $\chi^2$ of is calculated as:

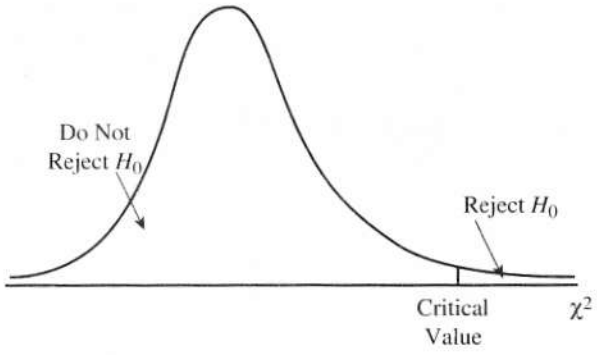
$$\chi^2 = \frac{(5 - 7.5)^2}{7.5} + \frac{(10 - 7.5)^2}{7.5} + \frac{(10 - 7.5)^2}{7.5} + \frac{(5 - 7.5)^2}{7.5}$$
$$= 0.833 + 0.833 + 0.833 + 0.833$$
$$= 3.333$$

To determine whether a systematic association exists, the probability of obtaining a value of chi-square as large or larger than the one calculated from the cross-tabulation is estimated. An important characteristic of the chi-square statistic is the number of degrees of freedom (df) associated with it. In general, the number of degrees of freedom is equal to the number of observations less the number of constraints needed to calculate a statistical term. In the case of a chi-square statistic associated with a cross-tabulation, the number of degrees of freedom is equal to the product of number of rows ($r$) less one and the number of columns ($c$) less one. That is, df = $(r - 1) \times (c - 1)$.[9] The null hypothesis ($H_0$) of no association between the two variables will be rejected only when the calculated value of the test statistic is greater than the critical value of the chi-square distribution with the appropriate degrees of freedom, as shown in Figure 15.8.

***chi-square distribution***
A skewed distribution whose shape depends solely on the number of degrees of freedom. As the number of degrees of freedom increases, the chi-square distribution becomes more symmetrical.

The ***chi-square distribution*** is a skewed distribution whose shape depends solely on the number of degrees of freedom.[10] As the number of degrees of freedom increases, the chi-square distribution becomes more symmetrical. Table 3 in the Statistical Appendix contains upper-tail areas of the chi-square distribution for different degrees of freedom. In this table, the value at the top of each column indicates the area in the upper portion (the right side, as shown in Figure 15.8) of the chi-square distribution. To illustrate, for 1 degree of freedom, the value for an upper-tail area of 0.05 is 3.841. This indicates that for 1 degree

*Figure 15.8*
Chi-Square Test of Association




of freedom, the probability of exceeding a chi-square value of 3.841 is 0.05. In other words, at the 0.05 level of significance with 1 degree of freedom, the critical value of the chi-square statistic is 3.841.

For the cross-tabulation given in Table 15.3, there are $(2 - 1) \times (2 - 1) = 1$ degree of freedom. The calculated chi-square statistic had a value of 3.333. Because this is less than the critical value of 3.841, the null hypothesis of no association cannot be rejected, indicating that the association is not statistically significant at the 0.05 level. Note that this lack of significance is mainly due to the small sample size (30). If, instead, the sample size were 300 and each entry of Table 15.3 were multiplied by 10, it can be seen that the value of the chi-square statistic would be multiplied by 10 and would be 33.33, which is significant at the 0.05 level.

The chi-square statistic can also be used in goodness-of-fit tests to determine whether certain models fit the observed data. These tests are conducted by calculating the significance of sample deviations from assumed theoretical (expected) distributions, and can be performed on cross-tabulations as well as on frequencies (one-way tabulations). The calculation of the chi-square statistic and the determination of its significance is the same as illustrated above.

The chi-square statistic should be estimated only on counts of data. When the data are in percentage form, they should first be converted to absolute counts or numbers. In addition, an underlying assumption of the chi-square test is that the observations are drawn independently. As a general rule, chi-square analysis should not be conducted when the expected or theoretical frequencies in any of the cells is less than five. If the number of observations in any cell is less than 10, or if the table has two rows and two columns (a $2 \times 2$ table), a correction factor should be applied.[11] With the correction factor, the value is 2.133, which is not significant at the 0.05 level. In the case of a $2 \times 2$ table, the chi-square is related to the phi coefficient.

## Phi Coefficient

*phi coefficient*
A measure of the strength of association in the special case of a table with two rows and two columns (a $2 \times 2$ table).

The ***phi coefficient*** ($\phi$) is used as a measure of the strength of association in the special case of a table with two rows and two columns (a $2 \times 2$ table). The phi coefficient is proportional to the square root of the chi-square statistic. For a sample of size $n$, this statistic is calculated as:

$$\phi = \sqrt{\frac{\chi^2}{n}}$$


**SPSS Output File**

It takes the value of 0 when there is no association, which would be indicated by a chi-square value of 0 as well. When the variables are perfectly associated, phi assumes the value of 1 and all the observations fall just on the main or minor diagonal. (In some computer programs, phi assumes a value of $-1$ rather than 1 when there is perfect negative association.) In our case, because the association was not significant at the 0.05 level, we would not normally compute the phi value. However, for the purpose of illustration, we show how the values of phi and other measures of the strength of association would be computed. The value of phi is:

$$\phi = \sqrt{\frac{3.333}{30}}$$
$$= 0.333$$

Thus, the association is not very strong. In the more general case involving a table of any size, the strength of association can be assessed by using the contingency coefficient.

## Contingency Coefficient

***contingency coefficient (C)***
A measure of the strength of association in a table of any size.

Whereas the phi coefficient is specific to a $2 \times 2$ table, the ***contingency coefficient*** ($C$) can be used to assess the strength of association in a table of any size. This index is also related to chi-square, as follows:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

The contingency coefficient varies between 0 and 1. The 0 value occurs in the case of no association (i.e., the variables are statistically independent), but the maximum value of 1 is never achieved. Rather, the maximum value of the contingency coefficient depends on the size of the table (number of rows and number of columns). For this reason, it should be used only to compare tables of the same size. The value of the contingency coefficient for Table 15.3 is:

$$C = \sqrt{\frac{3.333}{3.333 + 30}}$$
$$= 0.316$$

This value of $C$ indicates that the association is not very strong. Another statistic that can be calculated for any table is Cramer's $V$.

## Cramer's V

***Cramer's V***
A measure of the strength of association used in tables larger than $2 \times 2$.

***Cramer's V*** is a modified version of the phi correlation coefficient, $\phi$, and is used in tables larger than $2 \times 2$. When phi is calculated for a table larger than $2 \times 2$, it has no upper limit. Cramer's $V$ is obtained by adjusting phi for either the number of rows or the number of columns in the table, based on which of the two is smaller. The adjustment is such that $V$ will range from 0 to 1. A large value of $V$ merely indicates a high degree of association. It does not indicate how the variables are associated. For a table with $r$ rows and $c$ columns, the relationship between Cramer's $V$ and the phi correlation coefficient is expressed as:

$$V = \sqrt{\frac{\phi^2}{\min (r-1), (c-1)}}$$

or

$$V = \sqrt{\frac{\chi^2 / n}{\min (r-1), (c-1)}}$$

The value of Cramer's $V$ for Table 15.3 is:

$$V = \sqrt{\frac{3.333/30}{1}}$$
$$= 0.333$$

Thus, the association is not very strong. As can be seen, in this case $V = \phi$. This is always the case for a $2 \times 2$ table. Another statistic commonly estimated is the lambda coefficient.

## Lambda Coefficient

***asymmetric lambda***
A measure of the percentage improvement in predicting the value of the dependent variable, given the value of the independent variable in contingency table analysis. Lambda also varies between 0 and 1.

Lambda assumes that the variables are measured on a nominal scale. ***Asymmetric lambda*** measures the percentage improvement in predicting the value of the dependent variable, given the value of the independent variable. Lambda also varies between 0 and 1. A value of 0 means no improvement in prediction. A value of 1 indicates that the prediction can be made without error. This happens when each independent variable category is associated with a single category of the dependent variable.

*symmetric lambda*
The symmetric lambda does not make an assumption about which variable is dependent. It measures the overall improvement when prediction is done in both directions.

Asymmetric lambda is computed for each of the variables (treating it as the dependent variable). In general, the two asymmetric lambdas are likely to be different because the marginal distributions are not usually the same. A ***symmetric lambda*** is also computed, which is a kind of average of the two asymmetric values. The symmetric lambda does not make an assumption about which variable is dependent. It measures the overall improvement when prediction is done in both directions.[12] The value of asymmetric lambda in Table 15.3, with usage as the dependent variable, is 0.333. This indicates that knowledge of sex increases our predictive ability by the proportion of 0.333, that is, a 33.3 percent improvement. The symmetric lambda is also 0.333.

## Other Statistics

Note that in the calculation of the chi-square statistic, the variables are treated as being measured on only a nominal scale. Other statistics such as tau *b*, tau *c*, and gamma are available to measure association between two ordinal-level variables. All these statistics use information about the ordering of categories of variables by considering every possible pair of cases in the table. Each pair is examined to determine if its relative ordering on the first variable is the same as its relative ordering on the second variable (concordant), if the ordering is reversed (discordant), or if the pair is tied. The manner in which the ties are treated is the basic difference between these statistics. Both tau *b* and tau *c* adjust for ties. ***Tau b*** is the most appropriate with square tables, in which the number of rows and the number of columns are equal. Its value varies between +1 and −1. Thus the direction (positive or negative) as well as the strength (how close the value is to 1) of the relationship can be determined. For a rectangular table in which the number of rows is different from the number of columns, ***tau c*** should be used. ***Gamma*** does not make an adjustment for either ties or table size. Gamma also varies between +1 and −1 and generally has a higher numerical value than tau *b* or tau *c*. For the data in Table 15.3, as sex is a nominal variable, it is not appropriate to calculate ordinal statistics. All these statistics can be estimated by using the appropriate computer programs for cross-tabulation. Other statistics for measuring the strength of association, namely product moment correlation and nonmetric correlation, are discussed in Chapter 17.

*tau b*
Test statistic that measures the association between two ordinal-level variables. It makes an adjustment for ties and is most appropriate when the table of variables is square.

*tau c*
Test statistic that measures the association between two ordinal-level variables. It makes an adjustment for ties and is most appropriate when the table of variables is not square but a rectangle.

*gamma*
Test statistic that measures the association between two ordinal-level variables. It does not make an adjustment for ties.

# CROSS-TABULATION IN PRACTICE

When conducting cross-tabulation analysis in practice, it is useful to proceed along the following steps.

1. Test the null hypothesis that there is no association between the variables using the chi-square statistic. If you fail to reject the null hypothesis, then there is no relationship.
2. If $H_0$ is rejected, then determine the strength of the association using an appropriate statistic (phi coefficient, contingency coefficient, Cramer's *V*, lambda coefficient, or other statistics).
3. If $H_0$ is rejected, interpret the pattern of the relationship by computing the percentages in the direction of the independent variable, across the dependent variable.
4. If the variables are treated as ordinal rather than nominal, use tau *b*, tau *c*, or gamma as the test statistic. If $H_0$ is rejected, then determine the strength of the association using the magnitude, and the direction of the relationship using the sign of the test statistic.
5. Translate the results of hypothesis testing, strength of association, and pattern of association into managerial implications and recommendations where meaningful.

**ACTIVE RESEARCH**

Visit *www.loreal.com* and conduct an Internet search using a search engine and your library's online database to obtain information on the heavy users, light users, and nonusers of cosmetics.

How would you analyze the data to determine whether the heavy, light, and nonusers differ in terms of demographic characteristics?

As the marketing director for L'Oreal, what marketing strategies would you adopt to reach the heavy users, light users, and nonusers of cosmetics?

# HYPOTHESIS TESTING RELATED TO DIFFERENCES

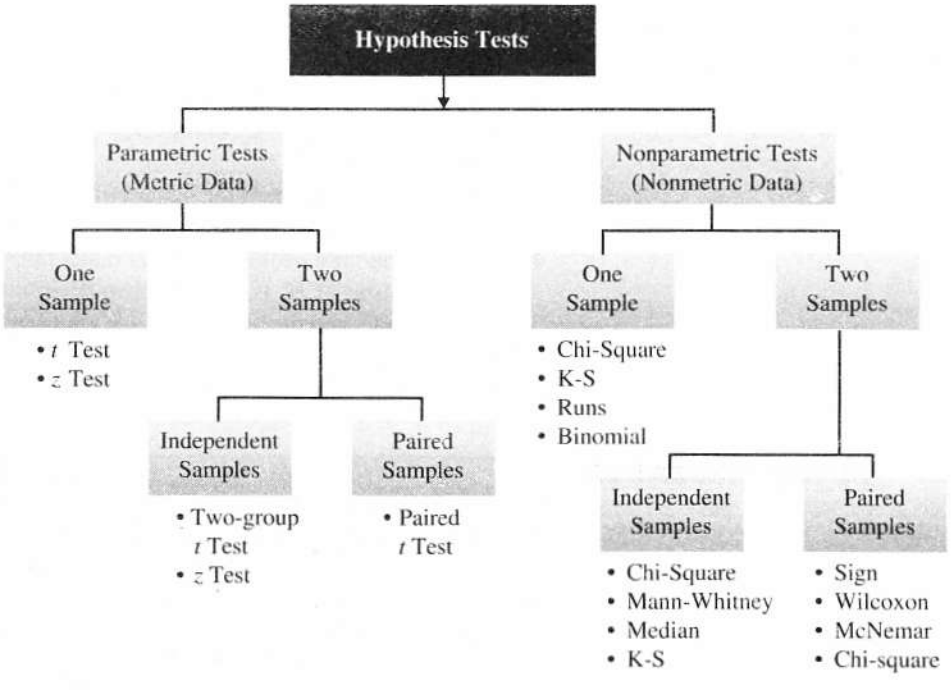
The previous section considered hypothesis testing related to associations. We now focus on hypothesis testing related to differences. A classification of hypothesis-testing procedures for examining differences is presented in Figure 15.9. Note that Figure 15.9 is consistent with the classification of univariate techniques presented in Figure 14.6. The major difference is that Figure 14.6 also accommodates more than two samples and thus deals with techniques such as one-way ANOVA and K-W ANOVA (Chapter 14), whereas Figure 15.9 is limited to no more than two samples. Also, one-sample techniques such as frequencies, which do not involve statistical testing, are not covered in Figure 15.9. Hypothesis-testing procedures can be broadly classified as parametric or nonparametric, based on the measurement scale of the variables involved. ***Parametric tests*** assume that the variables of interest are measured on at least an interval scale. ***Nonparametric tests*** assume that the variables are measured on a nominal or ordinal scale. These tests can be further classified based on whether one, two, or more samples are involved. As explained in Chapter 14, the number of samples is determined based on how the data are treated for the purpose of analysis, not based on how the data were collected. The samples are *independent* if they are drawn randomly from different populations. For the purpose of analysis, data pertaining to different groups of respondents, for example, males and females, are generally treated as independent samples. On the other hand, the samples are *paired* when the data for the two samples relate to the same group of respondents.

***parametric tests***
Hypothesis-testing procedures that assume that the variables of interest are measured on at least an interval scale.

***nonparametric tests***
Hypothesis-testing procedures that assume that the variables are measured on a nominal or ordinal scale.

The most popular parametric test is the $t$ test, conducted for examining hypotheses about means. The $t$ test could be conducted on the mean of one sample or two samples of observations. In the case of two samples, the samples could be independent or paired. The $z$ test can be used for one sample or two independent samples as well. Nonparametric tests based on observations drawn from one sample include the Kolmogorov-Smirnov test, the chi-square test, the runs test, and the binomial test. In case of two independent samples, the Mann-Whitney $U$ test, the median test, and the Kolmogorov-Smirnov two-sample test are used for examining hypotheses about location. These tests are nonparametric counterparts of the two-group $t$ test. The chi-square test can be used for examining differences in proportions. For paired samples, nonparametric tests include the Wilcoxon matched-pairs signed-ranks test and the sign

*Figure 15.9*
Hypothesis Tests Related to Differences



test. These tests are the counterparts of the paired $t$ test. Alternatively, the chi-square test can be used for binary variables. Parametric as well as nonparametric tests are also available for evaluating hypotheses relating to more than two samples. These tests are considered in later chapters.

# PARAMETRIC TESTS

***t* test**
A univariate hypothesis test using the $t$ distribution, which is used when the standard deviation is unknown and the sample size is small.

***t* statistic**
A statistic that assumes that the variable has a symmetric bell-shaped distribution, the mean is known (or assumed to be known), and the population variance is estimated from the sample.

***t* distribution**
A symmetric bell-shaped distribution that is useful for small sample ($n < 30$) testing.

Parametric tests provide inferences for making statements about the means of parent populations. A ***t* test** is commonly used for this purpose. This test is based on the Student's $t$ statistic. The ***t* statistic** assumes that the variable is normally distributed and the mean is known (or assumed to be known), and the population variance is estimated from the sample. Assume that the random variable $X$ is normally distributed, with mean $\mu$ and unknown population variance $\sigma^2$, which is estimated by the sample variance $s^2$. Recall that the standard deviation of the sample mean, $\overline{X}$, is estimated as $s_{\overline{x}} = s/\sqrt{n}$. Then $t = (\overline{X} - \mu)/s_{\overline{X}}$ is $t$ distributed with $n - 1$ degrees of freedom.

The ***t* distribution** is similar to the normal distribution in appearance. Both distributions are bell shaped and symmetric. However, as compared to the normal distribution, the $t$ distribution has more area in the tails and less in the center. This is because population variance $\sigma^2$ is unknown and is estimated by the sample variance $s^2$. Given the uncertainty in the value of $s^2$, the observed values of $t$ are more variable than those of $z$. Thus, we must go a larger number of standard deviations from 0 to encompass a certain percentage of values from the $t$ distribution than is the case with the normal distribution. Yet, as the number of degrees of freedom increases, the $t$ distribution approaches the normal distribution. In fact, for large samples of 120 or more, the $t$ distribution and the normal distribution are virtually indistinguishable. Table 4 in the Statistical Appendix shows selected percentiles of the $t$ distribution. Although normality is assumed, the $t$ test is quite robust to departures from normality.

The procedure for hypothesis testing, for the special case when the $t$ statistic is used, is as follows.

1. Formulate the null ($H_0$) and the alternative ($H_1$) hypotheses.
2. Select the appropriate formula for the $t$ statistic.
3. Select a significance level, $\alpha$, for testing $H_0$. Typically, the 0.05 level is selected.[13]
4. Take one or two samples and compute the mean and standard deviation for each sample.
5. Calculate the $t$ statistic assuming $H_0$ is true.
6. Calculate the degrees of freedom and estimate the probability of getting a more extreme value of the statistic from Table 4. (Alternatively, calculate the critical value of the $t$ statistic.)
7. If the probability computed in step 6 is smaller than the significance level selected in step 3, reject $H_0$. If the probability is larger, do not reject $H_0$. (Alternatively, if the value of the calculated $t$ statistic in step 5 is larger than the critical value determined in step 6, reject $H_0$. If the calculated value is smaller than the critical value, do not reject $H_0$.) Failure to reject $H_0$ does not necessarily imply that $H_0$ is true. It only means that the true state is not significantly different from that assumed by $H_0$.[14]
8. Express the conclusion reached by the $t$ test in terms of the marketing research problem.

## One Sample

In marketing research, the researcher is often interested in making statements about a single variable against a known or given standard. Examples of such statements include: The market share for a new product will exceed 15 percent; at least 65 percent of customers will like a new package design; 80 percent of dealers will prefer the new pricing policy. These statements can be translated to null hypotheses that can be tested using a one-sample test, such as the $t$ test or the $z$ test. In the case of a $t$ test for a single mean, the researcher is interested in testing whether the population mean conforms to a given hypothesis ($H_0$). For the data in Table 15.1, suppose we wanted to test the hypothesis that


**SPSS Output File**

the mean familiarity rating exceeds 4.0, the neutral value on a 7-point scale. A significance level of $\alpha = 0.05$ is selected. The hypotheses may be formulated as:

$$H_0: \mu \leq 4.0$$
$$H_1: \mu > 4.0$$
$$t = \frac{(\bar{X} - \mu)}{s_{\bar{X}}}$$
$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$
$$s_{\bar{X}} = 1.579/\sqrt{29} = 1.579/5.385 = 0.293$$
$$t = (4.724 - 4.0)/0.293 = 0.724/0.293 = 2.471$$

The degrees of freedom for the $t$ statistic to test the hypothesis about one mean are $n - 1$. In this case, $n - 1 = 29 - 1$ or 28. From Table 4 in the Statistical Appendix, the probability of getting a more extreme value than 2.471 is less than 0.05. (Alternatively, the critical $t$ value for 28 degrees of freedom and a significance level of 0.05 is 1.7011, which is less than the calculated value.) Hence, the null hypothesis is rejected. The familiarity level does exceed 4.0.

***z* test**
A univariate hypothesis test using the standard normal distribution.

Note that if the population standard deviation was assumed to be known as 1.5, rather than estimated from the sample, a ***z* test** would be appropriate. In this case, the value of the $z$ statistic would be:

$$z = (\bar{X} - \mu)/\sigma_{\bar{X}}$$

where

$$\sigma_{\bar{X}} = 1.5/\sqrt{29} = 1.5/5.385 = 0.279$$

and

$$z = (4.724 - 4.0)/0.279 = 0.724/0.279 = 2.595$$

From Table 2 in the Statistical Appendix, the probability of getting a more extreme value of $z$ than 2.595 is less than 0.05. (Alternatively, the critical $z$ value for a one-tailed test and a significance level of 0.05 is 1.645, which is less than the calculated value.) Therefore, the null hypothesis is rejected, reaching the same conclusion arrived at earlier by the $t$ test.

The procedure for testing a null hypothesis with respect to a proportion was illustrated earlier in this chapter when we introduced hypothesis testing.

## Two Independent Samples

Several hypotheses in marketing relate to parameters from two different populations: for example, the users and nonusers of a brand differ in terms of their perceptions of the brand, the high-income consumers spend more on entertainment than low-income consumers, or the proportion of brand-loyal users in segment I is more than the proportion in segment II. Samples drawn randomly from different populations are termed ***independent samples.*** As in the case for one sample, the hypotheses could relate to means or proportions.

***independent samples***
Two samples that are not experimentally related. The measurement of one sample has no effect on the values of the second sample.

***Means.*** In the case of means for two independent samples, the hypotheses take the following form.

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

The two populations are sampled and the means and variances computed based on samples of sizes $n_1$ and $n_2$. If both populations are found to have the same variance, a pooled variance estimate is computed from the two sample variances as follows:

$$s^2 = \frac{\sum_{i=1}^{n_1} (X_{i1} - \bar{X}_1)^2 + \sum_{i=1}^{n_2} (X_{i2} - \bar{X}_2)^2}{n_1 + n_2 - 2}$$

or

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

The standard deviation of the test statistic can be estimated as:

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

The appropriate value of $t$ can be calculated as:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{X}_1 - \bar{X}_2}}$$

The degrees of freedom in this case are $(n_1 + n_2 - 2)$.

If the two populations have unequal variances, an exact $t$ cannot be computed for the difference in sample means. Instead, an approximation to $t$ is computed. The number of degrees of freedom in this case is usually not an integer, but a reasonably accurate probability can be obtained by rounding to the nearest integer.[15]

***F test***
A statistical test of the equality of the variances of two populations.

An ***F test*** of sample variance may be performed if it is not known whether the two populations have equal variance. In this case the hypotheses are:

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_1: \sigma_1^2 \neq \sigma_2^2$$

***F statistic***
The $F$ statistic is computed as the ratio of two sample variances.

The ***F statistic*** is computed from the sample variances as follows:

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

where

$n_1$ = size of sample 1
$n_2$ = size of sample 2
$n_1 - 1$ = degrees of freedom for sample 1
$n_2 - 1$ = degrees of freedom for sample 2
$s_1^2$ = sample variance for sample 1
$s_2^2$ = sample variance for sample 2

***F distribution***
A frequency distribution that depends upon two sets of degrees of freedom—the degrees of freedom in the numerator and the degrees of freedom in the denominator.

As can be seen, the critical value of the ***F distribution*** depends upon two sets of degrees of freedom—those in the numerator and those in the denominator. The critical values of $F$ for various degrees of freedom for the numerator and denominator are given in Table 5 of the Statistical Appendix. If the probability of $F$ is greater than the significance level $\alpha$, $H_0$ is not rejected, and $t$ based on the pooled variance estimate can be used. On the other hand, if the probability of $F$ is less than or equal to $\alpha$, $H_0$ is rejected and $t$ based on a separate variance estimate is used.

Using the data of Table 15.1, suppose we wanted to determine whether Internet usage was different for males as compared to females. A two-independent-samples $t$ test was conducted. The results are presented in Table 15.14. Note that the $F$ test of sample variances has a probability that is less than 0.05. Accordingly, $H_0$ is rejected, and the $t$ test based on the "equal variances not assumed" should be used. The $t$ value is $-4.492$ and, with 18.014 degrees of freedom, this gives a probability of 0.000, which is less than the significance level of 0.05. Therefore, the null hypothesis of equal means is rejected. Because the mean usage for males (sex = 1) is 9.333 and that for females (sex = 2) is 3.867, males use the Internet to a significantly greater extent than females. We also show the $t$ test assuming equal variances because most computer programs automatically conduct the $t$ test both ways. Instead of the small sample of 30, if this were a large and representative sample, there are profound implications for Internet service providers such as AOL, EarthLink, and the various telephone (e.g., Verizon) and cable (e.g., Comcast)

SPSS Output File

**TABLE 15.14**

Two-Independent-Samples *t* Test

| Summary Statistics | | | |
|---|---|---|---|
| | Number of Cases | Mean | Standard Error Mean |
| Male | 15 | 9.333 | 1.137 |
| Female | 15 | 3.867 | 0.435 |

| *F* Test for Equality of Variances | |
|---|---|
| *F* Value | 2-Tail Probability |
| 15.507 | 0.000 |

| *t* Test | | | | | |
|---|---|---|---|---|---|
| Equal Variances Assumed | | | Equal Variances Not Assumed | | |
| *t* Value | Degrees of Freedom | 2-Tail Probability | *t* Value | Degrees of Freedom | 2-Tail Probability |
| −4.492 | 28 | 0.000 | −4.492 | 18.014 | 0.000 |

companies. In order to target the heavy Internet users, these companies should focus on males. Thus, more advertising dollars should be spent on magazines that cater to male audiences than those that target females.

## REAL RESEARCH

### *Stores Seek to Suit Elderly to a "t"*

A study based on a national sample of 789 respondents who were age 65 or older attempted to determine the effect that lack of mobility has on patronage behavior. A major research question related to the differences in the physical requirements of dependent and self-reliant elderly persons. That is, did the two groups require different things to get to the store or after they arrived at the store? A more detailed analysis of the physical requirements conducted by two-independent-sample *t* tests (shown in the accompanying table) indicated that dependent elderly persons are more likely to look for stores that offer home delivery and phone orders, and stores to which they have accessible transportation. They are also more likely to look for a variety of stores located close together. Retailers, now more than ever, are realizing the sales potential in the elderly market. With the Baby Boomer generation nearing retirement in 2008, stores such as Wal-Mart, Coldwater Creek, and Williams-Sonoma see "the icing on the cake." The elderly shoppers are more likely to spend more money and become patrons of a store. However, to attract them, stores should offer home delivery and phone orders, and arrange accessible transportation.[16]

Differences in Physical Requirements Between Dependent and Self-Reliant Elderly

| | *Mean*[a] | | |
|---|---|---|---|
| ***Physical Requirement Items*** | ***Self-Reliant*** | ***Dependent*** | **t *Test Probability*** |
| Delivery to home | 1.787 | 2.000 | 0.023 |
| Phone-in order | 2.030 | 2.335 | 0.003 |
| Transportation to store | 2.188 | 3.098 | 0.000 |
| Convenient parking | 4.001 | 4.095 | 0.305 |
| Location close to home | 3.177 | 3.325 | 0.137 |
| Variety of stores close together | 3.456 | 3.681 | 0.023 |

[a]Measured on a 5-point scale from not important (1) to very important (5). ■

In this example, we tested the difference between means. A similar test is available for testing the difference between proportions for two independent samples.

***Proportions.*** The case involving proportions for two independent samples is also illustrated using the data of Table 15.1, which gives the number of males and females who

use the Internet for shopping. Is the proportion of respondents using the Internet for shopping the same for males and females? The null and alternative hypotheses are:

$$H_0: \pi_1 = \pi_2$$
$$H_1: \pi_1 \neq \pi_2$$

A $z$ test is used as in testing the proportion for one sample. However, in this case the test statistic is given by:

$$z = \frac{P_1 - P_2}{s_{P_1 - P_2}}$$

In the test statistic, the numerator is the difference between the proportions in the two samples, $P_1$ and $P_2$. The denominator is the standard error of the difference in the two proportions and is given by

$$s_{P_1 - P_2} = \sqrt{P(1 - P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where

$$P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$$

A significance level of $\alpha = 0.05$ is selected. Given the data of Table 15.1, the test statistic can be calculated as:

$$P_1 - P_2 = (11/15) - (6/15)$$
$$= 0.733 - 0.400 = 0.333$$
$$P = (15 \times 0.733 + 15 \times 0.4)/(15 + 15) = 0.567$$
$$s_{P_1 - P_2} = \sqrt{0.567 \times 0.433\left(\frac{1}{15} + \frac{1}{15}\right)} = 0.181$$
$$z = 0.333/0.181 = 1.84$$

Given a two-tail test, the area to the right of the critical value is $\alpha/2$ or 0.025. Hence, the critical value of the test statistic is 1.96. Because the calculated value is less than the critical value, the null hypothesis cannot be rejected. Thus, the proportion of users (0.733) for males and (0.400) for females is not significantly different for the two samples. Note that although the difference is substantial, it is not statistically significant due to the small sample sizes (15 in each group).

## Paired Samples

***paired samples***
In hypothesis testing, the observations are paired so that the two sets of observations relate to the same respondents.

***paired samples t test***
A test for differences in the means of paired samples.

In many marketing research applications, the observations for the two groups are not selected from independent samples. Rather, the observations relate to ***paired samples*** in that the two sets of observations relate to the same respondents. A sample of respondents may rate two competing brands, indicate the relative importance of two attributes of a product, or evaluate a brand at two different times. The difference in these cases is examined by a ***paired samples t test.*** To compute $t$ for paired samples, the paired difference variable, denoted by $D$, is formed and its mean and variance calculated. Then the $t$ statistic is computed. The degrees of freedom are $n - 1$, where $n$ is the number of pairs. The relevant formulas are:

$$H_0: \mu_D = 0$$
$$H_1: \mu_D \neq 0$$

$$t_{n-1} = \frac{\overline{D} - \mu_D}{\frac{s_D}{\sqrt{n}}}$$

**TABLE 15.15**

**Paired Samples *t* Test**

| Variable | Number of Cases | Mean | Standard Deviation | Standard Error |
|---|---|---|---|---|
| Internet Attitude | 30 | 5.167 | 1.234 | 0.225 |
| Technology Attitude | 30 | 4.100 | 1.398 | 0.255 |

Difference = Internet – Technology

| Difference Mean | Standard Deviation | Standard Error | Correlation | 2-Tail Probability | *t* Value | Degrees of Freedom | 2-Tail Probability |
|---|---|---|---|---|---|---|---|
| 1.067 | 0.828 | 0.1511 | 0.809 | 0.000 | 7.059 | 29 | 0.000 |


**SPSS Output File**

where

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n}$$

$$s_D = \sqrt{\frac{\sum_{i=1}^{n} (D_i - \overline{D})^2}{n-1}}$$

$$S_{\overline{D}} = \frac{S_D}{\sqrt{n}}$$

In the Internet usage example (Table 15.1), a paired *t* test could be used to determine if the respondents differed in their attitude toward the Internet and attitude toward technology. The resulting output is shown in Table 15.15. The mean attitude toward the Internet is 5.167 and that toward technology is 4.10. The mean difference between the variables is 1.067, with a standard deviation of 0.828 and a standard error of 0.1511. This results in a *t* value of (1.067/0.1511) 7.06, with 30 − 1 = 29 degrees of freedom and a probability of less than 0.001. Therefore, the respondents have a more favorable attitude toward the Internet as compared to technology in general. An implication, if this were a large and representative sample, would be that Internet service providers should not hesitate to market their services to consumers who do not have a very positive attitude toward technology and do not consider themselves to be technologically savvy. Another application is provided in the context of determining the relative effectiveness of 15-second versus 30-second television commercials.

## REAL RESEARCH

### *Seconds Count*

A survey of 83 media directors of the largest Canadian advertising agencies was conducted to determine the relative effectiveness of 15-second versus 30-second commercial advertisements. Using a 5-point rating scale (1 being excellent and 5 being poor), 15- and 30-second commercials were rated by each respondent for brand awareness, main idea recall, persuasion, and ability to tell an emotional story. The accompanying table indicates that 30-second commercials were rated more favorably on all the dimensions. Paired *t* tests indicated that these differences were significant, and the 15-second commercials were evaluated as less effective. Thus, 15-second commercials may not be the answer marketers are looking for. Actually, today, the problem may not be how effective television commercials are, but if the consumers actually will be watching the commercials. One in five users never watched a commercial in 2005, and there is a threat that this number will increase in the future. Heavy advertisers such as General Motors will have to come up with more effective and creative ways to show their commercials.[17]

**ACTIVE RESEARCH**

Visit *www.reebok.com* and conduct an Internet search using a search engine and your library's online database to obtain information on the factors consumers use to evaluate competing brands of athletic shoes.

As the marketing director for Reebok, how would you improve the image and competitive positioning of your brand?

The users and nonusers of Reebok evaluated the brand on five factors using Likert-type scales. How would you analyze these data using two independent samples and paired samples $t$ tests?

Mean Rating of 15- and 30-Second Commercials on the Four Communication Variables

| *Brand Awareness* | | *Main Idea Recall* | | *Persuasion* | | *Ability to Tell Emotional Story* | |
|---|---|---|---|---|---|---|---|
| *15* | *30* | *15* | *30* | *15* | *30* | *15* | *30* |
| 2.5 | 1.9 | 2.7 | 2.0 | 3.7 | 2.1 | 4.3 | 1.9 ■ |

The difference in proportions for paired samples can be tested by using the McNemar test or the chi-square test, as explained in the following section on nonparametric tests.

# NONPARAMETRIC TESTS

Nonparametric tests are used when the independent variables are nonmetric. Like parametric tests, nonparametric tests are available for testing variables from one sample, two independent samples, or two related samples.

## One Sample

***Kolmogorov-Smirnov one-sample test***
A one-sample nonparametric goodness-of-fit test that compares the cumulative distribution function for a variable with a specified distribution.

Sometimes the researcher wants to test whether the observations for a particular variable could reasonably have come from a particular distribution, such as the normal, uniform, or Poisson distribution. Knowledge of the distribution is necessary for finding probabilities corresponding to known values of the variable or variable values corresponding to known probabilities (see Appendix 12A). The ***Kolmogorov-Smirnov (K-S) one-sample test*** is one such goodness-of-fit test. The K-S compares the cumulative distribution function for a variable with a specified distribution. $A_i$ denotes the cumulative relative frequency for each category of the theoretical (assumed) distribution, and $O_i$ the comparable value of the sample frequency. The K-S test is based on the maximum value of the absolute difference between $A_i$ and $O_i$. The test statistic is

$$K = \text{Max} \mid A_i - O_i \mid$$

The decision to reject the null hypothesis is based on the value of $K$. The larger the $K$ is, the more confidence we have that $H_0$ is false. For $\alpha = 0.05$, the critical value of $K$ for large samples (over 35) is given by $1.36/\sqrt{n}$ .[18] Alternatively, $K$ can be transformed into a normally distributed $z$ statistic and its associated probability determined.

In the context of the Internet usage example, suppose we wanted to test whether the distribution of Internet usage was normal. A K-S one-sample test is conducted, yielding the data shown in Table 15.16. The largest absolute difference between the observed and normal distribution was $K = 0.222$. Although our sample size is only 30 (less than 35), we can use the approximate formula and the critical value for $K$ is $1.36/\sqrt{30} = 0.248$ Because the calculated value of $K$ is smaller than the critical value, the null hypothesis cannot be rejected. Alternatively, Table 15.16 indicates that the probability of observing a $K$ value of 0.222, as determined by the normalized $z$ statistic, is 0.103. Because this is more than the significance level of 0.05, the null hypothesis cannot be rejected, leading to the same conclusion. Hence, the distribution of Internet


**SPSS Output File**

**TABLE 15.16**

**K-S One-Sample Test for Normality for Internet Usage**

| | | | | |
|---|---|---|---|---|
| TEST DISTRIBUTION—NORMAL | | | | |
| Mean: | 6.600 | | | |
| Standard Deviation: | 4.296 | | | |
| Cases: | 30 | | | |
| | MOST EXTREME DIFFERENCES | | | |
| ABSOLUTE | POSITIVE | NEGATIVE | K-S *z* | 2-TAILED *p* |
| 0.222 | 0.222 | −0.142 | 1.217 | 0.103 |

usage does not deviate significantly from the normal distribution. The implication is that we are safe in using statistical tests (e.g., the *z* test) and procedures that assume the normality of this variable.

As mentioned earlier, the chi-square test can also be performed on a single variable from one sample. In this context, the chi-square serves as a goodness-of-fit test. It tests whether a significant difference exists between the observed number of cases in each category and the expected number. Other one-sample nonparametric tests include the runs test and the binomial test. The ***runs test*** is a test of randomness for the dichotomous variables. This test is conducted by determining whether the order or sequence in which observations are obtained is random. The ***binomial test*** is also a goodness-of-fit test for dichotomous variables. It tests the goodness of fit of the observed number of observations in each category to the number expected under a specified binomial distribution. For more information on these tests, refer to standard statistical literature.[19]

***runs test***
A test of randomness for a dichotomous variable.

***binomial test***
A goodness-of-fit statistical test for dichotomous variables. It tests the goodness of fit of the observed number of observations in each category to the number expected under a specified binomial distribution.

## Two Independent Samples

When the difference in the location of two populations is to be compared based on observations from two independent samples, and the variable is measured on an ordinal scale, the ***Mann-Whitney U test*** can be used.[20] This test corresponds to the two-independent-sample *t* test for interval scale variables, when the variances of the two populations are assumed equal.

***Mann-Whitney U test***
A statistical test for a variable measured on an ordinal scale, comparing the difference in the location of two populations based on observations from two independent samples.

In the Mann-Whitney *U* test, the two samples are combined and the cases are ranked in order of increasing size. The test statistic, *U*, is computed as the number of times a score from sample 1 or group 1 precedes a score from group 2. If the samples are from the same population, the distribution of scores from the two groups in the rank list should be random. An extreme value of *U* would indicate a nonrandom pattern, pointing to the inequality of the two groups. For samples of less than 30, the exact significance level for *U* is computed. For larger samples, *U* is transformed into a normally distributed *z* statistic. This *z* can be corrected for ties within ranks.

We examine again the difference in the Internet usage of males and females. This time, though, the Mann-Whitney *U* test is used. The results are given in Table 15.17. Again, a significant difference is found between the two groups, corroborating the results of the two-independent-samples *t* test reported earlier. Because the ranks are assigned from the smallest observation to the largest, the higher mean rank (20.93) of males indicates that they use the Internet to a greater extent than females (mean rank = 10.07).

Researchers often wish to test for a significant difference in proportions obtained from two independent samples. As an alternative to the parametric *z* test considered earlier, one could also use the cross-tabulation procedure to conduct a chi-square test.[21] In this case, we will have a 2 × 2 table. One variable will be used to denote the sample and will assume the value 1 for sample 1 and the value of 2 for sample 2. The other variable will be the binary variable of interest.


**SPSS Output File**

**TABLE 15.17**

**Mann-Whitney *U* Test**

MANN-WHITNEY *U*—WILCOXON RANK SUM *W* TEST
INTERNET USAGE BY SEX

| SEX | MEAN RANK | CASES |
|---|---|---|
| Male | 20.93 | 15 |
| Female | 10.07 | 15 |
| Total | | 30 |

| *U* | *W* | *z* | CORRECTED FOR TIES 2-TAILED *P* |
|---|---|---|---|
| 31.000 | 151.000 | −3.406 | 0.001 |

Note
*U* = Mann-Whitney test statistic
*W* = Wilcoxon *W* Statistic
*z* = *U* transformed into a normally distributed *z* statistic

***two-sample median test***
Nonparametric test statistic that determines whether two groups are drawn from populations with the same median. This test is not as powerful as the Mann-Whitney *U*.

***Kolmogorov-Smirnov two-sample test***
Nonparametric test statistic that determines whether two distributions are the same. It takes into account any differences in the two distributions, including median, dispersion, and skewness.

Two other independent-samples nonparametric tests are the median test and Kolmogorov-Smirnov test. The ***two-sample median test*** determines whether the two groups are drawn from populations with the same median. It is not as powerful as the Mann-Whitney *U* test because it merely uses the location of each observation relative to the median, and not the rank of each observation. The ***Kolmogorov-Smirnov two-sample test*** examines whether the two distributions are the same. It takes into account any differences between the two distributions, including the median, dispersion, and skewness, as illustrated by the following example.

## REAL RESEARCH

### *Directors Change Direction*

How do marketing research directors and users in Fortune 500 manufacturing firms perceive the role of marketing research in initiating changes in marketing strategy formulation? It was found that the marketing research directors were more strongly in favor of initiating changes in strategy and less in favor of holding back than were users of marketing research. The percentage of responses to one of the items, "Initiate change in the marketing strategy of the firm whenever possible," are given in the following table. Using the Kolmogorov-Smirnov (K-S) test, these differences of role definition were statistically significant at the 0.05 level, as shown in the table.

The users of marketing research had become even more reluctant to initiate marketing strategy changes during the uncertain economy of 2005. In today's business climate, however, the reluctance of these marketing research users must be overcome to help gain a better understanding of the buyer's power. Thus, marketing research firms should devote considerable effort to convincing the users (generally marketing managers) of the value of marketing research.[22]

The Role of Marketing Research in Strategy Formulation

| | | *Responses (%)* | | | | |
|---|---|---|---|---|---|---|
| ***Sample*** | ***n*** | ***Absolutely Must*** | ***Preferably Should*** | ***May or May Not*** | ***Preferably Should Not*** | ***Absolutely Must Not*** |
| *D* | 77 | 7 | 26 | 43 | 19 | 5 |
| *U* | 68 | 2 | 15 | 32 | 35 | 16 |

K-S significance = 0.05
*D = directors, U = users ■

SPSS Output File

**TABLE 15.18**

**Wilcoxon Matched-Pairs Signed-Rank Test**

| | INTERNET WITH TECHNOLOGY | |
|---|---|---|
| (TECHNOLOGY—INTERNET) | CASES | MEAN RANK |
| − Ranks | 23 | 12.72 |
| + Ranks | 1 | 7.50 |
| Ties | 6 | |
| Total | 30 | |
| $z = -4.207$ | | 2-tailed $p = 0.0000$ |

In this example, the marketing research directors and users comprised two independent samples. However, the samples are not always independent. In the case of paired samples, a different set of tests should be used.

## Paired Samples

***Wilcoxon matched-pairs signed-ranks test***
A nonparametric test that analyzes the differences between the paired observations, taking into account the magnitude of the differences.

An important nonparametric test for examining differences in the location of two populations based on paired observations is the ***Wilcoxon matched-pairs signed-ranks test.*** This test analyzes the differences between the paired observations, taking into account the magnitude of the differences. It computes the differences between the pairs of variables and ranks the absolute differences. The next step is to sum the positive and negative ranks. The test statistic, $z$, is computed from the positive and negative rank sums. Under the null hypothesis of no difference, $z$ is a standard normal variate with mean 0 and variance 1 for large samples. This test corresponds to the paired $t$ test considered earlier.[23]

The example considered for the paired $t$ test, whether the respondents differed in terms of attitude toward the Internet and attitude toward technology, is considered again. Suppose we assume that both these variables are measured on ordinal rather than interval scales. Accordingly, we use the Wilcoxon test. The results are shown in Table 15.18. Again, a significant difference is found in the variables, and the results are in accordance with the conclusion reached by the paired $t$ test. There are 23 negative differences (attitude toward technology is less favorable than attitude toward Internet). The mean rank of these negative differences is 12.72. On the other hand, there is only one positive difference (attitude toward technology is more favorable than attitude toward Internet). The mean rank of this difference is 7.50. There are six ties, or observations with the same value for both variables. These numbers indicate that the attitude toward the Internet is more favorable than toward technology. Furthermore, the probability associated with the $z$ statistic is less than 0.05, indicating that the difference is indeed significant.

## DECISION RESEARCH

### *General Mills' Harmony: Helping Women Achieve Nutritional Harmony*

#### The Situation

Stephen W. Sanger, CEO of General Mills, is constantly being faced with the challenge of how to keep up with the changing tastes and preferences of consumers. General Mills recently did thorough focus group research on the most important consumer in grocery stores today: a woman. It is a known fact that 3 out of 4 of grocery shoppers in the United States are women, and many of these females are focusing more on their health and the nutrition value of foods. Although there are many cereals on the market with the same amount of valuable vitamins and minerals, such as Total or Kellogg's Smart Start,

General Mills decided to design a product specifically for a woman. Harmony would be its name.

Dietician Roberta Duyff claims that women do not get enough nutrients like calcium or folic acid from day to day. "It is great that a woman can now increase her intake of these important nutrients with a simple bowl of cereal for breakfast, and if you add milk, the vitamin D in milk makes the calcium in both the fortified cereal and milk itself more absorbable." This is one way in which General Mills saw an advantage—convenience for the woman. She can grab a bowl in the morning and start off her day with the nutrients she needs. Not only is the convenience of the product an incentive to market it but also the fact that it was found in the focus groups that women like to have a product of their own. In fact, according to Megan Nightingale, assistant marketing manager at General Mills, "Our research has shown that women are looking for something that's nutritious, fast, convenient, and has a good taste." Women even enjoy the calming effect of the box with its pale yellow background and blue female figure. The fact that General Mills is pursuing a single sector of the market and meeting their needs shows the benefits of this cereal. Since its nationwide release, Harmony cereal has helped increase total sales to $11.2 billion for 2005. Researching and customizing the market for women was the right way for General Mills to go.

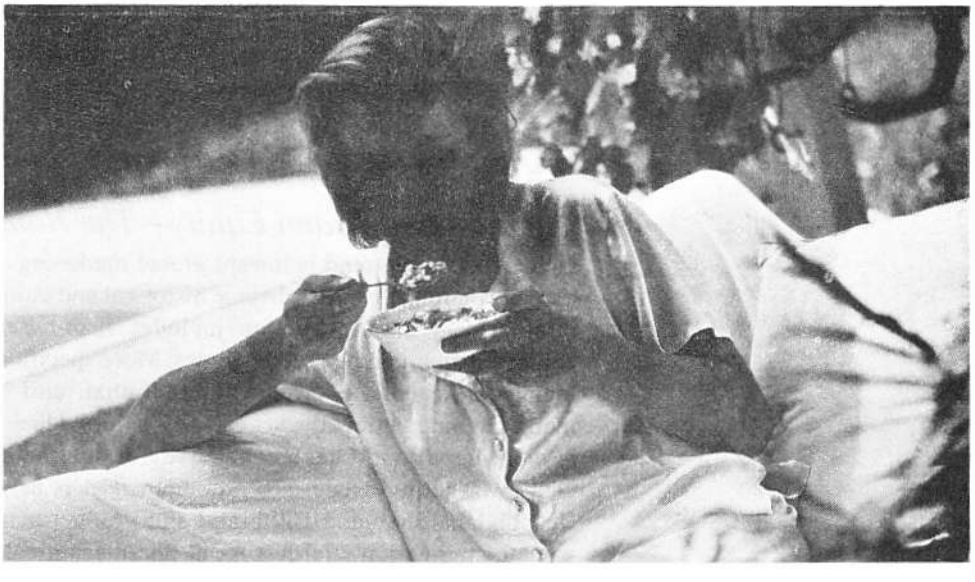
A telephone survey was conducted to determine the preference for and consumption of Harmony and the relative importance that women attached to a cereal being nutritious, fast, convenient, and good tasting.

### The Marketing Research Decision

1. What is the relative importance of the four variables (nutritious, fast, convenient, and good taste) in influencing women to buy Harmony? What type of analysis should be conducted?
2. Discuss the role of the type of data analysis you recommend in enabling Stephen W. Sanger to understand women's preference for and consumption of Harmony.

### The Marketing Management Decision

1. The advertising for Harmony should stress which of the four factors (nutritious, fast, convenient, or good taste)?
2. Discuss how the marketing management decision action that you recommend to Stephen W. Sanger is influenced by the type of data analysis that you suggested earlier and by the findings of that analysis.[24] ■



Frequency distribution, cross-tabulation, and hypothesis testing can help General Mills understand women's cereal preferences and develop appropriate advertising.

**TABLE 15.19**

**A Summary of Hypothesis Tests Related to Differences**

| SAMPLE | APPLICATION | LEVEL OF SCALING | TEST/COMMENTS |
|---|---|---|---|
| ONE SAMPLE | | | |
| One sample | Distributions | Nonmetric | K-S and chi-square for goodness of fit<br>Runs test for randomness<br>Binomial test for goodness of fit for dichotomous variables |
| One sample | Means | Metric | *t* test, if variance is unknown<br>*z* test, if variance is known |
| One Sample | Proportions | Metric | *z* test |
| TWO INDEPENDENT SAMPLES | | | |
| Two independent samples | Distributions | Nonmetric | K-S two-sample test for examining the equivalence of two distributions |
| Two independent samples | Means | Metric | Two-group *t* test<br>*F* test for equality of variances |
| Two independent samples | Proportions | Metric<br>Nonmetric | *z* test<br>Chi-square test |
| Two independent samples | Rankings/Medians | Nonmetric | Mann-Whitney *U* test is more powerful than the median test |
| PAIRED SAMPLES | | | |
| Paired samples | Means | Metric | Paired *t* test |
| Paired samples | Proportions | Nonmetric | McNemar test for binary variables<br>Chi-square test |
| Paired samples | Rankings/Medians | Nonmetric | Wilcoxon matched-pairs ranked-signs test is more powerful than the sign test |

***sign test***
A nonparametric test for examining differences in the location of two populations, based on paired observations, that compares only the signs of the differences between pairs of variables without taking into account the magnitude of the differences.

Another paired sample nonparametric test is the ***sign test.***[25] This test is not as powerful as the Wilcoxon matched-pairs signed-ranks test, as it compares only the signs of the differences between pairs of variables without taking into account the ranks. In the special case of a binary variable where the researcher wishes to test differences in proportions, the McNemar test can be used. Alternatively, the chi-square test can also be used for binary variables. The various parametric and nonparametric tests for differences are summarized in Table 15.19. The tests in Table 15.19 can be easily related to those in Figure 15.9. Table 15.19 classifies the tests in more detail because parametric tests (based on metric data) are classified separately for means and proportions. Likewise, nonparametric tests (based on nonmetric data) are classified separately for distributions and rankings/medians. The next example illustrates the use of hypothesis testing in international branding strategy, and the example after that cites the use of descriptive statistics in research on ethics.

## REAL RESEARCH

### *International Brand Equity—The Name of the Game*

In the 2000s, the trend is toward global marketing. How can marketers market a brand abroad where there exists diverse historical and cultural differences? In general, a firm's international brand structure includes firm-based characteristics, product market characteristics, and market dynamics. More specifically, according to Bob Kroll, the former president of Del Monte International, uniform packaging may be an asset to marketing internationally, yet catering to individual countries' culinary taste preferences is more important. One recent survey on international product marketing makes this clear. Marketing executives now believe it is best to think globally but act locally. Respondents included 100 brand and product managers and marketing people from some of the nation's largest food, pharmaceutical, and personal product companies.

Thirty-nine percent said that it would not be a good idea to use uniform packaging in foreign markets, whereas 38 percent were in favor of it. Those in favor of regionally targeted packaging, however, mentioned the desirability of maintaining as much brand equity and package consistency as possible from market to market. But they also believed it was necessary to tailor the package to fit the linguistic and regulatory needs of different markets. Based on this finding, a suitable research question can be: Do consumers in different countries prefer to buy global name brands with different packaging customized to suit their local needs? Based on this research question, one can frame a hypothesis that, other things being constant, standardized branding with customized packaging for a well-established name brand will result in greater market share. The hypotheses may be formulated as follows:

$H_0$: Standardized branding with customized packaging for a well-established name brand will not lead to greater market share in the international market.

$H_1$: Other factors remaining equal, standardized branding with customized packaging for a well-established name brand will lead to greater market share in the international market.

To test the null hypothesis, a well-established brand such as Colgate toothpaste, which has followed a mixed strategy, can be selected. The market share in countries with standardized branding and standardized packaging can be compared with market share in countries with standardized branding and customized packaging, after controlling for the effect of other factors. A two-independent-samples $t$ test can be used.[26] ■

**REAL RESEARCH**

### *Statistics Describe Distrust*

Descriptive statistics indicate that the public perception of ethics in business, and thus ethics in marketing, is poor. In a poll conducted by *Business Week,* 46 percent of those surveyed said that the ethical standards of business executives are only fair. A *Time* magazine survey revealed that 76 percent of Americans felt that business managers (and thus researchers) lacked ethics and that this lack contributes to the decline of moral standards in the United States. However, the general public is not alone in its disparagement of business ethics. In a Touche Ross survey of businesspeople, results showed that the general feeling was that ethics were a serious concern and media portrayal of the lack of ethics in business has not been exaggerated. However, a recent research study conducted by the Ethics Resource Center of Washington, D.C., found that 90 percent of American businesspeople expected their organization to do what is right, not just what is profitable. Twelve percent of those polled said they felt pressure to compromise their organization's ethical standards. Twenty-six percent of those polled cited the most common ethical slip in the workplace to be lying to customers, other employees, vendors, or the public, whereas 25 percent cited withholding needed information from those parties. A mere 5 percent of those polled have seen people giving or taking bribes or inappropriate gifts. Despite the fact that American businesspeople expect their organization to conduct business in an ethical manner, these studies reveal that unethical behavior remains a common practice in the workplace.[27] ■

## STATISTICAL SOFTWARE

The major programs for frequency distribution are FREQUENCIES (SPSS) and UNIVARIATE (SAS). Other programs provide only the frequency distribution (FREQ in SAS) or only some of the associated statistics (Exhibit 15.1).[28] In MINITAB, the main function is Stats>Descriptive Statistics. The output values include the mean, median, standard deviation, minimum, maximum, and quartiles. Histograms in a bar chart or graph can be produced from the Graph>Histogram selection. Several of the spreadsheets can also be used to obtain frequencies and descriptive statistics. In Excel, the Tools>Data Analysis function computes the descriptive statistics. The output produces the mean, standard error, median, mode, standard deviation, variance, kurtosis, skewness, range, minimum,

**Exhibit 15.1**
Computer Programs for Frequencies

***SPSS***
The main program in SPSS is FREQUENCIES. It produces a table of frequency counts, percentages, and cumulative percentages for the values of each variable. It gives all of the associated statistics except for the coefficient of variation. If the data are interval scaled and only the summary statistics are desired, the DESCRIPTIVES procedure can be used. All of the statistics computed by DESCRIPTIVES are available in FREQUENCIES. However, DESCRIPTIVES is more efficient because it does not sort values into a frequency table. An additional program, MEANS, computes means and standard deviations for a dependent variable over subgroups of cases defined by independent variables.

***SAS***
The main program in SAS is UNIVARIATE. In addition to providing a frequency table, this program provides all of the associated statistics. Another procedure available is FREQ. For a one-way frequency distribution, FREQ does not provide any associated statistics. If only summary statistics are desired, procedures such as MEANS, SUMMARY, and TABULATE can be used. It should be noted that FREQ is not available as an independent program in the microcomputer version.

***MINITAB***
The main function is Stats>Descriptive Statistics. The output values include the mean, median, mode, standard deviation, minimum, maximum, and quartiles. Histograms in a bar chart or graph can be produced from the Graph>Histogram selection.

***Excel***
The Tools>Data Analysis function computes the descriptive statistics. The output produces the mean, standard error, median, mode, standard deviation, variance, kurtosis, skewness, range, minimum, maximum, sum, count, and confidence level. Frequencies can be selected under the Histogram function. A histogram can be produced in bar format.

maximum, sum, count, and confidence level. Frequencies can be selected under the Histogram function. A histogram can be produced in bar format.

The major cross-tabulation programs are CROSSTABS (SPSS) and FREQ (SAS). All these programs will display the cross-classification tables and provide cell counts, row and column percentages, the chi-square test for significance, and all the measures of the strength of the association that have been discussed. In addition, the TABULATE (SAS) program can be used for obtaining cell counts and row and column percentages, although it does not provide any of the associated statistics. In MINITAB, cross-tabulations (cross tabs) and chi-square are under the Stats>Tables function. Each of these features must be selected separately under the Tables function. The Data>Pivot Table function performs cross tabs in Excel. To do additional analysis or customize data, select a different summary function such as max, min, average, or standard deviation. In addition, a custom calculation can be selected to perform values based on other cells in the data plane. ChiTest can be accessed under the Insert>Function>Statistical>ChiTest function.

The major program for conducting $t$ tests in SPSS is T-TEST. This program can be used to conduct $t$ tests on independent as well as paired samples. All the nonparametric tests that we have discussed can be conducted by using the NPAR TESTS program. In SAS, the program T TEST can be used. The nonparametric tests may be conducted by using NPAR1WAY. This program will conduct the two-independent-samples tests (Mann-Whitney, median, and K-S) as well as the Wilcoxon test for paired samples. Parametric tests available in MINITAB in descriptive stat function are $z$ test mean, $t$ test of the mean, and two-sample $t$ test. The nonparametric tests can be accessed under the Stat>Time Series function. The output includes the one-sample sign, one-sample Wilcoxon, Mann-Whitney, Kruskal-Wallis, Mood's Median test, Friedman, runs test, pairwise average, pairwise differences, and pairwise slopes. The available parametric tests in Excel and other spreadsheets include the $t$ test: paired two sample for means; $t$ test: two independent samples

assuming equal variances; *t* test: two independent samples assuming unequal variances; *z* test: two samples for means; and *F* test: two samples for variances. Nonparametric tests are not available.

## SPSS WINDOWS

The main program in SPSS is FREQUENCIES. It produces a table of frequency counts, percentages, and cumulative percentages for the values of each variable. It gives all of the associated statistics. If the data are interval scaled and only the summary statistics are desired, the DESCRIPTIVES procedure can be used. All of the statistics computed by DESCRIPTIVES are available in FREQUENCIES. However, DESCRIPTIVES is more efficient because it does not sort values into a frequency table. Moreover, the DESCRIPTIVES procedure displays summary statistics for several variables in a single table and can also calculate standardized values (*z* scores). The EXPLORE procedure produces summary statistics and graphical displays, either for all the cases or separately for groups of cases. Mean, median, variance, standard deviation, minimum, maximum, and range are some of the statistics that can be calculated.

To select these procedures, click:

Analyze>Descriptive Statistics>Frequencies
Analyze>Descriptive Statistics>Descriptives
Analyze>Descriptive Statistics>Explore

We give detailed steps for running frequencies on Familiarity with the Internet (Table 15.1) and plotting the histogram (Figure 15.1). The corresponding screen captures for these steps can be downloaded from the Web site for this book.

1. Select ANALYZE on the SPSS menu bar.
2. Click DESCRIPTIVE STATISTICS and select FREQUENCIES.
3. Move the variable "Familiarity [familiar]" to the VARIABLE(s) box.
4. Click STATISTICS.
5. Select MEAN, MEDIAN, MODE, STD. DEVIATION, VARIANCE, and RANGE.
6. Click CONTINUE.
7. Click CHARTS.
8. Click HISTOGRAMS, then click CONTINUE.
9. Click OK.

The major cross-tabulation program is CROSSTABS. This program will display the cross-classification tables and provide cell counts, row and column percentages, the chi-square test for significance, and all the measures of the strength of the association that have been discussed.

To select these procedures, click:

Analyze>Descriptive Statistics>Crosstabs

We give detailed steps for running the cross-tabulation of sex and usage of the Internet given in Table 15.3 and calculating the chi-square, contingency coefficient, and Cramer's *V*. The corresponding screen captures for these steps can be downloaded from the Web site for this book.

1. Select ANALYZE on the SPSS menu bar.
2. Click on DESCRIPTIVE STATISTICS and select CROSSTABS.
3. Move the variable "Internet Usage Group [iusagegr]" to the ROW(S) box.
4. Move the variable "Sex[sex]" to the COLUMN(S) box.
5. Click on CELLS.
6. Select OBSERVED under COUNTS and COLUMN under PERCENTAGES.
7. Click CONTINUE.
8. Click STATISTICS.

9. Click on CHI-SQUARE, PHI and CRAMER'S *V*.
10. Click CONTINUE.
11. Click OK.

The major program for conducting parametric tests in SPSS is COMPARE MEANS. This program can be used to conduct *t* tests on one sample or independent or paired samples. To select these procedures using SPSS for Windows, click:

Analyze>Compare Means>Means . . .
Analyze>Compare Means>One-Sample T Test . . .
Analyze>Compare Means>Independent-Samples T Test . . .
Analyze>Compare Means>Paired-Samples T Test . . .

We give the detailed steps for running a one-sample test on the data of Table 15.1. We wanted to test the hypothesis that the mean familiarity rating exceeds 4.0. The corresponding screen captures for these steps can be downloaded from the Web site for this book. The null hypothesis is that the mean preference for sample one before entering the theme park is 5.0.

1. Select ANALYZE from the SPSS menu bar.
2. Click COMPARE MEANS and then ONE-SAMPLE T TEST.
3. Move "Familiarity [familiar]" into the TEST VARIABLE(S) box.
4. Type "4" in the TEST VALUE box.
5. Click OK.

We give the detailed steps for running a two-independent-samples *t*-test on the data of Table 15.1 The null hypothesis is that the Internet usage for males and females is the same.

1. Select ANALYZE from the SPSS menu bar.
2. Click COMPARE MEANS and then INDEPENDENT-SAMPLES T TEST.
3. Move "Internet Usage Hrs/Week [iusage]" into the TEST VARIABLE(S) box.
4. Move "Sex[sex]" to the GROUPING VARIABLE box.
5. Click DEFINE GROUPS.
6. Type "1" in box GROUP 1 and "2" in box GROUP 2.
7. Click CONTINUE.
8. Click OK.

We give the detailed steps for running a paired samples *t*-test on the data of Table 15.1 The null hypothesis is that there is no difference in the attitude toward the Internet and attitude toward technology.

1. Select ANALYZE from the SPSS menu bar.
2. Click COMPARE MEANS and then PAIRED-SAMPLES T TEST.
3. Select "Attitude toward Internet [iattitude]" and then select "Attitude toward technology [tattitude]." Move these variables into the PAIRED VARIABLE(S) box.
4. Click OK.

The nonparametric tests discussed in this chapter can be conducted using NONPARAMETRIC TESTS. To select these procedures using SPSS for Windows, click:

Analyze>Nonparametric Tests>Chi-Square . . .
Analyze>Nonparametric Tests>Binomial . . .
Analyze>Nonparametric Tests>Runs . . .
Analyze>Nonparametric Tests>1-Sample K-S . . .
Analyze>Nonparametric Tests>2 Independent Samples . . .
Analyze>Nonparametric Tests>2 Related Samples . . .

The detailed steps for the nonparametric tests are similar to those for parametric tests and are not shown due to space constraints.

## PROJECT RESEARCH

### *Basic Data Analysis*

In the department store project, basic data analysis formed the foundation for conducting subsequent multivariate analysis. Data analysis began by obtaining a frequency distribution and descriptive statistics for each variable. In addition to identifying possible problems with the data (see Chapter 14), this information provided a good feel for the data and insights into how specific variables should be treated in subsequent analyses. For example, should some variables be treated as categorical, and, if so, how many categories should there be? Several two- and three-variable cross-tabulations were also conducted to identify associations in the data. The effects of variables with two categories on the metric dependent variables of interest were examined by means of $t$ tests and other hypothesis-testing procedures.


**SPSS Data File**

### Project Activities

Download the SPSS data file Sears Data 14 from the Web site for this book. See Chapter 14 for a description of this file.

1. Run a frequency distribution for each familiarity variable and the overall familiarity score (as computed in Chapter 14) with all accompanying descriptive statistics.
2. Recode the overall familiarity score as follows: 32 or less = 1; 33 to 37 = 2; 38 to 43 = 3; 44 to 60 = 4. Run cross tabs of the recoded overall familiarity score with demographic variables as recoded in Chapter 14. Interpret the results.
3. Test the null hypothesis that the average overall familiarity score is less than or equal to 30.
4. Do a parametric and a corresponding nonparametric test to determine whether the married and not married (recoded marital status) differ in their overall familiarity score.
5. Do a parametric and a corresponding nonparametric test to determine whether the respondents differ in their familiarity with Neiman Marcus and JCPenney. ■

## EXPERIENTIAL RESEARCH


**SPSS Data File**

Download the Dell case and questionnaire from the Web site for this book. This information is also given at the end of the book. Download the Dell SPSS data file.

1. Calculate the frequency distribution for each variable in the data file. Examine the distribution to get a feel for the data.
2. Cross-tabulate the recoded questions q4 (Overall satisfaction with Dell), q5 (Would recommend Dell), and q6 (Likelihood of choosing Dell) with the recoded demographic characteristics. Interpret the results.
3. Cross-tabulate the recoded questions on price sensitivity (q9_5per and q9_10per) with the recoded demographic characteristics. Interpret the results.
4. The mean response on which of the evaluations of Dell (q8_1 to q8_13) exceeds 5 (the midpoint of the scale)?
5. The response on which of the evaluations of Dell (q8_1 to q8_13) is normally distributed? What are the implications of your results for data analysis?
6. Are the two overall satisfaction groups derived based on the recoding of q4 as specified in Chapter 14 different in terms of each of the evaluations of Dell (q8_1 to q8_13)? How would your analysis change if the evaluations of Dell (q8_1 to q8_13) are to be treated as ordinal rather than interval scaled?
7. Are the two Likely to recommend groups derived based on the recoding of q5 as specified in Chapter 14 different in terms of each of the evaluations of Dell (q8_1 to q8_13)? How would your analysis change if the evaluations of Dell (q8_1 to q8_13) are to be treated as ordinal rather than interval scaled?

8. Are the two Likelihood of choosing Dell groups derived based on the recoding of q6 as specified in Chapter 14 different in terms of each of the evaluations of Dell (q8_1 to q8_13)? How would your analysis change if the evaluations of Dell (q8_1 to q8_13) are to be treated as ordinal rather than interval scaled?
9. Is the mean of responses to q8b_1 (Make ordering a computer system easy) and q8b_2 (Let customers order computer systems customized to their specifications) different? How would your analysis change if the evaluations of Dell (q8_1 and q8_2) are to be treated as ordinal rather than interval scaled?
10. Is the mean of responses to q8b_9 ("Bundle" its computers with appropriate software) and q8b_10 ("Bundle" its computers with Internet access) different? How would your analysis change if the evaluations of Dell (q8_9 and q8_10) are to be treated as ordinal rather than interval scaled?
11. Is the mean of responses to q8b_6 (Have computers that run programs quickly) and q8b_7 (Have high-quality computers with no technical problems) different? How would your analysis change if the evaluations of Dell (q8_6 and q8_7) are to be treated as ordinal rather than interval scaled? ■

## SUMMARY

Basic data analysis provides valuable insights and guides the rest of the data analysis as well as the interpretation of the results. A frequency distribution should be obtained for each variable in the data. This analysis produces a table of frequency counts, percentages, and cumulative percentages for all the values associated with that variable. It indicates the extent of out-of-range, missing, or extreme values. The mean, mode, and median of a frequency distribution are measures of central tendency. The variability of the distribution is described by the range, the variance or standard deviation, coefficient of variation, and interquartile range. Skewness and kurtosis provide an idea of the shape of the distribution.

Cross-tabulations are tables that reflect the joint distribution of two or more variables. In cross-tabulation, the percentages can be computed either columnwise, based on column totals, or rowwise, based on row totals. The general rule is to compute the percentages in the direction of the independent variable, across the dependent variable. Often the introduction of a third variable can provide additional insights. The chi-square statistic provides a test of the statistical significance of the observed association in a cross-tabulation. The phi coefficient, contingency coefficient, Cramer's $V$, and the lambda coefficient provide measures of the strength of association between the variables.

Parametric and nonparametric tests are available for testing hypotheses related to differences. In the parametric case, the $t$ test is used to examine hypotheses related to the population mean. Different forms of the $t$ test are suitable for testing hypotheses based on one sample, two independent samples, or paired samples. In the nonparametric case, popular one-sample tests include the Kolmogorov-Smirnov, chi-square, runs test, and the binomial test. For two independent nonparametric samples, the Mann-Whitney $U$ test, median test, and the Kolmogorov-Smirnov test can be used. For paired samples, the Wilcoxon matched-pairs signed-ranks test and the sign test are useful for examining hypotheses related to measures of location.

## KEY TERMS AND CONCEPTS

frequency distribution, *458*
measures of location, *460*
mean, *460*
mode, *460*
median, *460*
measures of variability, *461*
range, *461*
interquartile range, *461*
variance, *461*
standard deviation, *461*
coefficient of variation, *462*
skewness, *462*
kurtosis, *462*
null hypothesis, *464*
alternative hypothesis, *464*
one-tailed test, *465*
two-tailed test, *465*
test statistic, *465*
type I error, *466*
level of significance, *466*
type II error, *466*

power of a test, ***466***
cross-tabulation, ***468***
contingency tables, ***469***
chi-square statistic, ***474***
chi-square distribution, ***474***
phi coefficient, ***475***
contingency coefficient (C), ***476***
Cramer's *V*, ***476***
asymmetric lambda, ***476***
symmetric lambda, ***477***
tau *b*, ***477***
tau *c*, ***477***
gamma, ***477***
parametric tests, ***478***
nonparametric tests, ***478***
*t* test, ***479***
*t* statistic, ***479***
*t* distribution, ***479***
*z* test, ***480***
independent samples, ***480***
*F* test, ***481***
*F* statistic, ***481***
*F* distribution, ***481***
paired samples, ***483***
paired samples *t* test, ***483***
Kolmogorov-Smirnov (K-S) one-sample test, ***485***
runs test, ***486***
binomial test, ***486***
Mann-Whitney *U* test, ***486***
two-sample median test, ***487***
Kolmogorov-Smirnov two-sample test, ***487***
Wilcoxon matched-pairs signed-ranks test, ***488***
sign test, ***490***

## SUGGESTED CASES, VIDEO CASES, AND HBS CASES

### *Cases*

Case 3.1 Is Celebrity Advertising Worth Celebrating?
Case 3.3 Matsushita Retargets the U.S.A.
Case 3.4 Pampers Curing Its Rash of Market Share
Case 3.5 DaimlerChrysler Seeks a New Image
Case 3.6 Cingular Wireless: A Singular Focus
Case 3.7 IBM: The World's Top Provider of Computer Hardware, Software, and Services
Case 3.8 Kimberly-Clark: Competing Through Innovation
Case 4.1 Wachovia: "Watch Ovah Ya" Finances
Case 4.2 Wendy's: History and Life After Dave Thomas
Case 4.3 Astec: Continuing to Grow
Case 4.4 Is Marketing Research the Cure for Norton Healthcare Kosair Children's Hospital's Ailments?

### *Video Cases*

Video Case 3.1 The Mayo Clinic: Staying Healthy with Marketing Research
Video Case 4.1 Subaru: "Mr. Survey" Monitors Customer Satisfaction
Video Case 4.2 Procter & Gamble: Using Marketing Research to Build Brands

## LIVE RESEARCH: CONDUCTING A MARKETING RESEARCH PROJECT

1. Each team can conduct the entire analysis, or the data analysis can be split between teams with each team conducting a different type of analysis.
2. It is helpful to run a frequency count for every variable. This gives a good feel for the data.
3. Calculate the measures of location (mean, median, mode), measures of variability (range and standard deviation), as well as the measures of shape (skewness and kurtosis) for each variable.
4. Relevant associations can be examined by conducting cross-tabulations. Procedures should be specified for categorizing interval or ratio scaled variables.
5. Differences between groups are of interest in most projects. In case of two groups, these can be examined by using independent samples t tests.
6. Often each respondent evaluates many stimuli. For example, each respondent may evaluate different brands or provide importance ratings for different attributes. In such cases, differences between pairs of stimuli may be examined using the paired samples *t* test.

# ACRONYMS

The statistics associated with frequencies may be summarized by the acronym FREQUENCIES:

**F** latness or peakedness: kurtosis
**R** ange
**E** stimate of location: mean
**Q** uotients: percentages
**U** ndulation: variance
**E** stimate of location: mode
**N** umbers or counts
**C** oefficient of variation
**I** nterquartile range
**E** stimate of location: median
**S** kewness

The salient characteristics of cross-tabulations may be summarized by the acronym C TABULATIONS:

**C** ube: chi-square, contingency coefficient, and Cramer's *V*
**T** wo by two table statistic: phi coefficient
**A** dditional insights or refinements provided by third variable
**B** ased on cell count of at least five
**U** nchanged association with third variable introduction
**L** ambda coefficient
**A** ssociation and not causation is measured
**T** wo- and three-variable cases
**I** nitial relationship may be spurious
**O** ver three variables poses problems
**N** umbers and percentages
**S** uppressed association may be revealed

# EXERCISES

## *Questions*

1. Describe the procedure for computing frequencies.
2. What measures of location are commonly computed?
3. Define the interquartile range. What does it measure?
4. What is meant by the coefficient of variation?
5. How is the relative flatness or peakedness of a distribution measured?
6. What is a skewed distribution? What does it mean?
7. What is the major difference between cross-tabulation and frequency distribution?
8. What is the general rule for computing percentages in cross-tabulation?
9. Define a spurious correlation.
10. What is meant by a suppressed association? How is it revealed?
11. Discuss the reasons for the frequent use of cross-tabulations. What are some of its limitations?
12. Present a classification of hypothesis-testing procedures.
13. Describe the general procedure for conducting a *t* test.
14. What is the major difference between parametric and nonparametric tests?
15. Which nonparametric tests are the counterparts of the two-independent-samples *t* test for parametric data?
16. Which nonparametric tests are the counterparts of the paired samples *t* test for parametric data?

## *Problems*

1. In each of the following situations, indicate the statistical analysis you would conduct and the appropriate test or test statistic that should be used.
   **a.** Consumer preferences for Camay bathing soap were obtained on an 11-point Likert scale. The same consumers were then shown a commercial about Camay. After the commercial, preferences for Camay were again measured. Has the commercial been successful in inducing a change in preferences?
   **b.** Does the preference for Camay soap follow a normal distribution?
   **c.** Respondents in a survey of 1,000 households were classified as heavy, medium, light, or nonusers of ice cream. They were also classified as being in high-, medium-, or low-income categories. Is the consumption of ice cream related to income level?
   **d.** In a survey using a representative sample of 2,000 households from the Market Facts consumer mail panel, the respondents were asked to rank 10 department stores, including Sears, in order of preference. The sample was divided into small and large households based on a median split of the household size. Does preference for shopping in Sears vary by household size?

2. The current advertising campaign for a major soft drink brand would be changed if less than 30 percent of the consumers like it.
   a. Formulate the null and alternative hypotheses.
   b. Discuss the type I and type II errors that could occur in hypothesis testing.
   c. Which statistical test would you use? Why?
   d. A random sample of 300 consumers was surveyed, and 84 respondents indicated that they liked the campaign. Should the campaign be changed? Why?

3. A major department store chain is having an end-of-season sale on refrigerators. The number of refrigerators sold during this sale at a sample of 10 stores was:

   80 110 0 40 70 80 100 50 80 30

   a. Is there evidence that an average of more than 50 refrigerators per store were sold during this sale? Use $\alpha = 0.05$.
   b. What assumption is necessary to perform this test?

## INTERNET AND COMPUTER EXERCISES

SPSS Data File

1. In a pretest, data on Nike were obtained from 45 respondents. These data are given in the following table, which gives the usage, sex, awareness, attitude, preference, intention, and loyalty toward Nike of a sample of Nike users. Usage has been coded as 1, 2, or 3, representing light, medium, or heavy users. The sex has been coded as 1 for females and 2 for males. Awareness, attitude, preference, intention, and loyalty are measured on 7-point Likert-type scales (1 = very unfavorable, 7 = very favorable). Note that five respondents have missing values that are denoted by 9.

| *Number* | *Usage* | *Sex* | *Awareness* | *Attitude* | *Preference* | *Intention* | *Loyalty* |
|---|---|---|---|---|---|---|---|
| 1 | 3 | 2 | 7 | 6 | 5 | 5 | 6 |
| 2 | 1 | 1 | 2 | 2 | 4 | 6 | 5 |
| 3 | 1 | 1 | 3 | 3 | 6 | 7 | 6 |
| 4 | 3 | 2 | 6 | 5 | 5 | 3 | 2 |
| 5 | 3 | 2 | 5 | 4 | 7 | 4 | 3 |
| 6 | 2 | 2 | 4 | 3 | 5 | 2 | 3 |
| 7 | 2 | 1 | 5 | 4 | 4 | 3 | 2 |
| 8 | 1 | 1 | 2 | 1 | 3 | 4 | 5 |
| 9 | 2 | 2 | 4 | 4 | 3 | 6 | 5 |
| 10 | 1 | 1 | 3 | 1 | 2 | 4 | 5 |
| 11 | 3 | 2 | 6 | 7 | 6 | 4 | 5 |
| 12 | 3 | 2 | 6 | 5 | 6 | 4 | 4 |
| 13 | 1 | 1 | 4 | 3 | 3 | 1 | 1 |
| 14 | 3 | 2 | 6 | 4 | 5 | 3 | 2 |
| 15 | 1 | 2 | 4 | 3 | 4 | 5 | 6 |
| 16 | 1 | 2 | 3 | 4 | 2 | 4 | 2 |
| 17 | 3 | 1 | 7 | 6 | 4 | 5 | 3 |
| 18 | 2 | 1 | 6 | 5 | 4 | 3 | 2 |
| 19 | 1 | 1 | 1 | 1 | 3 | 4 | 5 |
| 20 | 3 | 1 | 5 | 7 | 4 | 1 | 2 |
| 21 | 3 | 2 | 6 | 6 | 7 | 7 | 5 |
| 22 | 2 | 2 | 2 | 3 | 1 | 4 | 2 |
| 23 | 1 | 1 | 1 | 1 | 3 | 2 | 2 |
| 24 | 3 | 1 | 6 | 7 | 6 | 7 | 6 |
| 25 | 1 | 2 | 3 | 2 | 2 | 1 | 1 |
| 26 | 2 | 2 | 5 | 3 | 4 | 4 | 5 |
| 27 | 3 | 2 | 7 | 6 | 6 | 5 | 7 |
| 28 | 2 | 1 | 6 | 4 | 2 | 5 | 6 |
| 29 | 1 | 1 | 9 | 2 | 3 | 1 | 3 |
| 30 | 2 | 2 | 5 | 9 | 4 | 6 | 5 |
| 31 | 1 | 2 | 1 | 2 | 9 | 3 | 2 |
| 32 | 1 | 2 | 4 | 6 | 5 | 9 | 3 |
| 33 | 2 | 1 | 3 | 4 | 3 | 2 | 9 |
| 34 | 2 | 1 | 4 | 6 | 5 | 7 | 6 |
| 35 | 3 | 1 | 5 | 7 | 7 | 3 | 3 |
| 36 | 3 | 1 | 6 | 5 | 7 | 3 | 4 |
| 37 | 3 | 2 | 6 | 7 | 5 | 3 | 4 |
| 38 | 3 | 2 | 5 | 6 | 4 | 3 | 2 |
| 39 | 3 | 2 | 7 | 7 | 6 | 3 | 4 |
| 40 | 1 | 1 | 4 | 3 | 4 | 6 | 5 |
| 41 | 1 | 1 | 2 | 3 | 4 | 5 | 6 |
| 42 | 1 | 1 | 1 | 3 | 2 | 3 | 4 |
| 43 | 1 | 1 | 2 | 4 | 3 | 6 | 7 |
| 44 | 1 | 1 | 3 | 3 | 4 | 6 | 5 |
| 45 | 1 | 1 | 1 | 1 | 4 | 5 | 3 |

Analyze the Nike data to answer the following questions. In each case, formulate the null and the alternative hypotheses and conduct the appropriate statistical test(s).

   a. Obtain a frequency distribution for each of the following variables and calculate the relevant statistics: awareness, attitude, preference, intention, and loyalty toward Nike.
   b. Conduct a cross-tabulation of the usage with sex. Interpret the results.
   c. Does the awareness for Nike exceed 3.0?
   d. Do the males and females differ in their awareness for Nike? Their attitude toward Nike? Their loyalty for Nike?
   e. Do the respondents in the pretest have a higher level of awareness than loyalty?
   f. Does awareness of Nike follow a normal distribution?
   g. Is the distribution of preference for Nike normal?
   h. Assume that awareness toward Nike was measured on an ordinal scale rather than an interval scale. Do males and females differ in their awareness toward Nike?
   i. Assume that loyalty toward Nike was measured on an ordinal scale rather than an interval scale. Do males and females differ in their loyalty toward Nike?
   j. Assume that attitude and loyalty toward Nike were measured on an ordinal scale rather than an interval scale. Do the respondents have greater awareness of Nike than loyalty for Nike?

2. In a pretest, respondents were asked to express their preference for an outdoor lifestyle using a 7-point scale: 1 = not at all preferred, to 7 = greatly preferred (V1). They were also asked to indicate the importance of the following variables on a 7-point scale: 1 = not at all important, to 7 = very important.

V2 = enjoying nature
V3 = relating to the weather
V4 = living in harmony with the environment
V5 = exercising regularly
V6 = meeting other people

The sex of the respondent (V7) was coded as 1 for females and 2 for males. The location of residence (V8) was coded as: 1 = midtown/downtown, 2 = suburbs, and 3 = countryside. The data obtained are given in the following table:

| V1 | V2 | V3 | V4 | V5 | V6 | V7 | V8 |
|---|---|---|---|---|---|---|---|
| 7.00 | 3.00 | 6.00 | 4.00 | 5.00 | 2.00 | 1.00 | 1.00 |
| 1.00 | 1.00 | 1.00 | 2.00 | 1.00 | 2.00 | 1.00 | 1.00 |
| 6.00 | 2.00 | 5.00 | 4.00 | 4.00 | 5.00 | 1.00 | 1.00 |
| 4.00 | 3.00 | 4.00 | 6.00 | 3.00 | 2.00 | 1.00 | 1.00 |
| 1.00 | 2.00 | 2.00 | 3.00 | 1.00 | 2.00 | 1.00 | 1.00 |
| 6.00 | 3.00 | 5.00 | 4.00 | 6.00 | 2.00 | 1.00 | 1.00 |
| 5.00 | 3.00 | 4.00 | 3.00 | 4.00 | 5.00 | 1.00 | 1.00 |
| 6.00 | 4.00 | 5.00 | 4.00 | 5.00 | 1.00 | 1.00 | 1.00 |
| 3.00 | 3.00 | 2.00 | 2.00 | 2.00 | 2.00 | 1.00 | 1.00 |
| 2.00 | 4.00 | 2.00 | 6.00 | 2.00 | 2.00 | 1.00 | 1.00 |
| 6.00 | 4.00 | 5.00 | 3.00 | 5.00 | 5.00 | 1.00 | 2.00 |
| 2.00 | 3.00 | 1.00 | 4.00 | 2.00 | 1.00 | 1.00 | 2.00 |
| 7.00 | 2.00 | 6.00 | 4.00 | 5.00 | 6.00 | 1.00 | 2.00 |
| 4.00 | 6.00 | 4.00 | 5.00 | 3.00 | 3.00 | 1.00 | 2.00 |
| 1.00 | 3.00 | 1.00 | 2.00 | 1.00 | 4.00 | 1.00 | 2.00 |
| 6.00 | 6.00 | 6.00 | 3.00 | 4.00 | 5.00 | 2.00 | 2.00 |
| 5.00 | 5.00 | 6.00 | 4.00 | 4.00 | 6.00 | 2.00 | 2.00 |
| 7.00 | 7.00 | 4.00 | 4.00 | 7.00 | 7.00 | 2.00 | 2.00 |
| 2.00 | 6.00 | 3.00 | 7.00 | 4.00 | 3.00 | 2.00 | 2.00 |
| 3.00 | 7.00 | 3.00 | 6.00 | 4.00 | 4.00 | 2.00 | 2.00 |
| 1.00 | 5.00 | 2.00 | 6.00 | 3.00 | 3.00 | 2.00 | 3.00 |
| 5.00 | 6.00 | 4.00 | 7.00 | 5.00 | 6.00 | 2.00 | 3.00 |
| 2.00 | 4.00 | 1.00 | 5.00 | 4.00 | 4.00 | 2.00 | 3.00 |
| 4.00 | 7.00 | 4.00 | 7.00 | 4.00 | 6.00 | 2.00 | 3.00 |
| 6.00 | 7.00 | 4.00 | 2.00 | 1.00 | 7.00 | 2.00 | 3.00 |
| 3.00 | 6.00 | 4.00 | 6.00 | 4.00 | 4.00 | 2.00 | 3.00 |
| 4.00 | 7.00 | 7.00 | 4.00 | 2.00 | 5.00 | 2.00 | 3.00 |
| 3.00 | 7.00 | 2.00 | 6.00 | 4.00 | 3.00 | 2.00 | 3.00 |
| 4.00 | 6.00 | 3.00 | 7.00 | 2.00 | 7.00 | 2.00 | 3.00 |
| 5.00 | 6.00 | 2.00 | 6.00 | 7.00 | 2.00 | 2.00 | 3.00 |

Using a statistical package of your choice, please answer the following questions. In each case, formulate the null and the alternative hypotheses and conduct the appropriate statistical test(s).

**a.** Does the mean preference for an outdoor lifestyle exceed 3.0?
**b.** Does the mean importance of enjoying nature exceed 3.5?
**c.** Does the mean preference for an outdoor lifestyle differ for males and females?
**d.** Does the importance attached to V2 to V6 differ for males and females?
**e.** Do the respondents attach more importance to enjoying nature than they do to relating to the weather?
**f.** Do the respondents attach more importance to relating to the weather than they do to meeting other people?
**g.** Do the respondents attach more importance to living in harmony with the environment than they do to exercising regularly?
**h.** Does the importance attached to V2 to V6 differ for males and females if these variables are treated as ordinal rather than interval scaled?
**i.** Do the respondents attach more importance to relating to the weather than they do to meeting other people if these variables are treated as ordinal rather than interval?

**3.** Use one of the statistical packages (SPSS, SAS, MINITAB, or Excel) to conduct the following analysis for the soft drink data that you have collected as part of your fieldwork (described later).

**a.** Obtain a frequency distribution of the weekly soft drink consumption.
**b.** Obtain the summary statistics related to the weekly amount spent on soft drinks.
**c.** Conduct a cross-tabulation of the weekly consumption of soft drinks with sex of the respondent. Does your data show any association?
**d.** Do a two-independent-sample $t$ test to determine whether the weekly amount spent on soft drinks is different for males and females.
**e.** Conduct a test to determine whether there is any difference between the weekly amount spent on soft drinks and that spent on other nonalcoholic beverages. What is your conclusion?

# ACTIVITIES

## *Role Playing*

**1.** You have been hired as a marketing research analyst by a major industrial marketing company in the country. Your boss, the market research manager, is a high-powered statistician who does not believe in using rudimentary techniques such as frequency distributions, cross-tabulations, and simple $t$ tests. Convince your boss (a student in your class) of the merits of conducting these analyses.

## *Fieldwork*

**1.** Develop a questionnaire to obtain the following information from students on your campus.

**a.** Average amount per week spent on the consumption of soft drinks
**b.** Average amount per week spent on the consumption of other nonalcoholic beverages (milk, coffee, tea, fruit juices, etc.)

**c.** Frequency of weekly soft drink consumption. Measure this as a categorical variable with the following question: "How often do you consume soft drinks? (1) once a week or less often, (2) two or three times a week, (3) four to six times a week, and (4) more than six times a week."

**d.** Sex of the respondent

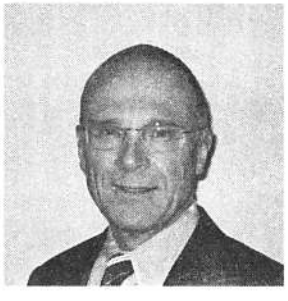
Administer this questionnaire to 40 students. Code the data and transcribe them for computer analysis. As compared to males, do females: (i) spend more on soft drinks, (ii) spend more on other nonalcoholic beverages, (iii) consume more soft drinks?

## *Group Discussion*

1. "Because cross-tabulation has certain basic limitations, this technique should not be used extensively in commercial marketing research." Discuss as a small group.
2. "Why waste time doing basic data analysis? Why not just conduct sophisticated multivariate data analysis?" Discuss.

CHAPTER 16

# Analysis of Variance and Covariance



"Analysis of variance is a straightforward way to look at differences among more than two groups of responses measured on interval or ratio scales."

*Terry Grapentine, Grapentine Company, Inc.*

## Objectives

After reading this chapter, the student should be able to:

1. Discuss the scope of the analysis of variance (ANOVA) technique and its relationship to the *t* test and regression.
2. Describe one-way analysis of variance, including decomposition of the total variation, measurement of effects, significance testing, and interpretation of results.
3. Describe *n*-way analysis of variance and the testing of the significance of the overall effect, the interaction effect, and the main effect of each factor.
4. Describe analysis of covariance and show how it accounts for the influence of uncontrolled independent variables.
5. Explain key factors pertaining to the interpretation of results with emphasis on interactions, relative importance of factors, and multiple comparisons.
6. Discuss specialized ANOVA techniques applicable to marketing such as repeated measures ANOVA, nonmetric analysis of variance, and multivariate analysis of variance (MANOVA).

## Overview

In Chapter 15, we examined tests of differences between two means or two medians. In this chapter, we discuss procedures for examining differences between more than two means or medians. These procedures are called *analysis of variance* and *analysis of covariance*. Although these procedures have traditionally been used for analyzing experimental data, they are also used for analyzing survey or observational data.

We describe the analysis of variance and covariance procedures and discuss their relationship to other techniques. Then we describe one-way analysis of variance, the simplest of these procedures, followed by *n*-way analysis of variance and analysis of covariance. Special attention is given to issues in interpretation of results as they relate to interactions, relative importance of factors, and multiple comparisons. Some specialized topics, such as repeated measures analysis of variance, nonmetric analysis of variance, and multivariate analysis of variance, are briefly discussed.

**REAL RESEARCH**

### *Analysis of Tourism Destinations*

A marketing research survey conducted by EgeBank in Istanbul, Turkey, focused on the importance of U.S. tour operators' and travel agents' perceptions of selected Mediterranean tourist destinations (Egypt, Greece, Italy, and Turkey). This study was conducted with the help of the Department of Tourism and Convention Administration at the University of Nevada–Las Vegas (*www.unlv.edu*).

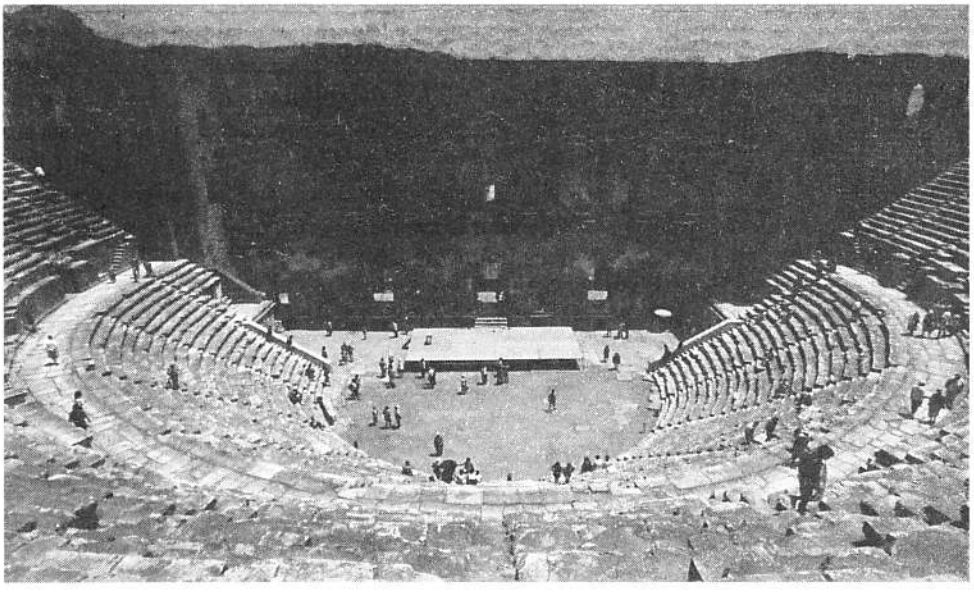
Operators/travel agents were mailed surveys based on the locations of tours, broken down as follows: Egypt (53), Greece (130), Italy (150), and Turkey (65). The survey consisted of questions on affective and perceptual/cognitive evaluations of the four destinations. The four affective questions were asked on a 7-point semantic differential scale, whereas the 14 perceptual/cognitive evaluations were measured on a 5-point Likert scale (1 = offers very little, 2 = offers somewhat little, 3 = offers neither little nor much, 4 = offers somewhat much, and 5 = offers very much). The differences in the evaluations of the four locations were examined using one-way analysis of variance (ANOVA), as seen in the following table.

Image Variations of Destinations Promoted to Tour Operators and Travel Agencies

| *Image Items* | *Turkey (n = 36)* | *Egypt (n = 29)* | *Greece (n = 37)* | *Italy (n = 34)* | *Significance* |
|---|---|---|---|---|---|
| **Affective (Scale 1–7)** | | | | | |
| Unpleasant–pleasant | 6.14 | 5.62 | 6.43 | 6.50 | 0.047[a] |
| Sleepy–arousing | 6.24 | 5.61 | 6.14 | 6.56 | 0.053 |
| Distressing–relaxing | 5.60 | 4.86 | 6.05 | 6.09 | 0.003[a] |
| Gloomy–exciting | 6.20 | 5.83 | 6.32 | 6.71 | 0.061 |
| **Perceptual (Scale 1–5)** | | | | | |
| Good value for money | 4.62 | 4.32 | 3.89 | 3.27 | 0.000[a] |
| Beautiful scenery and natural attractions | 4.50 | 4.04 | 4.53 | 4.70 | 0.011[a] |
| Good climate | 4.29 | 4.00 | 4.41 | 4.35 | 0.133 |
| Interesting cultural attractions | 4.76 | 4.79 | 4.67 | 4.79 | 0.781 |
| Suitable accommodations | 4.17 | 4.28 | 4.35 | 4.62 | 0.125 |
| Appealing local food (cuisine) | 4.44 | 3.57 | 4.19 | 4.85 | 0.000[a] |
| Great beaches and water sports | 3.91 | 3.18 | 4.27 | 3.65 | 0.001[a] |
| Quality of infrastructure | 3.49 | 2.97 | 3.68 | 4.09 | 0.000[a] |
| Personal safety | 3.83 | 3.28 | 4.19 | 4.15 | 0.000[a] |
| Interesting historical attractions environment | 4.71 | 4.86 | 4.81 | 4.82 | 0.650 |
| Unpolluted and unspoiled | 3.54 | 3.34 | 3.43 | 3.59 | 0.784 |
| Good nightlife and entertainment | 3.44 | 3.15 | 4.06 | 4.27 | 0.000[a] |
| Standard hygiene and cleanliness | 3.29 | 2.79 | 3.76 | 4.29 | 0.000[a] |
| Interesting and friendly people | 4.34 | 4.24 | 4.35 | 4.32 | 0.956 |

[a]Significant at 0.05 level

The ANOVA table shows that "unpleasant–pleasant" and "distressing–relaxing" affective factors have significant differences among the four destinations. For instance, Greece and Italy were perceived as being significantly more relaxing than Egypt. As for the perceptual factors, eight of the 14 factors were significant. Turkey was perceived as a significantly better value for money than Greece and Italy. Turkey's main strength appears to be "good value," and the country's tourism agencies should promote this in their marketing strategies. On the other hand, Turkey needs to improve the perception of its infrastructure, cleanliness, and entertainment to encourage more tour operators and travel agencies in the United States to offe travel packages to Turkey. Tourism revenues in Turkey exceeded U.S. $11 billion in 2005.[1] ■



Analysis of variance techniques can help identify affective and perceptual factors that differentiate alternative tourist destinations.

**REAL RESEARCH**

## *Electronic Shopping Risks*

Analysis of variance was used to test differences in preferences for electronic shopping for products with different economic and social risks. In a 2 × 2 design, economic risk and social risk were varied at two levels each (high, low). Preference for electronic shopping served as the dependent variable. The results indicated a significant interaction of social risk with economic risk. Electronic shopping was not perceived favorably for high-social-risk products, regardless of the level of economic product risk, but it was preferred for low-economic-risk products over high-economic-risk products when the level of social risk was low.

Despite the results of this study, the number of online shoppers continues to grow. According to a 2005 study by Forrester Research, e-commerce transactions are expected to reach $316 billion by 2010. The increase in shoppers can be attributed to bargain-seeking consumers, convenience of using the Internet, and, surprisingly, an added sense of safety associated with purchasing online. Improved Web sites, streamlined order taking and delivery, and assurances of more secure payment systems have increased the flow of new shoppers to the Internet while decreasing the traditional risk associated with online transaction purchases.[2] ■

The tourist destination example presented a situation with four categories. The *t* test was not appropriate for examining the overall difference in category means, so analysis of variance was used instead. The electronic shopping study involved a comparison of means when there were two factors (independent variables), each of which was varied at two levels. In this example, *t* tests were not appropriate, because the effect of each factor was not independent of the effect of the other factor (in other words, interactions were significant). Analysis of variance provided meaningful conclusions in these studies. The relationship of analysis of variance to the *t* test and other techniques is considered in the next section.

# RELATIONSHIP AMONG TECHNIQUES

***analysis of variance (ANOVA)***
A statistical technique for examining the differences among means for two or more populations.

***factors***
Categorical independent variables. The independent variables must be all categorical (nonmetric) to use ANOVA.

***treatment***
In ANOVA, a particular combination of factor levels or categories.

***one-way analysis of variance***
An ANOVA technique in which there is only one factor.

***n-way analysis of variance***
An ANOVA model where two or more factors are involved.

***analysis of covariance (ANCOVA)***
An advanced analysis of variance procedure in which the effects of one or more metric-scaled extraneous variables are removed from the dependent variable before conducting the ANOVA.

Analysis of variance and analysis of covariance are used for examining the differences in the mean values of the dependent variable associated with the effect of the controlled independent variables, after taking into account the influence of the uncontrolled independent variables. Essentially, ***analysis of variance (ANOVA)*** is used as a test of means for two or more populations. The null hypothesis, typically, is that all means are equal. For example, suppose the researcher was interested in examining whether heavy, medium, light, and nonusers of cereals differed in their preference for Total cereal, measured on a 9-point Likert scale. The null hypothesis that the four groups were not different in preference for Total could be tested using analysis of variance.

In its simplest form, analysis of variance must have a dependent variable (preference for Total cereal) that is metric (measured using an interval or ratio scale). There must also be one or more independent variables (product use: heavy, medium, light, and nonusers). The independent variables must be all categorical (nonmetric). Categorical independent variables are also called ***factors.*** A particular combination of factor levels, or categories, is called a ***treatment. One-way analysis of variance*** involves only one categorical variable, or a single factor. The differences in preference of heavy, medium, light, and nonusers would be examined by one-way ANOVA. In one-way analysis of variance, a treatment is the same as a factor level (medium users constitute a treatment). If two or more factors are involved, the analysis is termed ***n-way analysis of variance.*** If, in addition to product use, the researcher also wanted to examine the preference for Total cereal of customers who are loyal and those who are not, an *n*-way analysis of variance would be conducted.

If the set of independent variables consists of both categorical and metric variables, the technique is called ***analysis of covariance (ANCOVA).*** For example, analysis of covariance would be required if the researcher wanted to examine the preference of product use groups and loyalty groups, taking into account the respondents' attitudes

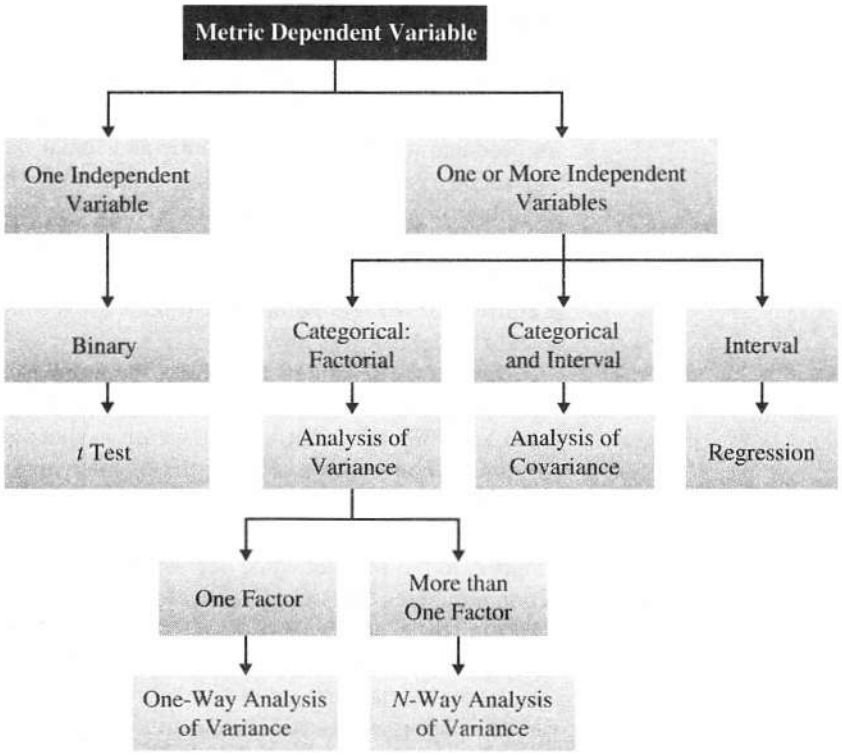



***Figure 16.1***
Relationship Between t Test, Analysis of Variance, Analysis of Covariance, and Regression

toward nutrition and the importance they attached to breakfast as a meal. The latter two variables would be measured on 9-point Likert scales. In this case, the categorical independent variables (product use and brand loyalty) are still referred to as factors, whereas the metric-independent variables (attitude toward nutrition and importance attached to breakfast) are referred to as ***covariates.***

***covariate***
A metric independent variable used in ANCOVA.

The relationship of analysis of variance to $t$ tests and other techniques, such as regression (see Chapter 17), is shown in Figure 16.1. All of these techniques involve a metric dependent variable. ANOVA and ANCOVA can include more than one independent variable (product use, brand loyalty, attitude, and importance). Furthermore, at least one of the independent variables must be categorical, and the categorical variables may have more than two categories (in our example, product use has four categories). A $t$ test, on the other hand, involves a single, binary independent variable. For example, the difference in the preferences of loyal and nonloyal respondents could be tested by conducting a $t$ test. Regression analysis, like ANOVA and ANCOVA, can also involve more than one independent variable. However, all the independent variables are generally interval scaled, although binary or categorical variables can be accommodated using dummy variables. For example, the relationship between preference for Total cereal, attitude toward nutrition, and importance attached to breakfast could be examined via regression analysis, with preference for Total serving as the dependent variable and attitude and importance as independent variables.

## ONE-WAY ANALYSIS OF VARIANCE

Marketing researchers are often interested in examining the differences in the mean values of the dependent variable for several categories of a single independent variable or factor. For example:

- Do the various segments differ in terms of their volume of product consumption?
- Do the brand evaluations of groups exposed to different commercials vary?
- Do retailers, wholesalers, and agents differ in their attitudes toward the firm's distribution policies?

- How do consumers' intentions to buy the brand vary with different price levels?
- What is the effect of consumers' familiarity with the store (measured as high, medium, and low) on preference for the store?

The answer to these and similar questions can be determined by conducting one-way analysis of variance. Before describing the procedure, we define the important statistics associated with one-way analysis of variance.[3]

# STATISTICS ASSOCIATED WITH ONE-WAY ANALYSIS OF VARIANCE

***eta*$^2$ ($\eta^2$).** The strength of the effects of $X$ (independent variable or factor) on $Y$ (dependent variable) is measured by *eta*$^2$ ($\eta^2$). The value of $\eta^2$ varies between 0 and 1.

***F* statistic.** The null hypothesis that the category means are equal in the population is tested by an $F$ statistic based on the ratio of mean square related to $X$ and mean square related to error.

***Mean square.*** The mean square is the sum of squares divided by the appropriate degrees of freedom.

$SS_{between}$. Also denoted as $SS_x$, this is the variation in $Y$ related to the variation in the means of the categories of $X$. This represents variation between the categories of $X$, or the portion of the sum of squares in $Y$ related to $X$.

$SS_{within}$. Also referred to as $SS_{error}$, this is the variation in $Y$ due to the variation within each of the categories of $X$. This variation is not accounted for by $X$.

$SS_y$. The total variation in $Y$ is $SS_y$.

# CONDUCTING ONE-WAY ANALYSIS OF VARIANCE

The procedure for conducting one-way analysis of variance is described in Figure 16.2. It involves identifying the dependent and independent variables, decomposing the total variation, measuring effects, testing significance, and interpreting results. We consider these steps in detail and illustrate them with some applications.

## Identify the Dependent and Independent Variables

The dependent variable is denoted by $Y$ and the independent variable by $X$. $X$ is a categorical variable having $c$ categories. There are $n$ observations on $Y$ for each category of $X$, as shown in Table 16.1. As can be seen, the sample size in each category of $X$ is $n$, and the total sample size $N = n \times c$. Although the sample sizes in the categories of $X$ (the group sizes) are assumed to be equal for the sake of simplicity, this is not a requirement.

*Figure 16.2*
Conducting One-Way ANOVA

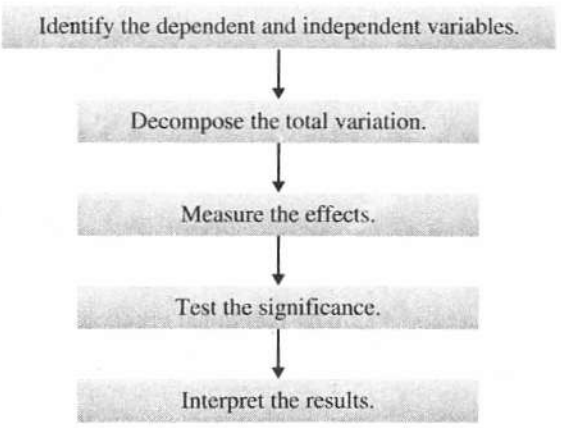
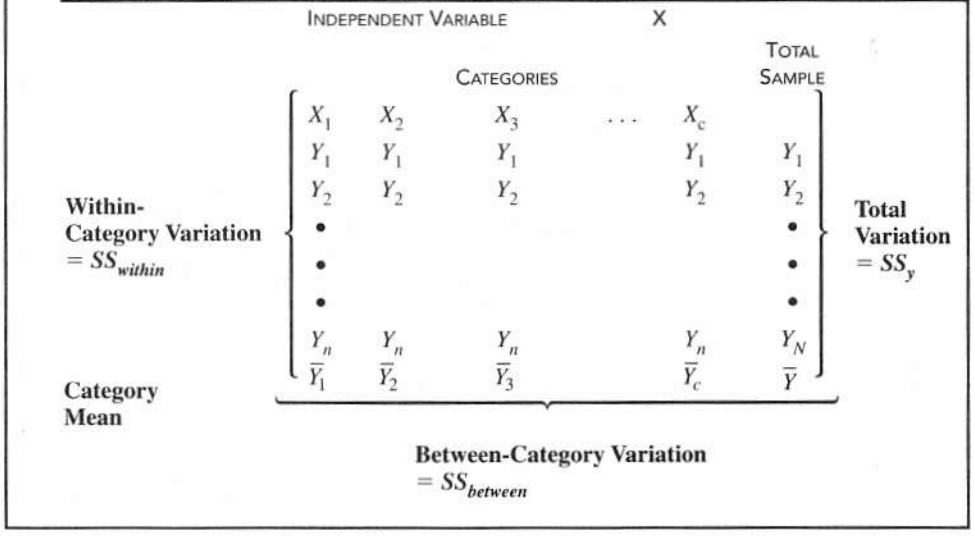


**TABLE 16.1**

**Decomposition of the Total Variation: One-Way ANOVA**




## Decompose the Total Variation

***decomposition of the total variation***
In one-way ANOVA, separation of the variation observed in the dependent variable into the variation due to the independent variables plus the variation due to error.

In examining the differences among means, one-way analysis of variance involves the ***decomposition of the total variation*** observed in the dependent variable. This variation is measured by the sums of squares corrected for the mean (*SS*). Analysis of variance is so named because it examines the variability or variation in the sample (dependent variable) and, based on the variability, determines whether there is reason to believe that the population means differ.

The total variation in $Y$, denoted by $SS_y$, can be decomposed into two components:

$$SS_y = SS_{between} + SS_{within}$$

where the subscripts *between* and *within* refer to the categories of $X$. $SS_{between}$ is the variation in $Y$ related to the variation in the means of the categories of $X$. It represents variation between the categories of $X$. In other words, $SS_{between}$ is the portion of the sum of squares in $Y$ related to the independent variable or factor $X$. For this reason, $SS_{between}$ is also denoted as $SS_x$. $SS_{within}$ is the variation in $Y$ related to the variation within each category of $X$. $SS_{within}$ is not accounted for by $X$. Therefore it is referred to as $SS_{error}$. The total variation in $Y$ may be decomposed as:

$$SS_y = SS_x + SS_{error}$$

where

$$SS_y = \sum_{i=1}^{N} (Y_i - \overline{Y})^2$$

$$SS_x = \sum_{j=1}^{c} n(\overline{Y}_j - \overline{Y})^2$$

$$SS_{error} = \sum_{j}^{c} \sum_{i}^{n} (Y_{ij} - \overline{Y}_j)^2$$

$Y_i$ = individual observation
$\overline{Y}_j$ = mean for category $j$
$\overline{Y}$ = mean over the whole sample, or grand mean
$Y_{ij}$ = $i$th observation in the $j$th category

The logic of decomposing the total variation in $Y$, $SS_y$, into $SS_{between}$ and $SS_{within}$ in order to examine differences in group means can be intuitively understood. Recall from Chapter 15 that if the variation of the variable in the population was known or estimated, one could estimate how much the sample mean should vary because of random variation alone. In analysis of variance, there are several different groups (e.g., heavy, medium, light, and nonusers). If the null hypothesis is true and all the groups have the same mean in the population, one can estimate how much the sample means should vary because of sampling (random) variations alone. If the observed variation in the sample means is more than what would be expected by sampling variation, it is reasonable to conclude that this extra variability is related to differences in group means in the population.

In analysis of variance, we estimate two measures of variation: within groups ($SS_{within}$) and between groups ($SS_{between}$). Within-group variation is a measure of how much the observations, $Y$ values, within a group vary. This is used to estimate the variance within a group in the population. It is assumed that all the groups have the same variation in the population. However, because it is not known that all the groups have the same mean, we cannot calculate the variance of all the observations together. The variance for each of the groups must be calculated individually, and these are combined into an "average" or "overall" variance. Likewise, another estimate of the variance of the $Y$ values may be obtained by examining the variation between the means. (This process is the reverse of determining the variation in the means, given the population variances.) If the population mean is the same in all the groups, then the variation in the sample means and the sizes of the sample groups can be used to estimate the variance of $Y$. The reasonableness of this estimate of the $Y$ variance depends on whether the null hypothesis is true. If the null hypothesis is true and the population means are equal, the variance estimate based on between-group variation is correct. On the other hand, if the groups have different means in the population, the variance estimate based on between-group variation will be too large. Thus, by comparing the $Y$ variance estimates based on between-group and within-group variation, we can test the null hypothesis. Decomposition of the total variation in this manner also enables us to measure the effects of $X$ on $Y$.

## Measure the Effects

The effects of $X$ on $Y$ are measured by $SS_x$. Because $SS_x$ is related to the variation in the means of the categories of $X$, the relative magnitude of $SS_x$ increases as the differences among the means of $Y$ in the categories of $X$ increase. The relative magnitude of $SS_x$ also increases as the variations in $Y$ within the categories of $X$ decrease. The strength of the effects of $X$ on $Y$ are measured as follows:

$$\eta^2 = \frac{SS_x}{SS_y} = \frac{(SS_y - SS_{error})}{SS_y}$$

The value of $\eta^2$ varies between 0 and 1. It assumes a value of 0 when all the category means are equal, indicating that $X$ has no effect of $X$ on $Y$. The value of $\eta^2$ will be 1 when there is no variability within each category of $X$ but there is some variability between categories. Thus, $\eta^2$ is a measure of the variation in $Y$ that is explained by the independent variable $X$. Not only can we measure the effects of $X$ on $Y$, but we can also test for their significance.

## Test the Significance

In one-way analysis of variance, the interest lies in testing the null hypothesis that the category means are equal in the population.[4] In other words,

$$H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_c$$

Under the null hypothesis, $SS_x$ and $SS_{error}$ come from the same source of variation. In such a case, the estimate of the population variance of $Y$ can be based on either between-category variation or within-category variation. In other words, the estimate of the population variance of $Y$,

$$\begin{aligned} S_y^2 &= \frac{SS_x}{(c-1)} \\ &= \text{mean square due to } X \\ &= MS_x \end{aligned}$$

or

$$\begin{aligned} S_y^2 &= \frac{SS_{error}}{(N-c)} \\ &= \text{mean square due to error} \\ &= MS_{error} \end{aligned}$$

The null hypothesis may be tested by the $F$ statistic based on the ratio between these two estimates:

$$F = \frac{SS_x/(c-1)}{SS_{error}/(N-c)} = \frac{MS_x}{MS_{error}}$$

This statistic follows the $F$ distribution, with $(c-1)$ and $(N-c)$ degrees of freedom (df). A table of the $F$ distribution is given as Table 5 in the Statistical Appendix at the end of the book. As mentioned in Chapter 15, the $F$ distribution is a probability distribution of the ratios of sample variances. It is characterized by degrees of freedom for the numerator and degrees of freedom for the denominator.[5]

## Interpret the Results

If the null hypothesis of equal category means is not rejected, then the independent variable does not have a significant effect on the dependent variable. On the other hand, if the null hypothesis is rejected, then the effect of the independent variable is significant. In other words, the mean value of the dependent variable will be different for different categories of the independent variable. A comparison of the category mean values will indicate the nature of the effect of the independent variable. Other salient issues in the interpretation of results, such as examination of differences among specific means, are discussed later.

# ILLUSTRATIVE DATA

We illustrate the concepts discussed in this chapter using the data presented in Table 16.2. For illustrative purposes, we consider only a small number of observations. In actual practice, analysis of variance is performed on a much larger sample such as that in the Dell Experiential Research considered later. These data were generated by an experiment in which a major department store chain wanted to examine the effect of the level of in-store promotion and a storewide coupon on sales. In-store promotion was varied at three levels: high (1), medium (2), and low (3). Couponing was manipulated at two levels. Either a $20 storewide coupon was distributed to potential shoppers (denoted by 1) or it was not (denoted by 2 in Table 16.2). In-store promotion and couponing were crossed, resulting in a $3 \times 2$ design with six cells. Thirty stores were randomly selected, and five stores were randomly assigned to each treatment condition, as shown in Table 16.2. The experiment was run for two months. Sales in each store were measured, normalized to account for extraneous factors (store size, traffic, etc.), and converted to a 1-to-10 scale. In addition, a qualitative assessment was made of the relative affluence of the clientele of each store, again using a 1-to-10 scale. In these scales, higher numbers denote higher sales or more affluent clientele.


SPSS Data File

**TABLE 16.2**

**Coupon Level, In-Store Promotion, Sales, and Clientele Rating**

| Store Number | Coupon Level | In-Store Promotion | Sales | Clientele Rating |
|---|---|---|---|---|
| 1 | 1 | 1 | 10 | 9 |
| 2 | 1 | 1 | 9 | 10 |
| 3 | 1 | 1 | 10 | 8 |
| 4 | 1 | 1 | 8 | 4 |
| 5 | 1 | 1 | 9 | 6 |
| 6 | 1 | 2 | 8 | 8 |
| 7 | 1 | 2 | 8 | 4 |
| 8 | 1 | 2 | 7 | 10 |
| 9 | 1 | 2 | 9 | 6 |
| 10 | 1 | 2 | 6 | 9 |
| 11 | 1 | 3 | 5 | 8 |
| 12 | 1 | 3 | 7 | 9 |
| 13 | 1 | 3 | 6 | 6 |
| 14 | 1 | 3 | 4 | 10 |
| 15 | 1 | 3 | 5 | 4 |
| 16 | 2 | 1 | 8 | 10 |
| 17 | 2 | 1 | 9 | 6 |
| 18 | 2 | 1 | 7 | 8 |
| 19 | 2 | 1 | 7 | 4 |
| 20 | 2 | 1 | 6 | 9 |
| 21 | 2 | 2 | 4 | 6 |
| 22 | 2 | 2 | 5 | 8 |
| 23 | 2 | 2 | 5 | 10 |
| 24 | 2 | 2 | 6 | 4 |
| 25 | 2 | 2 | 4 | 9 |
| 26 | 2 | 3 | 2 | 4 |
| 27 | 2 | 3 | 3 | 6 |
| 28 | 2 | 3 | 2 | 10 |
| 29 | 2 | 3 | 1 | 9 |
| 30 | 2 | 3 | 2 | 8 |

# ILLUSTRATIVE APPLICATIONS OF ONE-WAY ANALYSIS OF VARIANCE

We illustrate one-way ANOVA first with an example showing calculations done by hand and then using computer analysis. Suppose that only one factor, namely in-store promotion, was manipulated, that is, let us ignore couponing for the purpose of this illustration. The department store is attempting to determine the effect of in-store promotion ($X$) on sales ($Y$). For the purpose of illustrating hand calculations, the data of Table 16.2 are transformed in Table 16.3 to show the store ($Y_{ij}$) for each level of promotion.

The null hypothesis is that the category means are equal:

$$H_0: \mu_1 = \mu_2 = \mu_3$$

To test the null hypothesis, the various sums of squares are computed as follows:

$$\begin{aligned} SS_y = {} & (10 - 6.067)^2 + (9 - 6.067)^2 + (10 - 6.067)^2 + (8 - 6.067)^2 + (9 - 6.067)^2 \\ & + (8 - 6.067)^2 + (9 - 6.067)^2 + (7 - 6.067)^2 + (7 - 6.067)^2 + (6 - 6.067)^2 \\ & + (8 - 6.067)^2 + (8 - 6.067)^2 + (7 - 6.067)^2 + (9 - 6.067)^2 + (6 - 6.067)^2 \\ & + (4 - 6.067)^2 + (5 - 6.067)^2 + (5 - 6.067)^2 + (6 - 6.067)^2 + (4 - 6.067)^2 \\ & + (5 - 6.067)^2 + (7 - 6.067)^2 + (6 - 6.067)^2 + (4 - 6.067)^2 + (5 - 6.067)^2 \\ & + (2 - 6.067)^2 + (3 - 6.067)^2 + (2 - 6.067)^2 + (1 - 6.067)^2 + (2 - 6.067)^2 \end{aligned}$$

**TABLE 16.3**

**Effect of In-Store Promotion on Sales**

| Store No. | Level of In-Store Promotion | | |
|---|---|---|---|
| | High | Medium | Low |
| | | Normalized Sales | |
| 1 | 10 | 8 | 5 |
| 2 | 9 | 8 | 7 |
| 3 | 10 | 7 | 6 |
| 4 | 8 | 9 | 4 |
| 5 | 9 | 6 | 5 |
| 6 | 8 | 4 | 2 |
| 7 | 9 | 5 | 3 |
| 8 | 7 | 5 | 2 |
| 9 | 7 | 6 | 1 |
| 10 | 6 | 4 | 2 |
| Column Totals | 83 | 62 | 37 |
| Category means: $\overline{Y}_j$ | $\frac{83}{10}$ | $\frac{62}{10}$ | $\frac{37}{10}$ |
| | $= 8.3$ | $= 6.2$ | $= 3.7$ |
| Grand mean, $\overline{Y}$ | | $= \frac{(83 + 62 + 37)}{30} = 6.067$ | |

$$
\begin{aligned}
&= (3.933)^2 + (2.933)^2 + (3.933)^2 + (1.933)^2 + (2.933)^2 \\
&\quad + (1.933)^2 + (2.933)^2 + (0.933)^2 + (0.933)^2 + (-0.067)^2 \\
&\quad + (1.933)^2 + (1.933)^2 + (0.933)^2 + (2.933)^2 + (-0.067)^2 \\
&\quad + (-2.067)^2 + (-1.067)^2 + (-1.067)^2 + (-0.067)^2 + (-2.067)^2 \\
&\quad + (-1.067)^2 + (0.933)^2 + (-0.067)^2 + (-2.067)^2 + (-1.067)^2 \\
&\quad + (-4.067)^2 + (-3.067)^2 + (-4.067)^2 + (-5.067)^2 + (-4.067)^2 \\
&= 185.867
\end{aligned}
$$

$$
\begin{aligned}
SS_x &= 10(8.3 - 6.067)^2 + 10(6.2 - 6.067)^2 + 10(3.7 - 6.067)^2 \\
&= 10(2.233)^2 + 10(0.133)^2 + 10(-2.367)^2 \\
&= 106.067
\end{aligned}
$$

$$
\begin{aligned}
SS_{error} &= (10 - 8.3)^2 + (9 - 8.3)^2 + (10 - 8.3)^2 + (8 - 8.3)^2 + (9 - 8.3)^2 \\
&\quad + (8 - 8.3)^2 + (9 - 8.3)^2 + (7 - 8.3)^2 + (7 - 8.3)^2 + (6 - 8.3)^2 \\
&\quad + (8 - 6.2)^2 + (8 - 6.2)^2 + (7 - 6.2)^2 + (9 - 6.2)^2 + (6 - 6.2)^2 \\
&\quad + (4 - 6.2)^2 + (5 - 6.2)^2 + (5 - 6.2)^2 + (6 - 6.2)^2 + (4 - 6.2)^2 \\
&\quad + (5 - 3.7)^2 + (7 - 3.7)^2 + (6 - 3.7)^2 + (4 - 3.7)^2 + (5 - 3.7)^2 \\
&\quad + (2 - 3.7)^2 + (3 - 3.7)^2 + (2 - 3.7)^2 + (1 - 3.7)^2 + (2 - 3.7)^2 \\
&= (1.7)^2 + (0.7)^2 + (1.7)^2 + (-0.3)^2 + (0.7)^2 \\
&\quad + (-0.3)^2 + (0.7)^2 + (-1.3)^2 + (-1.3)^2 + (-2.3)^2 \\
&\quad + (1.8)^2 + (1.8)^2 + (0.8)^2 + (2.8)^2 + (-0.2)^2 \\
&\quad + (-2.2)^2 + (-1.2)^2 + (-1.2)^2 + (-0.2)^2 + (-2.2)^2 \\
&\quad + (1.3)^2 + (3.3)^2 + (2.3)^2 + (0.3)^2 + (1.3)^2 \\
&\quad + (-1.7)^2 + (-0.7)^2 + (-1.7)^2 + (-2.7)^2 + (-1.7)^2 \\
&= 79.80
\end{aligned}
$$

It can be verified that

$$SS_y = SS_x + SS_{error}$$

as follows:

$$185.867 = 106.067 + 79.80$$

The strength of the effects of $X$ on $Y$ are measured as follows:

$$\eta^2 = \frac{SS_x}{SS_y} = \frac{106.067}{185.867} = 0.571$$

In other words, 57.1 percent of the variation in sales ($Y$) is accounted for by in-store promotion ($X$), indicating a modest effect. The null hypothesis may now be tested.

$$F = \frac{SS_x/(c-1)}{SS_{error}/(N-c)} = \frac{MS_x}{MS_{error}}$$

$$F = \frac{106.067/(3-1)}{79.800/(30-3)} = 17.944$$

From Table 5 in the Statistical Appendix, we see that for 2 and 27 degrees of freedom, the critical value of $F$ is 3.35 for $\alpha = 0.05$. Because the calculated value of $F$ is greater than the critical value, we reject the null hypothesis. We conclude that the population means for the three levels of in-store promotion are indeed different. The relative magnitudes of the means for the three categories indicate that a high level of in-store promotion leads to significantly higher sales.

We now illustrate the analysis-of-variance procedure using a computer program. The results of conducting the same analysis by computer are presented in Table 16.4. The value of $SS_x$ denoted by main effects is 106.067 with 2 df; that of $SS_{error}$ denoted by residual is 79.80 with 27 df. Therefore, $MS_x = 106.067/2 = 53.033$, and $MS_{error} = 79.80/27 = 2.956$. The value of $F = 53.033/2.956 = 17.944$ with 2 and 27 degrees of freedom, resulting in a probability of 0.000. Because the associated probability is less than the significance level of 0.05, the null hypothesis of equal population means is rejected. Alternatively, it can be seen from Table 5 in the Statistical Appendix that the critical value of $F$ for 2 and 27 degrees of freedom is 3.35. Because the calculated value of $F$ (17.944) is larger than the critical value, the null hypothesis is rejected. As can be seen from Table 16.4, the sample means, with values of 8.3, 6.2, and 3.7, are quite different. Stores with a high level of in-store promotion have the highest average sales (8.3) and stores with a low level of in-store promotion have the lowest average sales (3.7). Stores with a medium level of in-store promotion have an intermediate level of average sales (6.2). These findings

SPSS

**SPSS Output File**

**TABLE 16.4**

**One-Way ANOVA: Effect of In-Store Promotion on Store Sales**

| Source of Variation | Sum of Squares | DF | Mean Square | F Ratio | F Prob. |
|---|---|---|---|---|---|
| Between groups (In-store promotion) | 106.067 | 2 | 53.033 | 17.944 | 0.000 |
| Within groups (Error) | 79.800 | 27 | 2.956 | | |
| TOTAL | 185.867 | 29 | 6.409 | | |

| Cell Means | | |
|---|---|---|
| Level of In-store promotion | Count | Mean |
| High (1) | 10 | 8.300 |
| Medium (2) | 10 | 6.200 |
| Low (3) | 10 | 3.700 |
| TOTAL | 30 | 6.067 |

**ACTIVE RESEARCH**

Visit *www.dell.com* and conduct an Internet search using a search engine and your library's online database to obtain information on computer usage in U.S. households.

As the marketing director for Dell, how would you segment the home computer market?

As a marketing research analyst working for Dell, how would you determine whether the three home computer usage segments (experts, novices, and nonusers) differ in terms of each of 10 psychographic characteristics, each measured on a 7-point scale?

seem plausible. Instead of 30 stores, if this were a large and representative sample, the implications would be that management seeking to increase sales should emphasize in-store promotion.

The procedure for conducting one-way analysis of variance and the illustrative application help us understand the assumptions involved.

## ASSUMPTIONS IN ANALYSIS OF VARIANCE

The salient assumptions in analysis of variance can be summarized as follows.

1. Ordinarily, the categories of the independent variable are assumed to be fixed. Inferences are made only to the specific categories considered. This is referred to as the *fixed-effects model.* Other models are also available. In the *random-effects model,* the categories or treatments are considered to be random samples from a universe of treatments. Inferences are made to other categories not examined in the analysis. A *mixed-effects model* results if some treatments are considered fixed and others random.[6]
2. The error term is normally distributed, with a zero mean and a constant variance. The error is not related to any of the categories of $X$. Modest departures from these assumptions do not seriously affect the validity of the analysis. Furthermore, the data can be transformed to satisfy the assumption of normality or equal variances.
3. The error terms are uncorrelated. If the error terms are correlated (i.e., the observations are not independent), the $F$ ratio can be seriously distorted.

In many data analysis situations, these assumptions are reasonably met. Analysis of variance is therefore a common procedure, as illustrated by the following example.

**REAL RESEARCH**

### *Videologs Put Marketers in the Picture*

Although the videolog, a shop-at-home video catalog, is still in its infancy, many direct marketers have shown an interest in its use. Companies such as Spiegel (*www.spiegel.com*) and Neiman Marcus (*www.neimanmarcus.com*) either plan to offer or already have offered video catalogs to consumers.

A study was designed to investigate the effectiveness of videolog retailing as a form of direct marketing. Subjects were randomly assigned to one of three treatments: (a) videolog only; (b) both videolog and catalog; or (c) catalog only. The dependent variables of interest, consisting of attitudes and opinions, were: (1) assessments of product (clothing) attributes; (2) assessments of the videolog/catalog sponsoring company; (3) assessments of price information; and (4) intentions to purchase.

One-way analysis of variance was conducted separately for each dependent variable. The results showed that respondents exposed to the videolog, or videolog and catalog, perceived the clothing more positively than did those exposed only to the catalog. Although the videolog-only treatment enhanced perceptions of the sponsoring company, the results were not as striking as were those for clothing perceptions. No significant differences were found in price perceptions and intentions to purchase. Yet the mean number of items respondents said they were likely to purchase was greater for those viewing both the videolog and catalog than those seeing just the videolog or the catalog.

Although this study was an exploratory effort, the positive results found in assessments of clothing seen in the videolog suggest that this is an area that may have potential for direct marketers.[7] ■

## N-WAY ANALYSIS OF VARIANCE

In marketing research, one is often concerned with the effect of more than one factor simultaneously.[8] For example:

- How do the consumers' intentions to buy a brand vary with different levels of price and different levels of distribution?
- How do advertising levels (high, medium, and low) interact with price levels (high, medium, and low) to influence a brand's sale?
- Do educational levels (less than high school, high school graduate, some college, and college graduate) and age (less than 35, 35–55, more than 55) affect consumption of a brand?
- What is the effect of consumers' familiarity with a department store (high, medium, and low) and store image (positive, neutral, and negative) on preference for the store?

In determining such effects, *n*-way analysis of variance can be used. A major advantage of this technique is that it enables the researcher to examine interactions between the factors. ***Interactions*** occur when the effects of one factor on the dependent variable depend on the level (category) of the other factors. The procedure for conducting *n*-way analysis of variance is similar to that for one-way analysis of variance. The statistics associated with *n*-way analysis of variance are also defined similarly. Consider the simple case of two factors, $X_1$ and $X_2$, having categories $c_1$ and $c_2$. The total variation in this case is partitioned as follows:

***interaction***
When assessing the relationship between two variables, an interaction occurs if the effect of $X_1$ depends on the level of $X_2$, and vice versa.

$$SS_{total} = SS \text{ due to } X_1 + SS \text{ due to } X_2 + SS \text{ due to interaction of } X_1 \text{ and } X_2 + SS_{within}$$

or

$$SS_y = SS_{x_1} + SS_{x_2} + SS_{x_1x_2} + SS_{error}$$

A larger effect of $X_1$ will be reflected in a greater mean difference in the levels of $X_1$ and a larger $SS_{x_1}$. The same is true for the effect of $X_2$. The larger the interaction between $X_1$ and $X_2$, the larger $SS_{x_1x_2}$ will be. On the other hand, if $X_1$ and $X_2$ are independent, the value of $SS_{x_1x_2}$ will be close to zero.[9]

***multiple* $\eta^2$**
The strength of the joint effect of two (or more) factors, or the overall effect.

The strength of the joint effect of two factors, called the overall effect, or ***multiple*** $\eta^2$, is measured as follows:

$$\text{multiple } \eta^2 = \frac{(SS_{x_1} + SS_{x_2} + SS_{x_1x_2})}{SS_y}$$

***significance of the overall effect***
A test that some differences exist between some of the treatment groups.

The ***significance of the overall effect*** may be tested by an $F$ test, as follows:

$$F = \frac{(SS_{x_1} + SS_{x_2} + SS_{x_1x_2})/\text{df}_n}{SS_{error}/\text{df}_d}$$

$$= \frac{SS_{x_1,x_2,x_1x_2}/\text{df}_n}{SS_{error}/\text{df}_d}$$

$$= \frac{MS_{x_1,x_2,x_1x_2}}{MS_{error}}$$

where

$\text{df}_n$ = degrees of freedom for the numerator
$= (c_1 - 1) + (c_2 - 1) + (c_1 - 1)(c_2 - 1)$
$= c_1c_2 - 1$
$\text{df}_d$ = degrees of freedom for the denominator
$= N - c_1c_2$
$MS$ = mean square

***significance of the interaction effect***
A test of the significance of the interaction between two or more independent variables.

If the overall effect is significant, the next step is to examine the ***significance of the interaction effect.*** Under the null hypothesis of no interaction, the appropriate $F$ test is:

$$F = \frac{SS_{x_1x_2}/\text{df}_n}{SS_{error}/\text{df}_d}$$

$$= \frac{MS_{x_1x_2}}{MS_{error}}$$

where

$$\text{df}_n = (c_1 - 1)(c_2 - 1)$$
$$\text{df}_d = N - c_1c_2$$

If the interaction effect is found to be significant, then the effect of $X_1$ depends on the level of $X_2$, and vice versa. Because the effect of one factor is not uniform, but varies with the level of the other factor, it is not generally meaningful to test the significance of the main effects. However, it is meaningful to test the significance of each main effect of each factor if the interaction effect is not significant.[10]

***significance of the main effect***
A test of the significance of the main effect for each individual factor.

The ***significance of the main effect*** of each factor may be tested as follows for $X_1$:

$$F = \frac{SS_{x_1}/\text{df}_n}{SS_{error}/\text{df}_d}$$

$$= \frac{MS_{x_1}}{MS_{error}}$$

where

$$\text{df}_n = c_1 - 1$$
$$\text{df}_d = N - c_1c_2$$

The foregoing analysis assumes that the design was orthogonal, or balanced (the number of cases in each cell was the same). If the cell size varies, the analysis becomes more complex.

# ILLUSTRATIVE APPLICATION OF *N*-WAY ANALYSIS OF VARIANCE

Returning to the data of Table 16.2, let us now examine the effect of the level of in-store promotion and couponing on store sales. The results of running a 3 × 2 ANOVA on the computer are presented in Table 16.5. For the main effect of level of promotion, the sum of squares $SS_{x_p}$, degrees of freedom, and mean square $MS_{x_p}$ are the same as earlier determined in Table 16.4. The sum of squares for couponing $SS_{x_c} = 53.333$ with 1 df, resulting in an identical value for the mean square $MS_{x_c}$. The combined main effect is determined by adding the sum of squares due to the two main effects ($SS_{x_p} + SS_{x_c} = 106.067 + 53.333 = 159.400$) as well as adding the degrees of freedom ($2 + 1 = 3$). For the promotion and coupon interaction effect, the sum of squares $SS_{x_px_c} = 3.267$ with $(3 - 1)(2 - 1) = 2$ degrees of freedom, resulting in $MS_{x_px_c} = 3.267/2 = 1.633$. For the overall (model) effect, the sum of squares is the addition of the sum of squares for promotion main effect, coupon main effect, and interaction effect = 106.067 + 53.333 + 3.267 = 162.667 with 2 + 1 + 2 = 5 degrees of freedom, resulting in a mean square of 162.667/5 = 32.533. Note, however, the error statistics are now different than in Table 16.4. This is due to the fact that we now have two factors instead of one. $SS_{error} = 23.2$ with $(30 - 3 \times 2)$ or 24 degrees of freedom resulting in $MS_{error} = 23.2/24 = 0.967$.

The test statistic for the significance of the overall effect is

$$F = \left(\frac{32.533}{0.967}\right)$$

$$= 33.655$$

with 5 and 24 degrees of freedom, which is significant at the 0.05 level.

SPSS

**SPSS Output File**

**TABLE 16.5**

Two-Way Analysis of Variance

| Source of Variation | Sum of Squares | df | Mean Square | F | Sig. of F | $\omega^2$ |
|---|---|---|---|---|---|---|
| Main Effects | | | | | | |
| In-store promotion | 106.067 | 2 | 53.033 | 54.862 | 0.000 | 0.557 |
| Coupon | 53.333 | 1 | 53.333 | 55.172 | 0.000 | 0.280 |
| Combined | 159.400 | 3 | 53.133 | 54.966 | 0.000 | |
| Two-way interaction | 3.267 | 2 | 1.633 | 1.690 | 0.206 | |
| Model | 162.667 | 5 | 32.533 | 33.655 | 0.000 | |
| Residual (Error) | 23.200 | 24 | 0.967 | | | |
| TOTAL | 185.867 | 29 | 6.409 | | | |
| **Cell Means** | | | | | | |
| **In-store Promotion** | **Coupon** | **Count** | **Mean** | | | |
| High | Yes | 5 | 9.200 | | | |
| High | No | 5 | 7.400 | | | |
| Medium | Yes | 5 | 7.600 | | | |
| Medium | No | 5 | 4.800 | | | |
| Low | Yes | 5 | 5.400 | | | |
| Low | No | 5 | 2.000 | | | |
| **Factor Level Means** | | | | | | |
| **Promotion** | **Coupon** | **Count** | **Mean** | | | |
| High | | 10 | 8.300 | | | |
| Medium | | 10 | 6.200 | | | |
| Low | | 10 | 3.700 | | | |
| | Yes | 15 | 7.400 | | | |
| | No | 15 | 4.733 | | | |
| Grand Mean | | 30 | 6.067 | | | |

The test statistic for the significance of the interaction effect is

$$F = \left(\frac{1.633}{0.967}\right)$$

$$= 1.690$$

with 2 and 24 degrees of freedom, which is not significant at the 0.05 level.

Because the interaction effect is not significant, the significance of the main effects can be evaluated. The test statistic for the significance of the main effect of promotion is

$$F = \left(\frac{53.033}{0.967}\right)$$

$$= 54.862$$

with 2 and 24 degrees of freedom, which is significant at the 0.05 level.

The test statistic for the significance of the main effect of coupon is

$$F = \left(\frac{53.333}{0.967}\right)$$

$$= 55.172$$

with 1 and 24 degrees of freedom, which is significant at the 0.05 level. Thus, a higher level of promotion results in higher sales. The distribution of a storewide coupon results in higher sales. The effect of each is independent of the other. If this were a large and representative sample, the implications are that management can increase sales by increasing in-store promotion and the use of coupons, independently of the other.

**REAL RESEARCH**

## *Country Affiliation Affects TV Reception*

A study examined the impact of country affiliation on the credibility of product-attribute claims for TVs. The dependent variables were the following product-attribute claims: good sound, reliability, crisp-clear picture, and stylish design. The independent variables that were manipulated consisted of price, country affiliation, and store distribution. A $2 \times 2 \times 2$ between-subjects design was used. Two levels of price, \$349.95 (low) and \$449.95 (high), two levels of country affiliation, Korea and the United States, and two levels of store distribution, Hudson's and without Hudson's, were specified.

Data were collected from two suburban malls in a large midwestern city. Thirty respondents were randomly assigned to each of the eight treatment cells for a total of 240 subjects. Table 1 presents the results for manipulations that had significant effects on each of the dependent variables.

TABLE 1 Analyses for Significant Manipulations

| *Effect* | *Dependent Variable* | *Univariate F* | *df* | *p* |
|---|---|---|---|---|
| Country × price | Good sound | 7.57 | 1,232 | 0.006 |
| Country × price | Reliability | 6.57 | 1,232 | 0.011 |
| Country × distribution | Crisp-clear picture | 6.17 | 1,232 | 0.014 |
| Country × distribution | Reliability | 6.57 | 1,232 | 0.011 |
| Country × distribution | Stylish design | 10.31 | 1,232 | 0.002 |

The directions of country-by-distribution interaction effects for the three dependent variables are shown in Table 2. Whereas the credibility ratings for the crisp-clear picture, reliability, and stylish design claims are improved by distributing the Korean-made TV set through Hudson's, rather than some other distributor, the same is not true of a U.S.-made set. Similarly, the directions of country-by-price interaction effects for the two dependent variables are shown in Table 3. At \$449.95, the credibility ratings for the "good sound" and "reliability" claims are higher for the U.S.-made TV set than for its Korean counterpart, but there is little difference related to country affiliation when the product is priced at \$349.95.

TABLE 2 Country-by-Distribution Interaction Means

| *Country × Distribution* | *Crisp-Clear Picture* | *Reliability* | *Stylish Design* |
|---|---|---|---|
| *Korea* | | | |
| Hudson's | 3.67 | 3.42 | 3.82 |
| Without Hudson's | 3.18 | 2.88 | 3.15 |
| *United States* | | | |
| Hudson's | 3.60 | 3.47 | 3.53 |
| Without Hudson's | 3.77 | 3.65 | 3.75 |

TABLE 3 Country-by-Price Interaction Means

| *Country × Price* | *Good Sound* | *Reliability* |
|---|---|---|
| \$349.95 | | |
| Korea | 3.75 | 3.40 |
| United States | 3.53 | 3.45 |
| \$449.95 | | |
| Korea | 3.15 | 2.90 |
| United States | 3.73 | 3.67 |

This study demonstrates that credibility of attribute claims, for products traditionally exported to the United States by a company in a newly industrialized country, can be significantly improved if the same company distributes the product through a prestigious U.S. retailer and considers making manufacturing investments in the United States.

**ACTIVE RESEARCH**

Visit *www.levis.com* and search the Internet using a search engine as well as your library's online databases to find information on consumer preferences for jeans.

Levi's would like to conduct marketing research to increase its share of the jeans market. Past studies suggest that the two most important factors determining the preferences for jeans are price (high, medium, and low) and quality (high, medium, and low). What design would you adopt and what analysis would you conduct to determine the effects of these factors on preference for jeans?

As Levi's marketing chief, what information would you need to formulate strategies aimed at increasing marketing share?

Specifically, three product attribute claims (crisp-clear picture, reliability, and stylish design) are perceived as more credible when the TVs are made in Korea if they are also distributed through a prestigious U.S. retailer. Also, the "good sound" and "reliability" claims for TVs are perceived to be more credible for a U.S-made set sold at a higher price, possibly offsetting the potential disadvantage of higher manufacturing costs in the United States. Thus, Thomson, the manufacturer of RCA products (*www.rca.com*), may be better off manufacturing its TV sets in the United States and selling them at a higher price. The Dayton Hudson Corporation was renamed the Target Corporation in 2000.[11] ■

## ANALYSIS OF COVARIANCE

When examining the differences in the mean values of the dependent variable related to the effect of the controlled independent variables, it is often necessary to take into account the influence of uncontrolled independent variables. For example:

- In determining how consumers' intentions to buy a brand vary with different levels of price, attitude toward the brand may have to be taken into consideration.
- In determining how different groups exposed to different commercials evaluate a brand, it may be necessary to control for prior knowledge.
- In determining how different price levels will affect a household's cereal consumption, it may be essential to take household size into account.

In such cases, analysis of covariance should be used. Analysis of covariance includes at least one categorical independent variable and at least one interval or metric independent variable. The categorical independent variable is called a *factor,* whereas the metric independent variable is called a *covariate.* The most common use of the covariate is to remove extraneous variation from the dependent variable, because the effects of the factors are of major concern. The variation in the dependent variable due to the covariates is removed by an adjustment of the dependent variable's mean value within each treatment condition. An analysis of variance is then performed on the adjusted scores.[12] The significance of the combined effect of the covariates, as well as the effect of each covariate, is tested by using the appropriate $F$ tests. The coefficients for the covariates provide insights into the effect that the covariates exert on the dependent variable. Analysis of covariance is most useful when the covariate is linearly related to the dependent variable and is not related to the factors.[13]

We again use the data of Table 16.2 to illustrate analysis of covariance. Suppose that we wanted to determine the effect of in-store promotion and couponing on sales while controlling for the effect of clientele. It is felt that the affluence of the clientele may also have an effect on sales of the department store. The dependent variable consists of store sales. As before, promotion has three levels and couponing has two. Clientele measured on an interval scale serves as the covariate. The results are shown in Table 16.6. As can be seen, the sum of squares attributable to the covariate is very small (0.838) with 1 df resulting in an identical value for the mean square. The associated $F$ value is 0.838/0.972 = 0.862, with 1 and 23 degrees of freedom, which is not significant at the 0.05 level. Thus, the conclusion is that the affluence of the clientele does not

SPSS Output File

**TABLE 16.6**
**Analysis of Covariance**

| Source of Variation | Sum of Squares | df | Mean Square | F | Sig. of F |
|---|---|---|---|---|---|
| Covariates | | | | | |
| Clientele | 0.838 | 1 | 0.838 | 0.862 | 0.363 |
| Main effects | | | | | |
| Promotion | 106.067 | 2 | 53.033 | 54.546 | 0.000 |
| Coupon | 53.333 | 1 | 53.333 | 54.855 | 0.000 |
| Combined | 159.400 | 3 | 53.133 | 54.649 | 0.000 |
| 2-way interaction | | | | | |
| Promotion × Coupon | 3.267 | 2 | 1.633 | 1.680 | 0.208 |
| Model | 163.505 | 6 | 27.251 | 28.028 | 0.000 |
| Residual (Error) | 22.362 | 23 | 0.972 | | |
| TOTAL | 185.867 | 29 | 6.409 | | |
| Covariate | Raw coefficient | | | | |
| Clientele | −0.078 | | | | |

have an effect on the sales of the department store. If the effect of the covariate is significant, the sign of the raw coefficient can be used to interpret the direction of the effect on the dependent variable.

# ISSUES IN INTERPRETATION

Important issues involved in the interpretation of ANOVA results include interactions, relative importance of factors, and multiple comparisons.

## Interactions

*ordinal interaction*
An interaction where the rank order of the effects attributable to one factor does not change across the levels of the second factor.

*disordinal interaction*
The change in the rank order of the effects of one factor across the levels of another.

The different interactions that can arise when conducting ANOVA on two or more factors are shown in Figure 16.3. One outcome is that ANOVA may indicate that there are no interactions (the interaction effects are not found to be significant). The other possibility is that the interaction is significant. An *interaction effect* occurs when the effect of an independent variable on a dependent variable is different for different categories or levels of another independent variable. The interaction may be ordinal or disordinal. In ***ordinal interaction,*** the rank order of the effects related to one factor does not change across the levels of the second factor. ***Disordinal interaction,*** on the other hand, involves a change in the rank order of the effects of one factor across the levels of another. If the interaction is disordinal, it could be of a noncrossover or crossover type.[14]

*Figure 16.3*
A Classification of Interaction Effects

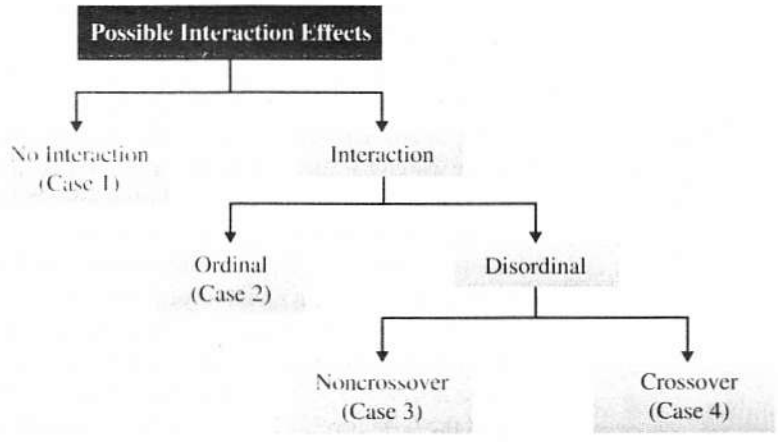
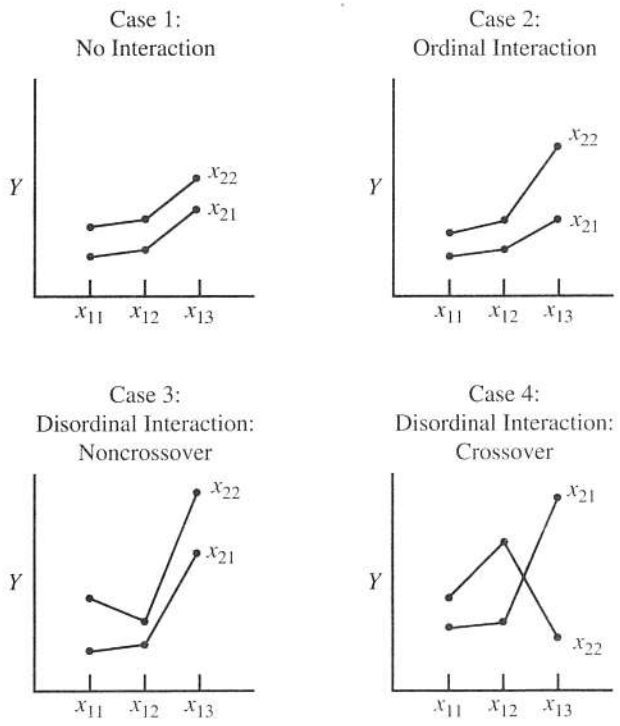


*Figure 16.4*
Patterns of Interaction




These interaction cases are displayed in Figure 16.4, which assumes that there are two factors, $X_1$ with three levels ($X_{11}$, $X_{12}$, and $X_{13}$), and $X_2$ with two levels ($X_{21}$ and $X_{22}$). Case 1 depicts no interaction. The effects of $X_1$ on $Y$ are parallel over the two levels of $X_2$. Although there is some departure from parallelism, this is not beyond what might be expected from chance. Parallelism implies that the net effect of $X_{22}$ over $X_{21}$ is the same across the three levels of $X_1$. In the absence of interaction, the joint effect of $X_1$ and $X_2$ is simply the sum of their individual main effects.

Case 2 depicts an ordinal interaction. The line segments depicting the effects of $X_1$ and $X_2$ are not parallel. The difference between $X_{22}$ and $X_{21}$ increases as we move from $X_{11}$ to $X_{12}$ and from $X_{12}$ to $X_{13}$, but the rank order of the effects of $X_1$ is the same over the two levels of $X_2$. This rank order, in ascending order, is $X_{11}$, $X_{12}$, $X_{13}$, and it remains the same for $X_{21}$ and $X_{22}$.

Disordinal interaction of a noncrossover type is displayed by case 3. The lowest effect of $X_1$ at level $X_{21}$ occurs at $X_{11}$, and the rank order of effects is $X_{11}$, $X_{12}$, and $X_{13}$. However, at level $X_{22}$, the lowest effect of $X_1$ occurs at $X_{12}$, and the rank order is changed to $X_{12}$, $X_{11}$, $X_{13}$. Because it involves a change in rank order, disordinal interaction is stronger than ordinal interaction.

In disordinal interactions of a crossover type, the line segments cross each other, as shown by case 4 in Figure 16.4. In this case, the relative effect of the levels of one factor changes with the levels of the other. Note that $X_{22}$ has a greater effect than $X_{21}$ when the levels of $X_1$ are $X_{11}$ and $X_{12}$. When the level of $X_1$ is $X_{13}$, the situation is reversed, and $X_{21}$ has a greater effect than $X_{22}$. (Note that in cases 1, 2, and 3, $X_{22}$ had a greater impact than $X_{21}$ across all three levels of $X_1$.) Hence, disordinal interactions of a crossover type represent the strongest interactions.[15]

## Relative Importance of Factors

Experimental designs are usually balanced in that each cell contains the same number of respondents. This results in an orthogonal design in which the factors are uncorrelated. Hence, it is possible to determine unambiguously the relative importance of each factor in explaining the variation in the dependent variable. The most commonly used measure in

***omega squared*** ($\omega^2$)
A measure indicating the proportion of the variation in the dependent variable explained by a particular independent variable or factor.

ANOVA is ***omega squared,*** ($\omega^2$). This measure indicates what proportion of the variation in the dependent variable is related to a particular independent variable or factor. The relative contribution of a factor $X$ is calculated as follows:[16]

$$\omega_x^2 = \frac{SS_x - (df_x \times MS_{error})}{SS_{total} + MS_{error}}$$

Note that the estimated value of $\omega^2$ can be negative, in which case the estimated value of $\omega^2$ is set equal to zero. Normally, $\omega^2$ is interpreted only for statistically significant effects.[17] In Table 16.5, $\omega^2$ associated with the level of in-store promotion is calculated as follows:

$$\omega_p^2 = \frac{106.067 - (2 \times 0.967)}{185.867 + 0.967}$$
$$= \frac{104.133}{186.834}$$
$$= 0.557$$

Note, in Table 16.5, that

$$SS_{total} = 106.067 + 53.333 + 3.267 + 23.2$$
$$= 185.867$$

Likewise, the $\omega^2$ associated with couponing is:

$$\omega_c^2 = \frac{53.333 - (1 \times 0.967)}{185.867 + 0.967}$$
$$= \frac{52.366}{186.834}$$
$$= 0.280$$

As a guide to interpreting $\omega^2$, a large experimental effect produces an $\omega^2$ of 0.15 or greater, a medium effect produces an index of around 0.06, and a small effect produces an index of 0.01.[18] In Table 16.5, although the effect of promotion and couponing are both large, the effect of promotion is much larger. Therefore, in-store promotion will be more effective in increasing sales than couponing.

## Multiple Comparisons

***contrasts***
In ANOVA, a method of examining differences among two or more means of the treatment groups.

***a priori contrasts***
Contrasts that are determined before conducting the analysis, based on the researcher's theoretical framework.

***a posteriori contrasts***
Contrasts made after the analysis. These are generally multiple comparison tests.

***multiple comparison test***
A posteriori contrasts that enable the researcher to construct generalized confidence intervals that can be used to make pairwise comparisons of all treatment means.

The ANOVA $F$ test examines only the overall difference in means. If the null hypothesis of equal means is rejected, we can conclude only that not all of the group means are equal. However, only some of the means may be statistically different and we may wish to examine differences among specific means. This can be done by specifying appropriate ***contrasts,*** or comparisons used to determine which of the means are statistically different. Contrasts may be a priori or a posteriori. ***A priori contrasts*** are determined before conducting the analysis, based on the researcher's theoretical framework. Generally, a priori contrasts are used in lieu of the ANOVA $F$ test. The contrasts selected are orthogonal (they are independent in a statistical sense).

***A posteriori contrasts*** are made after the analysis. These are generally ***multiple comparison tests.*** They enable the researcher to construct generalized confidence intervals that can be used to make pairwise comparisons of all treatment means. These tests, listed in order of decreasing power, include least significant difference, Duncan's multiple range test, Student-Newman-Keuls, Tukey's alternate procedure, honestly significant difference, modified least significant difference, and Scheffe's test. Of these tests, least significant difference is the most powerful, Scheffe's the most

conservative. For further discussion on a priori and a posteriori contrasts, refer to the literature.[19]

Our discussion so far has assumed that each subject is exposed to only one treatment or experimental condition. Sometimes subjects are exposed to more than one experimental condition, in which case repeated measures ANOVA should be used.

## REPEATED MEASURES ANOVA

In marketing research there are often large differences in the background and individual characteristics of respondents. If this source of variability can be separated from treatment effects (effects of the independent variable) and experimental error, then the sensitivity of the experiment can be enhanced. One way of controlling the differences between subjects is by observing each subject under each experimental condition (see Table 16.7). In this sense, each subject serves as its own control. For example, in a survey attempting to determine differences in evaluations of various airlines, each respondent evaluates all the major competing airlines. Because repeated measurements are obtained from each respondent, this design is referred to as within-subjects design or ***repeated measures analysis of variance.*** This differs from the assumption we made in our earlier discussion that each respondent is exposed to only one treatment condition, also referred to as *between-subjects design.*[20] Repeated measures analysis of variance may be thought of as an extension of the paired-samples $t$ test to the case of more than two related samples.

***repeated measures ANOVA***
An ANOVA technique used when respondents are exposed to more than one treatment condition and repeated measurements are obtained.

In the case of a single factor with repeated measures, the total variation, with $n(c-1)$ degrees of freedom, may be split into between-people variation and within-people variation.

$$SS_{total} = SS_{between\ people} + SS_{within\ people}$$

The between-people variation, which is related to the differences between the means of people, has $n-1$ degrees of freedom. The within-people variation has $n(c-1)$ degrees of freedom. The within-people variation may, in turn, be divided into two different sources of variation. One source is related to the differences between treatment means, and

**TABLE 16.7**

**Decomposition of the Total Variation: Repeated Measures ANOVA**

| | Independent Variable X | | | | | |
|---|---|---|---|---|---|---|
| | Subject No. | Categories | | | | Total Sample |
| | | $X_1$ | $X_2$ | $X_3$ | ... $X_c$ | |
| | 1 | $Y_{11}$ | $Y_{12}$ | $Y_{13}$ | $Y_{1c}$ | $Y_1$ |
| | 2 | $Y_{21}$ | $Y_{22}$ | $Y_{23}$ | $Y_{2c}$ | $Y_2$ |
| Between-People Variation $= SS_{between\ people}$ | | • | | | | • |
| | | • | | | | • |
| | | • | | | | • |
| | $n$ | $Y_{n1}$ | $Y_{n2}$ | $Y_{n3}$ | $Y_{nc}$ | $Y_N$ |
| Category Mean | | $\bar{Y}_1$ | $\bar{Y}_2$ | $\bar{Y}_3$ | $\bar{Y}_c$ | $\bar{Y}$ |

Total Variation $= SS_y$

Within-People Variation $= SS_{within\ people}$

the second consists of residual or error variation. The degrees of freedom corresponding to the treatment variation are $c - 1$, and those corresponding to residual variation are $(c - 1)(n - 1)$. Thus,

$$SS_{within\ people} = SS_x + SS_{error}$$

A test of the null hypothesis of equal means may now be constructed in the usual way:

$$F = \frac{SS_x/(c-1)}{SS_{error}/(n-1)(c-1)} = \frac{MS_x}{MS_{error}}$$

So far we have assumed that the dependent variable is measured on an interval or ratio scale. If the dependent variable is nonmetric, however, a different procedure should be used.

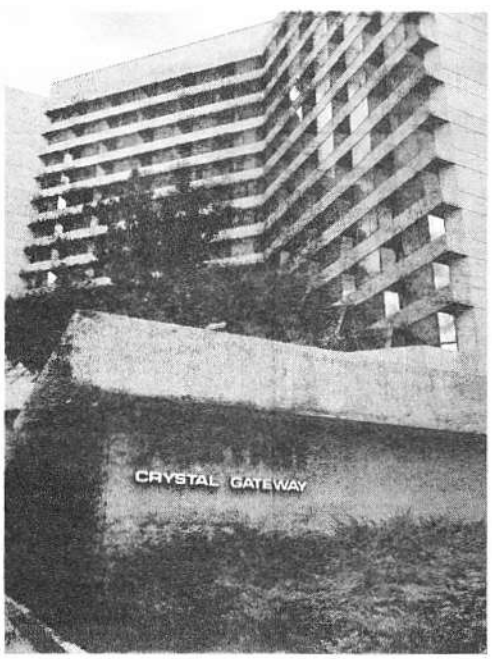
## DECISION RESEARCH

### *Marriott: Luring Business Travelers*

#### The Situation

Marriott International, Inc. is a leading worldwide hospitality company. Its heritage can be traced to a small root beer stand opened in Washington, D.C., in 1927 by J. Willard and Alice S. Marriott. As of June 17, 2005, the company operated or franchised 2,676 lodging properties located in the United States and 63 other countries and territories. Among Marriott's most frequent visitors are its business traveler customers. For many years, business travelers have faced a fundamental problem—figuring out a comfortable and convenient way to get their jobs accomplished in hotel rooms without a functional workspace. Though many were not very productive, they did hone important skills such as writing legibly on top of a comforter, stretching arms beyond maximum length to reach hidden outlets behind or underneath furniture, and squinting tightly to make documents readable.

Marriott recognized these needs of its business travelers and wanted to do something about it. Susan Hodapp, brand director for Marriott Hotels, Resorts, and Suites, commis-




Analysis of variance techniques can help Marriott determine what features of a hotel room are most important to business travelers.

sioned a survey to determine business travelers' preferences for hotels and the factors that are important in their hotel selection process. Part of the questionnaire focused attention on the features of the hotel room. Each respondent rated the relative importance of the following factors on a 7-point scale (1 = not at all important, 7 = extremely important): Room décor, room lighting, room furniture, voice and data access in the room, and price of the room per night.

### The Marketing Research Decision

1. Marriott would like to determine what features of a hotel room are most important in business travelers' choice of a hotel. Each feature could be offered at several levels, for example, room lighting could be bright, medium, or dim. What analysis should be conducted?
2. Discuss the role of the type of data analysis you recommend in enabling Susan Hodapp to understand business travelers' preferences for hotel rooms.

### The Marketing Management Decision

1. What should Marriott do in order to lure the business travelers? What features should it stress?
2. Discuss how the marketing management decision action that you recommend to Susan Hodapp is influenced by the type of data analysis that you suggested earlier and by the findings of that analysis.[21] ■

## NONMETRIC ANALYSIS OF VARIANCE

***nonmetric ANOVA***
An ANOVA technique for examining the difference in the central tendencies of more than two groups when the dependent variable is measured on an ordinal scale.

***k-sample median test***
Nonparametric test that is used to examine differences among groups when the dependent variable is measured on an ordinal scale.

***Kruskal-Wallis one-way analysis of variance***
A nonmetric ANOVA test that uses the rank value of each case, not merely its location relative to the median.

***Nonmetric analysis of variance*** examines the difference in the central tendencies of more than two groups when the dependent variable is measured on an ordinal scale. One such procedure is the ***k-sample median test.*** As its name implies, this is an extension of the median test for two groups, which was considered in Chapter 15. The null hypothesis is that the medians of the $k$ populations are equal. The test involves the computation of a common median over the $k$ samples. Then a $2 \times k$ table of cell counts based on cases above or below the common median is generated. A chi-square statistic is computed. The significance of the chi-square implies a rejection of the null hypothesis.

A more powerful test is the ***Kruskal-Wallis one-way analysis of variance.*** This is an extension of the Mann-Whitney test (Chapter 15). This test also examines the difference in medians. The null hypothesis is the same as in the $k$-sample median test, but the testing procedure is different. All cases from the $k$ groups are ordered in a single ranking. If the $k$ populations are the same, the groups should be similar in terms of ranks within each group. The rank sum is calculated for each group. From these, the Kruskal-Wallis $H$ statistic, which has a chi-square distribution, is computed.

The Kruskal-Wallis test is more powerful than the $k$-sample median test because it uses the rank value of each case, not merely its location relative to the median. However, if there are a large number of tied rankings in the data, the $k$-sample median test may be a better choice.

Nonmetric analysis of variance is not popular in commercial marketing research. Another procedure, which is also only rarely used, is multivariate analysis of variance.

## MULTIVARIATE ANALYSIS OF VARIANCE

***multivariate analysis of variance (MANOVA)***
An ANOVA technique using two or more metric dependent variables.

***Multivariate analysis of variance*** (MANOVA) is similar to analysis of variance (ANOVA), except that instead of one metric dependent variable, we have two or more. The objective is the same; MANOVA is also concerned with examining differences between groups. Whereas ANOVA examines group differences on a single dependent variable,

MANOVA examines group differences across multiple dependent variables simultaneously. In ANOVA, the null hypothesis is that the means of the dependent variable are equal across the groups. In MANOVA, the null hypothesis is that the vectors of means on multiple dependent variables are equal across groups. Multivariate analysis of variance is appropriate when there are two or more dependent variables that are correlated. If there are multiple dependent variables that are uncorrelated or orthogonal, ANOVA on each of the dependent variables is more appropriate than MANOVA.[22]

As an example, suppose that four groups, each consisting of 100 randomly selected individuals, were exposed to four different commercials about Tide detergent. After seeing the commercial, each individual provided ratings on preference for Tide, preference for Procter & Gamble (the company marketing Tide), and preference for the commercial itself. Because these three preference variables are correlated, multivariate analysis of variance should be conducted to determine which commercial is the most effective (produced the highest preference across the three preference variables). The next example illustrates the application of ANOVA and MANOVA in international marketing research, and the example after that shows an application of these techniques in examining ethics in marketing research.

**REAL RESEARCH**

## *The Commonality of Unethical Research Practices Worldwide*

As of 2006, mass media is continuing to focus more attention on the highly visible practices of unethical marketing research, and this poses a serious threat to marketing research practitioners. A study examined marketing professionals' perceptions of the commonality of unethical marketing research practices on a cross-national basis. The sample of marketing professionals was drawn from Australia, Canada, Great Britain, and the United States.

Respondents' evaluations were analyzed using computer programs for MANOVA and ANOVA. Country of respondent comprised the predictor variable in the analysis, and 15 commonality evaluations served as the criterion variables. The $F$ values from the ANOVA analyses indicated that only two of the 15 commonality evaluations achieved significance ($p < 0.05$ or better). Further, the MANOVA $F$ value was not statistically significant, implying the lack of overall differences in commonality evaluations across respondents of the four countries. Therefore, it was concluded that marketing professionals in the four countries evince similar perceptions of the commonality of unethical research practices. This finding is not surprising, given research evidence that organizations in the four countries reflect similar corporate cultures. Thus, the marketing research industry in these four countries should adopt a common platform in fighting unethical practices.[23] ■

**REAL RESEARCH**

## *"MAN"OVA Demonstrates That Man Is Different from Woman*

In order to investigate differences between research ethics judgments in men and women, the statistical techniques of MANOVA and ANOVA were used. Respondents were asked to indicate their degree of approval with regard to a series of scenarios involving decisions of an ethical nature. These evaluations served as the dependent variable in the analysis, and sex of the respondent served as the independent variable. MANOVA was used for multivariate analysis and its resultant $F$ value was significant at the $p < 0.001$ level—indicating that there was an "overall" difference between males and females in research ethics judgments. Univariate analysis was conducted via ANOVA, and $F$ values indicated that three items were the greatest contributors to the overall gender difference in ethical evaluations: the use of ultraviolet ink to precode a mail questionnaire, the use of an ad that encourages consumer misuse of a product, and unwillingness by a researcher to offer data help to an inner-city advisory group. Another

recent study examined how ethical beliefs are related to age and gender of business professionals. The results of this particular study indicated that overall, younger business professionals exhibited a lower standard of ethical beliefs. In the younger age group, females demonstrated a higher level of ethical beliefs compared to males. However, in the older age group, results showed that males had a slightly higher level of ethical beliefs. Thus, companies should emphasize ethical values and training to the younger professionals, especially men.[24] ■

## STATISTICAL SOFTWARE

The major computer packages (SPSS and SAS) have programs for conducting analysis of variance and covariance available in the microcomputer and mainframe versions. In addition to the basic analysis that we have considered, these programs can also perform more complex analysis. Minitab and Excel also offer some programs. Exhibit 16.1 contains

*Exhibit 16.1*
Computer Programs for ANOVA and ANCOVA

Given the importance of analysis of variance and covariance, several programs are available in each package.

***SPSS***

One-way ANOVA can be efficiently performed using the program ONEWAY. This program also allows the user to test a priori and a posteriori contrasts, which cannot be done in other SPSS programs. For performing $n$-way analysis of variance, the program ANOVA can be used. Although covariates can be specified, ANOVA does not perform a full analysis of covariance. For comprehensive analysis of variance or analysis of covariance, including repeated measures and multiple dependent measures, the MANOVA procedure is recommended. For nonmetric analysis of variance, including the $k$-sample median test and Kruskal-Wallis one-way analysis of variance, the program NPAR TESTS should be used.

***SAS***

The main program for performing analysis of variance in the case of a balanced design is ANOVA. This program can handle data from a wide variety of experimental designs, including multivariate analysis of variance and repeated measures. Both a priori and a posteriori contrasts can be tested. For unbalanced designs, the more general GLM procedure can be used. This program performs analysis of variance, analysis of covariance, repeated measures analysis of variance, and multivariate analysis of variance. It also allows the testing of a priori and a posteriori contrasts. Although GLM can also be used for analyzing balanced designs, it is not as efficient as ANOVA for such models. The VARCOMP procedure computes variance components. For nonmetric analysis of variance, the NPAR1WAY procedure can be used. For constructing designs and randomized plans, the PLAN procedure can be used.

***Minitab***

Analysis of variance and covariance can be assessed from the Stats>ANOVA function. This function performs one-way ANOVA, two-way ANOVA, analysis of means, balanced ANOVA, analysis of covariance, general linear model, main effects plot, interactions plot, and residual plots. In order to compute the mean and standard deviation, the CROSSTAB function must be used. To obtain $F$ and $p$ values, use the balanced ANOVA.

***Excel***

Both a one-way ANOVA and two-way ANOVA can be performed under the Tools>DATA ANALYSIS function. The two-way ANOVA has the features of a two-factor with replication and a two-factor without replication. The two-factor with replication includes more than one sample for each group of data. The two-factor without replication does not include more than one sampling per group.

a description of the relevant programs. Refer to the user manuals for these packages for more details.

## SPSS WINDOWS

One-way ANOVA can be efficiently performed using the program COMPARE MEANS and then ONE-WAY ANOVA. To select this procedure using SPSS for Windows, click:

Analyze>Compare Means>One-Way ANOVA . . .

The following are the detailed steps for running a one-way ANOVA on the data of Table 16.2. The corresponding screen captures for these steps can be downloaded from the Web site for this book. The null hypothesis is that there is no difference in mean normalized sales for the three levels of in-store promotion.

1. Select ANALYZE from the SPSS menu bar.
2. Click COMPARE MEANS and then ONE-WAY ANOVA.
3. Move "Sales [sales]" into the DEPENDENT LIST box.
4. Move "In-Store Promotion [promotion]" to the FACTOR box.
5. Click OPTIONS.
6. Click Descriptive.
7. Click CONTINUE.
8. Click OK.

*N*-way analysis of variance, analysis of covariance, MANOVA, and repeated measures ANOVA can be performed using GENERAL LINEAR MODEL. To select this procedure using SPSS for Windows, click:

Analyze>General Linear Model>Univariate . . .
Analyze>General Linear Model>Multivariate . . .
Analyze>General Linear Model>Repeated Measures . . .

We show the detailed steps for performing the analysis of covariance given in Table 16.6.

1. Select ANALYZE from the SPSS menu bar.
2. Click GENERAL LINEAR MODEL and then UNIVARIATE.
3. Move "Sales [sales]" into the DEPENDENT VARIABLE box.
4. Move "In-Store Promotion [promotion]" to the FIXED FACTOR(S) box. Then move "Coupon [coupon]" to the FIXED FACTOR(S) box.
5. Move "Clientele [clientele]" to the COVARIATE(S) box.
6. Click OK.

For nonmetric analysis of variance, including the *k*-sample median test and Kruskal-Wallis one way analysis of variance, the program Nonparametric Tests should be used.

Analyze>Nonparametric Tests>K Independent Samples . . .
Analyze>Nonparametric Tests>K Related Samples . . .

The detailed steps for the other procedures are similar to those shown and are not given here due to space constraints.

## PROJECT RESEARCH

### *Analysis of Variance*

In the department store project, several independent variables were examined as categorical variables having more than two categories. For example, familiarity with the department stores considered was respecified as high, medium, or low. The effects of these independent variables on metric dependent variables were examined using analysis of variance procedures. Several useful insights were obtained that guided subsequent data analysis and interpretation. For example, a three-category respecification of familiarity produced results that were not significant, whereas treating familiarity as a binary variable (high or low) produced significant results. This, along with the frequency distribution, indicated that treating familiarity as having only two categories was most appropriate.


**SPSS Data File**

#### Project Activities

Download the SPSS data file Sears Data 14 from the Web site for this book. See Chapter 14 for a description of this file.

1. Run separate one-way ANOVAs to determine which recoded demographic variables explain the overall familiarity score.

## EXPERIENTIAL RESEARCH


**SPSS Data File**

Download the Dell case and questionnaire from the Web site for this book. This information is also given at the end of the book. Download the Dell SPSS data file.

1. Are the three price-sensitive groups based on q9_5per as derived in Chapter 14 different in terms of each of the evaluations of Dell (q8_1 to q8_13)? Interpret the results.
2. Are the three price-sensitive groups based on q9_10per as derived in Chapter 14 different in terms of each of the evaluations of Dell (q8_1 to q8_13)? Interpret the results.
3. Do the demographic groups as recoded in Chapter 14 (recoded q11, q12, q13) and q14 differ in terms of overall satisfaction with Dell computers (q4)? Interpret the results.
4. Do the demographic groups as recoded in Chapter 14 (recoded q11, q12, q13) and q14 differ in terms of likelihood of recommending Dell computers (q5)? Interpret the results.
5. Do the demographic groups as recoded in Chapter 14 (recoded q11, q12, q13) and q14 differ in terms of likelihood of choosing Dell computers (q6)? Interpret the results. ■

# SUMMARY

In ANOVA and ANCOVA, the dependent variable is metric and the independent variables are all categorical, or combinations of categorical and metric variables. One-way ANOVA involves a single independent categorical variable. Interest lies in testing the null hypothesis that the category means are equal in the population. The total variation in the dependent variable is decomposed into two components: variation related to the independent variable and variation related to error. The variation is measured in terms of the sum of squares corrected for the mean (*SS*). The mean square is obtained by dividing the *SS* by the corresponding degrees of freedom (df). The null hypothesis of equal means is tested by an *F* statistic, which is the ratio of the mean square related to the independent variable to the mean square related to error.

*N*-way analysis of variance involves the simultaneous examination of two or more categorical independent variables. A major advantage is that the interactions between the independent variables can be examined. The significance

of the overall effect, interaction terms, and main effects of individual factors are examined by appropriate $F$ tests. It is meaningful to test the significance of main effects only if the corresponding interaction terms are not significant.

ANCOVA includes at least one categorical independent variable and at least one interval or metric independent variable. The metric independent variable, or covariate, is commonly used to remove extraneous variation from the dependent variable.

When analysis of variance is conducted on two or more factors, interactions can arise. An interaction occurs when the effect of an independent variable on a dependent variable is different for different categories or levels of another independent variable. If the interaction is significant, it may be ordinal or disordinal. Disordinal interaction may be of a noncrossover or crossover type. In balanced designs, the relative importance of factors in explaining the variation in the dependent variable is measured by omega squared ($\omega^2$). Multiple comparisons in the form of a priori or a posteriori contrasts can be used for examining differences among specific means.

In repeated measures analysis of variance, observations on each subject are obtained under each treatment condition. This design is useful for controlling for the differences in subjects that exist prior to the experiment. Nonmetric analysis of variance involves examining the differences in the central tendencies of two or more groups when the dependent variable is measured on an ordinal scale. Multivariate analysis of variance (MANOVA) involves two or more metric dependent variables.

## KEY TERMS AND CONCEPTS

analysis of variance (ANOVA), *505*
factors, *505*
treatment, *505*
one-way analysis of variance, *505*
$n$-way analysis of variance, *505*
analysis of covariance (ANCOVA), *505*
covariate, *506*
eta$^2$ ($\eta^2$), *507*
$F$ statistic, *507*
mean square, *507*
$SS_{between}$ ($SS_x$), *507*
$SS_{within}$ ($SS_{error}$), *507*
$SS_y$, *507*
decomposition of the total variation, *508*
interaction, *515*
multiple $\eta^2$, *515*
significance of the overall effect, *515*
significance of the interaction effect, *516*
significance of the main effect, *516*
ordinal interaction, *520*
disordinal interaction, *520*
omega squared ($\omega^2$), *522*
contrasts, *522*
a priori contrasts, *522*
a posteriori contrasts, *522*
multiple comparison tests, *522*
repeated measures ANOVA, *523*
nonmetric ANOVA, *525*
$k$-sample median test, *525*
Kruskal-Wallis one-way analysis of variance, *525*
multivariate analysis of variance (MANOVA), *525*

## SUGGESTED CASES, VIDEO CASES, AND HBS CASES

### *Cases*

Case 3.1 Is Celebrity Advertising Worth Celebrating?
Case 3.2 The Demographic Discovery of the New Millennium
Case 3.3 Matsushita Retargets the U.S.A.
Case 3.4 Pampers Curing Its Rash of Market Share
Case 3.6 Cingular Wireless: A Singular Focus
Case 3.7 IBM: The World's Top Provider of Computer Hardware, Software, and Services
Case 3.8 Kimberly-Clark: Competing Through Innovation
Case 4.1 Wachovia: "Watch Ovah Ya" Finances
Case 4.2 Wendy's: History and Life After Dave Thomas
Case 4.3 Astec: Continuing to Grow
Case 4.4 Is Marketing Research the Cure for Norton Healthcare Kosair Children's Hospital's Ailments?

### *Video Cases*

Video Case 3.1 The Mayo Clinic: Staying Healthy with Marketing Research
Video Case 4.1 Subaru: "Mr. Survey" Monitors Customer Satisfaction
Video Case 4.2 Procter & Gamble: Using Marketing Research to Build Brands

# LIVE RESEARCH: CONDUCTING A MARKETING RESEARCH PROJECT

1. Differences between groups are of interest in most projects. In case of two groups, these can be examined by using independent samples $t$ tests for two groups or one-way ANOVA for more than two groups.

# ACRONYMS

The major characteristics of analysis of variance may be described by the acronym ANOVA:

**A** nalysis of total variation
**N** ormally distributed errors which are uncorrelated
**O** ne or more categorical independent variables with fixed categories
**V** ariance is assumed to be constant
**A** single dependent variable which is metric

The major characteristics of analysis of covariance may be summarized by the acronym ANCOVA:

**A** nalysis of total variation
**N** ormally distributed errors which are uncorrelated
**C** ovariates: one or more metric independent variables
**O** ne or more categorical independent variables with fixed categories
**V** ariance is assumed to be constant
**A** single dependent variable which is metric

# EXERCISES

## *Questions*

1. Discuss the similarities and differences between analysis of variance and analysis of covariance.
2. What is the relationship between analysis of variance and the $t$ test?
3. What is total variation? How is it decomposed in a one-way analysis of variance?
4. What is the null hypothesis in one-way ANOVA? What basic statistic is used to test the null hypothesis in one-way ANOVA? How is this statistic computed?
5. How does $n$-way analysis of variance differ from the one-way procedure?
6. How is the total variation decomposed in $n$-way analysis of variance?
7. What is the most common use of the covariate in ANCOVA?
8. Define an interaction.
9. What is the difference between ordinal and disordinal interaction?
10. How is the relative importance of factors measured in a balanced design?
11. What is an a priori contrast?
12. What is the most powerful test for making a posteriori contrasts? Which test is the most conservative?
13. What is meant by repeated measures ANOVA? Describe the decomposition of variation in repeated measures ANOVA.
14. What are the differences between metric and nonmetric analyses of variance?
15. Describe two tests used for examining differences in central tendencies in nonmetric ANOVA.
16. What is multivariate analysis of variance? When is it appropriate?

## *Problems*

1. After receiving some complaints from the readers, your campus newspaper decides to redesign its front page. Two new formats, B and C, were developed and tested against the current format, A. A total of 75 students were randomly selected and 25 students were randomly assigned to each of three format conditions. The students were asked to evaluate the effectiveness of the format on an 11-point scale (1 = poor, 11 = excellent).
   a. State the null hypothesis.
   b. What statistical test should you use?
   c. What are the degrees of freedom associated with the test statistic?
2. A marketing researcher wants to test the hypothesis that, in the population, there is no difference in the importance attached to shopping by consumers living in the northern, southern, eastern, and western United States. A study is conducted and analysis of variance is used to analyze the data. The results obtained are presented in the following table.

| *Source* | *df* | *Sum of Squares* | *Mean Squares* | F *Ratio* | F *Probability* |
|---|---|---|---|---|---|
| Between groups | 3 | 70.212 | 23.404 | 1.12 | 0.3 |
| Within groups | 996 | 20812.416 | 20.896 | | |

   a. Is there sufficient evidence to reject the null hypothesis?
   b. What conclusion can be drawn from the table?
   c. If the average importance were computed for each group, would you expect the sample means to be similar or different?
   d. What was the total sample size in this study?

3. In a pilot study examining the effectiveness of three commercials (A, B, and C), 10 consumers were assigned to view each commercial and rate it on a 9-point Likert scale. The data obtained from the 30 respondents are shown in the table.

| *Commercial* | | | *Commercial* | | |
|---|---|---|---|---|---|
| *A* | *B* | *C* | *A* | *B* | *C* |
| 4 | 7 | 8 | 4 | 6 | 7 |
| 5 | 4 | 7 | 4 | 5 | 8 |
| 3 | 6 | 7 | 3 | 5 | 8 |
| 4 | 5 | 6 | 5 | 4 | 5 |
| 3 | 4 | 8 | 5 | 4 | 6 |

   a. Calculate the category means and the grand mean.
   b. Calculate $SS_y$, $SS_x$, and $SS_{error}$.
   c. Calculate $\eta^2$.
   d. Calculate the value of $F$.
   e. Are the three commercials equally effective?

4. An experiment tested the effects of package design and shelf display on the likelihood of purchase of Product 19 cereal. Package design and shelf display were varied at two levels each, resulting in a 2 × 2 design. Purchase likelihood was measured on a 7-point scale. The results are partially described in the following table.

| *Source of Variation* | *Sum of Squares* | *df* | *Mean Square* | F | *Significance of* F | $\omega^2$ |
|---|---|---|---|---|---|---|
| Package design | 68.76 | 1 | | | | |
| Shelf display | 320.19 | 1 | | | | |
| Two-way interaction | 55.05 | 1 | | | | |
| Residual error | 176.00 | 40 | | | | |

   a. Complete the table by calculating the mean square, $F$, significance of $F$, and $\omega^2$ values.
   b. How should the main effects be interpreted?

SPSS Data File

## INTERNET AND COMPUTER EXERCISES

1. Analyze the Nike data given in Internet and Computer Exercises 1 of Chapter 15. Do the three usage groups differ in terms of awareness, attitude, preference, intention, and loyalty toward Nike when these variables are considered individually, i.e., one at a time?
2. Conduct the following analyses for the outdoor lifestyle data given in Internet and Computer Exercises 2 of Chapter 15.
   a. Do the three groups based on location of residence differ in their preference for an outdoor lifestyle?
   b. Do the three groups based on location of residence differ in terms of the importance attached to enjoying nature?
   c. Do the three groups based on location of residence differ in terms of the importance attached to living in harmony with the environment?
   d. Do the three groups based on location of residence differ in terms of the importance attached to exercising regularly?
3. In an experiment designed to measure the effect of sex and frequency of travel on preference for foreign travel a 2 (sex) × 3 (frequency of travel) between-subjects design was adopted. Five respondents were assigned to each cell for a total sample size of 30. Preference for foreign travel was measured on a 9-point scale (1 = no preference, 9 = strong preference). Sex was coded as male = 1 and female = 2. Frequency of travel was coded as light = 1, medium = 2, and heavy = 3. The data obtained are shown here.

| *Number* | *Sex* | *Travel Group* | *Preference* |
|---|---|---|---|
| 1 | 1 | 1 | 2 |
| 2 | 1 | 1 | 3 |
| 3 | 1 | 1 | 4 |
| 4 | 1 | 1 | 4 |
| 5 | 1 | 1 | 2 |
| 6 | 1 | 2 | 4 |
| 7 | 1 | 2 | 5 |
| 8 | 1 | 2 | 5 |
| 9 | 1 | 2 | 3 |
| 10 | 1 | 2 | 3 |
| 11 | 1 | 3 | 8 |
| 12 | 1 | 3 | 9 |
| 13 | 1 | 3 | 8 |
| 14 | 1 | 3 | 7 |
| 15 | 1 | 3 | 7 |
| 16 | 2 | 1 | 6 |
| 17 | 2 | 1 | 7 |
| 18 | 2 | 1 | 6 |
| 19 | 2 | 1 | 5 |
| 20 | 2 | 1 | 7 |
| 21 | 2 | 2 | 3 |
| 22 | 2 | 2 | 4 |
| 23 | 2 | 2 | 5 |
| 24 | 2 | 2 | 4 |
| 25 | 2 | 2 | 5 |
| 26 | 2 | 3 | 6 |
| 27 | 2 | 3 | 6 |
| 28 | 2 | 3 | 6 |
| 29 | 2 | 3 | 7 |
| 30 | 2 | 3 | 8 |

   Using software of your choice, perform the following analysis.
   a. Do the males and the females differ in their preference for foreign travel?
   b. Do the light, medium, and heavy travelers differ in their preference for foreign travel?
   c. Conduct a 2 × 3 analysis of variance with preference for foreign travel as the dependent variable and sex and travel frequency as the independent variables or factors. Interpret the results.
4. Using the appropriate microcomputer and mainframe programs in the package of your choice (SPSS, SAS, Minitab, or Excel), analyze the data collected in Fieldwork assignment 1. Should the campus newspaper change the format of the cover page? What is your conclusion?

# ACTIVITIES

## *Fieldwork*

1. Contact your campus newspaper. Collect data for the experiment described in problem 1. Because this may be too much work for one student, this project may be handled in teams of three.

## *Group Discussion*

1. Which procedure is more useful in marketing research—analysis of variance or analysis of covariance? Discuss as a small group.

CHAPTER 17

# Correlation and Regression

"Correlation is a simple but powerful way to look at the linear relationship between two metric variables. Multiple regression extends this concept, enabling the researcher to examine the relationship between one variable and several others."

*Jim McGee, mission research specialist, Global Mapping International*

## Objectives

After reading this chapter, the student should be able to:

1. Discuss the concepts of product moment correlation, partial correlation, and part correlation and show how they provide a foundation for regression analysis.
2. Explain the nature and methods of bivariate regression analysis and describe the general model, estimation of parameters, standardized regression coefficient, significance testing, prediction accuracy, residual analysis, and model cross-validation.
3. Explain the nature and methods of multiple regression analysis and the meaning of partial regression coefficients.
4. Describe specialized techniques used in multiple regression analysis, particularly stepwise regression, regression with dummy variables, and analysis of variance and covariance with regression.
5. Discuss nonmetric correlation and measures such as Spearman's rho and Kendall's tau.

## Overview

Chapter 16 examined the relationship among the *t* test, analysis of variance and covariance, and regression. This chapter describes regression analysis, which is widely used for explaining variation in market share, sales, brand preference, and other marketing results in terms of marketing management variables such as advertising, price, distribution, and product quality. However, before discussing regression, we describe the concepts of product moment correlation and partial correlation coefficient, which lay the conceptual foundation for regression analysis.

In introducing regression analysis, we discuss the simple bivariate case first. We describe estimation, standardization of the regression coefficients, testing and examination of the strength and significance of association between variables, prediction accuracy, and the assumptions underlying the regression model. Next, we discuss the multiple regression model, emphasizing the interpretation of parameters, strength of association, significance tests, and examination of residuals.

Then we cover topics of special interest in regression analysis, such as stepwise regression, multicollinearity, relative importance of predictor variables, and cross-validation. We describe regression with dummy variables and the use of this procedure to conduct analysis of variance and covariance.

### REAL RESEARCH

### *Regression Rings the Right Bell for Avon*

Avon Products, Inc. (*www.avon.com*), was having significant problems with the sales staff. The company's business, dependent on sales representatives, was facing a shortage of sales reps without much hope of getting new ones. Regression models, operating on microcomputers, were developed to reveal the possible variables that were fueling this


Good products, well-trained sales reps, and sophisticated regression models have opened the doors for Avon, enabling it to penetrate the cosmetics market and become the world's largest direct seller.

situation. The models revealed that the most significant variable was the level of the appointment fee that reps pay for materials and second was the employee benefits. With data to back up its actions, the company lowered the fee. The company also hired senior manager Michele Schneider to improve the way Avon informed new hires of their employee benefits program. Schneider revamped Avon's benefits program information packet, which yielded an informative and easy to navigate "Guide to Your Personal Benefits." These changes resulted in an improvement in the recruitment and retention of sales reps. As of 2006, Avon was the world's largest direct seller, selling products in over 100 countries.[1] ■

**REAL RESEARCH**

### *Retailing Revolution*

Many retailing experts suggest that electronic shopping will be the next revolution in retailing. Whereas many traditional retailers experienced sluggish, single-digit sales growth in the early 2000s, online sales records were off the charts. Although e-tailing continues to make up a very small portion of overall retail sales (less than 3 percent in 2006), the trend looks very promising for the future. A research project investigating this trend looked for correlates of consumers' preferences for electronic shopping services via home videotext (computerized in-home shopping services). The explanation of consumers' preferences was sought in psychographic, demographic, and communication variables suggested in the literature.

Multiple regression was used to analyze the data. The overall multiple regression model was significant at the 0.05 level. Univariate *t* tests indicated that the following variables in the model were significant at the 0.05 level or better: price orientation, sex, age, occupation, ethnicity, and education. None of the three communication variables (mass media, word of mouth, and publicity) was significantly related to consumer preference, the dependent variable.

The results suggest that electronic shopping is preferred by white females who are older, better educated, working in supervisory or higher level occupations, and price-oriented shoppers. Information of this type is valuable in targeting marketing effort to electronic shoppers.[2] ■

These examples illustrate some of the uses of regression analysis in determining which independent variables explain a significant variation in the dependent variable of interest, the structure and form of the relationship, the strength of the relationship, and predicted values of the dependent variable. Fundamental to regression analysis is an understanding of the product moment correlation.

## PRODUCT MOMENT CORRELATION

In marketing research we are often interested in summarizing the strength of association between two metric variables, as in the following situations:

- How strongly are sales related to advertising expenditures?
- Is there an association between market share and size of the sales force?
- Are consumers' perceptions of quality related to their perceptions of prices?

***product moment correlation (r)*** A statistic summarizing the strength of association between two metric variables.

In situations like these, the ***product moment correlation,*** **r,** is the most widely used statistic, summarizing the strength of association between two metric (interval or ratio scaled) variables, say *X* and *Y*. It is an index used to determine whether a linear, or straight-line, relationship exists between *X* and *Y*. It indicates the degree to which the variation in one variable, *X*, is related to the variation in another variable, *Y*. Because it was originally proposed by Karl Pearson, it is also known as the *Pearson correlation coefficient.* It is also referred to as *simple correlation, bivariate correlation,* or merely the

*correlation coefficient*. From a sample of $n$ observations, $X$ and $Y$, the product moment correlation, $r$, can be calculated as:

$$r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}}$$

Division of the numerator and denominator by $n - 1$ gives

$$r = \frac{\sum_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}}{\sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} \sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{n-1}}}$$

$$= \frac{COV_{xy}}{S_x S_y}$$

***covariance***
A systematic relationship between two variables in which a change in one implies a corresponding change in the other ($COV_{xy}$).

In these equations, $\bar{X}$ and $\bar{Y}$ denote the sample means, and $S_x$ and $S_y$ the standard deviations. $COV_{xy}$, the ***covariance*** between $X$ and $Y$, measures the extent to which $X$ and $Y$ are related. The covariance may be either positive or negative. Division by $S_x S_y$ achieves standardization, so that $r$ varies between $-1.0$ and $+1.0$. Note that the correlation coefficient is an absolute number and is not expressed in any unit of measurement. The correlation coefficient between two variables will be the same regardless of their underlying units of measurement.

As an example, suppose a researcher wants to explain attitudes toward a respondent's city of residence in terms of duration of residence in the city. The attitude is measured on an 11-point scale (1 = do not like the city, 11 = very much like the city), and the duration of residence is measured in terms of the number of years the respondent has lived in the city. In a pretest of 12 respondents, the data shown in Table 17.1 are obtained. For illustrative purposes, we consider only a small number of observations. In actual practice, correlation and regression are performed on a much larger sample such as that in the Dell Experiential Research considered later.


**SPSS Data File**

**TABLE 17.1**
**Explaining Attitude Toward the City of Residence**

| Respondent No. | Attitude Toward the City | Duration of Residence | Importance Attached to Weather |
|---|---|---|---|
| 1 | 6 | 10 | 3 |
| 2 | 9 | 12 | 11 |
| 3 | 8 | 12 | 4 |
| 4 | 3 | 4 | 1 |
| 5 | 10 | 12 | 11 |
| 6 | 4 | 6 | 1 |
| 7 | 5 | 8 | 7 |
| 8 | 2 | 2 | 4 |
| 9 | 11 | 18 | 8 |
| 10 | 9 | 9 | 10 |
| 11 | 10 | 17 | 8 |
| 12 | 2 | 2 | 5 |

The correlation coefficient may be calculated as follows:

$$\bar{X} = \frac{(10 + 12 + 12 + 4 + 12 + 6 + 8 + 2 + 18 + 9 + 17 + 2)}{12}$$

$$= 9.333$$

$$\bar{Y} = \frac{(6 + 9 + 8 + 3 + 10 + 4 + 5 + 2 + 11 + 9 + 10 + 2)}{12}$$

$$= 6.583$$

$$\begin{aligned}\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) &= (10 - 9.33)(6 - 6.58) + (12 - 9.33)(9 - 6.58) \\ &\quad + (12 - 9.33)(8 - 6.58) + (4 - 9.33)(3 - 6.58) \\ &\quad + (12 - 9.33)(10 - 6.58) + (6 - 9.33)(4 - 6.58) \\ &\quad + (8 - 9.33)(5 - 6.58) + (2 - 9.33)(2 - 6.58) \\ &\quad + (18 - 9.33)(11 - 6.58) + (9 - 9.33)(9 - 6.58) \\ &\quad + (17 - 9.33)(10 - 6.58) + (2 - 9.33)(2 - 6.58) \\ &= -0.3886 + 6.4614 + 3.7914 + 19.0814 \\ &\quad + 9.1314 + 8.5914 + 2.1014 + 33.5714 \\ &\quad + 38.3214 - 0.7986 + 26.2314 + 33.5714 \\ &= 179.6668\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n} (X_i - \bar{X})^2 &= (10 - 9.33)^2 + (12 - 9.33)^2 + (12 - 9.33)^2 + (4 - 9.33)^2 \\ &\quad + (12 - 9.33)^2 + (6 - 9.33)^2 + (8 - 9.33)^2 + (2 - 9.33)^2 \\ &\quad + (18 - 9.33)^2 + (9 - 9.33)^2 + (17 - 9.33)^2 + (2 - 9.33)^2 \\ &= 0.4489 + 7.1289 + 7.1289 + 28.4089 \\ &\quad + 7.1289 + 11.0889 + 1.7689 + 53.7289 \\ &\quad + 75.1689 + 0.1089 + 58.8289 + 53.7289 \\ &= 304.6668\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n} (Y_i - \bar{Y})^2 &= (6 - 6.58)^2 + (9 - 6.58)^2 + (8 - 6.58)^2 + (3 - 6.58)^2 \\ &\quad + (10 - 6.58)^2 + (4 - 6.58)^2 + (5 - 6.58)^2 + (2 - 6.58)^2 \\ &\quad + (11 - 6.58)^2 + (9 - 6.58)^2 + (10 - 6.58)^2 + (2 - 6.58)^2 \\ &= 0.3364 + 5.8564 + 2.0164 + 12.8164 \\ &\quad + 11.6964 + 6.6564 + 2.4964 + 20.9764 \\ &\quad + 19.5364 + 5.8564 + 11.6964 + 20.9764 \\ &= 120.9168\end{aligned}$$

Thus,

$$r = \frac{179.6668}{\sqrt{(304.6668)(120.9168)}}$$

$$= 0.9361$$

In this example, $r = 0.9361$, a value close to 1.0. This means that respondents' duration of residence in the city is strongly associated with their attitude toward the city. Furthermore, the positive sign of $r$ implies a positive relationship; the longer the duration of residence, the more favorable the attitude and vice versa.

Because $r$ indicates the degree to which variation in one variable is related to variation in another, it can also be expressed in terms of the decomposition of the total variation (see Chapter 16). In other words,

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$
$$= \frac{SS_x}{SS_y}$$
$$= \frac{\text{Total variation} - \text{Error variation}}{\text{Total variation}}$$
$$= \frac{SS_y - SS_{error}}{SS_y}$$

Hence, $r^2$ measures the proportion of variation in one variable that is explained by the other. Both $r$ and $r^2$ are symmetric measures of association. In other words, the correlation of $X$ with $Y$ is the same as the correlation of $Y$ with $X$. It does not matter which variable is considered to be the dependent variable and which the independent. The product moment coefficient measures the strength of the linear relationship and is not designed to measure nonlinear relationships. Thus, $r = 0$ merely indicates that there is no linear relationship between $X$ and $Y$. It does not mean that $X$ and $Y$ are unrelated. There could well be a nonlinear relationship between them, which would not be captured by $r$ (see Figure 17.1).

When it is computed for a population rather than a sample, the product moment correlation is denoted by ρ, the Greek letter rho. The coefficient $r$ is an estimator of ρ. Note that the calculation of $r$ assumes that $X$ and $Y$ are metric variables whose distributions have the same shape. If these assumptions are not met, $r$ is deflated and underestimates ρ. In marketing research, data obtained by using rating scales with a small number of categories may not be strictly interval. This tends to deflate $r$, resulting in an underestimation of ρ.[3]


**SPSS Output File**

The statistical significance of the relationship between two variables measured by using $r$ can be conveniently tested. The hypotheses are:

$$H_0: \rho = 0$$
$$H_1: \rho \neq 0$$

The test statistic is:

$$t = r\left[\frac{n-2}{1-r^2}\right]^{1/2}$$

which has a $t$ distribution with $n - 2$ degrees of freedom.[4] For the correlation coefficient calculated based on the data given in Table 17.1,

$$t = 0.9361\left[\frac{12-2}{1-(0.9361)^2}\right]^{1/2}$$
$$= 8.414$$

and the degrees of freedom = 12 − 2 = 10. From the $t$ distribution table (Table 4 in the Statistical Appendix), the critical value of $t$ for a two-tailed test and $\alpha = 0.05$ is 2.228. Hence, the null hypothesis of no relationship between $X$ and $Y$ is rejected. This, along with the positive sign of $r$, indicates that attitude toward the city is positively related to the duration of residence in the city. Moreover, the high value of $r$ indicates that this relationship is strong. If this were a large and representative sample, the implication would be that

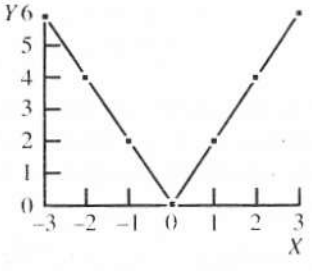



***Figure 17.1***
A Nonlinear Relationship for Which r = 0

managers, city officials, and politicians wishing to reach people with a favorable attitude toward the city should target long-time residents of that city.

In conducting multivariate data analysis, it is often useful to examine the simple correlation between each pair of variables. These results are presented in the form of a correlation matrix, which indicates the coefficient of correlation between each pair of variables. Usually, only the lower triangular portion of the matrix is considered. The diagonal elements all equal 1.00, because a variable correlates perfectly with itself. The upper triangular portion of the matrix is a mirror image of the lower triangular portion, because $r$ is a symmetric measure of association. The form of a correlation matrix for five variables, $V_1$ through $V_5$, is as follows.

| | $V_1$ | $V_2$ | $V_3$ | $V_4$ | $V_5$ |
|---|---|---|---|---|---|
| $V_1$ | | | | | |
| $V_2$ | 0.5 | | | | |
| $V_3$ | 0.3 | 0.4 | | | |
| $V_4$ | 0.1 | 0.3 | 0.6 | | |
| $V_5$ | 0.2 | 0.5 | 0.3 | 0.7 | |

Although a matrix of simple correlations provides insights into pairwise associations, sometimes researchers want to examine the association between two variables after controlling for one or more other variables. In the latter case, partial correlation should be estimated.

## PARTIAL CORRELATION

***partial correlation coefficient***
A measure of the association between two variables after controlling or adjusting for the effects of one or more additional variables.

Whereas the product moment or simple correlation is a measure of association describing the linear association between two variables, a ***partial correlation coefficient*** measures the association between two variables after controlling for or adjusting for the effects of one or more additional variables. This statistic is used to answer the following questions:

- How strongly are sales related to advertising expenditures when the effect of price is controlled?
- Is there an association between market share and size of the sales force after adjusting for the effect of sales promotion?
- Are consumers' perceptions of quality related to their perceptions of prices when the effect of brand image is controlled?

As in these situations, suppose one wanted to calculate the association between $X$ and $Y$ after controlling for a third variable, $Z$. Conceptually, one would first remove the effect of $Z$ from $X$. To do this, one would predict the values of $X$ based on a knowledge of $Z$ by using the product moment correlation between $X$ and $Z$, $r_{xz}$. The predicted value of $X$ is then subtracted from the actual value of $X$ to construct an adjusted value of $X$. In a similar manner, the values of $Y$ are adjusted to remove the effects of $Z$. The product moment correlation between the adjusted values of $X$ and the adjusted values of $Y$ is the partial correlation coefficient between $X$ and $Y$, after controlling for the effect of $Z$, and is denoted by $r_{xy.z}$. Statistically, because the simple correlation between two variables completely describes the linear relationship between them, the partial correlation coefficient can be calculated by a knowledge of the simple correlations alone, without using individual observations.

$$r_{xy.z} = \frac{r_{xy} - (r_{xz})(r_{yz})}{\sqrt{1 - r_{xz}^2}\sqrt{1 - r_{yz}^2}}$$

To continue our example, suppose the researcher wanted to calculate the association between attitude toward the city, $Y$, and duration of residence, $X_1$, after controlling for a third variable, importance attached to weather, $X_2$. These data are presented in Table 17.1. The simple correlations between the variables are:

$$r_{yx_1} = 0.9361 \qquad r_{yx_2} = 0.7334 \qquad r_{x_1x_2} = 0.5495$$

The required partial correlation is calculated as follows:

$$r_{yx_1 \cdot x_2} = \frac{0.9361 - (0.5495)(0.7334)}{\sqrt{1 - (0.5495)^2}\sqrt{1 - (0.7334)^2}}$$
$$= 0.9386$$

As can be seen, controlling for the effect of importance attached to weather has little effect on the association between attitude toward the city and duration of residence. Thus, regardless of the importance they attach to weather, those who have stayed in a city longer have more favorable attitudes towards the city and vice versa.

Partial correlations have an *order* associated with them. The order indicates how many variables are being adjusted or controlled. The simple correlation coefficient, $r$, has a zero-order, as it does not control for any additional variables when measuring the association between two variables. The coefficient $r_{xyz}$ is a first-order partial correlation coefficient, as it controls for the effect of one additional variable, $Z$. A second-order partial correlation coefficient controls for the effects of two variables, a third-order for the effects of three variables, and so on. The higher-order partial correlations are calculated similarly. The $(n + 1)$th-order partial coefficient may be calculated by replacing the simple correlation coefficients on the right side of the preceding equation with the $n$th-order partial coefficients.

Partial correlations can be helpful for detecting spurious relationships (see Chapter 15). The relationship between $X$ and $Y$ is spurious if it is solely due to the fact that $X$ is associated with $Z$, which is indeed the true predictor of $Y$. In this case, the correlation between $X$ and $Y$ disappears when the effect of $Z$ is controlled. Consider a case in which consumption of a cereal brand ($C$) is positively associated with income ($I$), with $r_{ci} = 0.28$. Because this brand was popularly priced, income was not expected to be a significant factor. Therefore, the researcher suspected that this relationship was spurious. The sample results also indicated that income is positively associated with household size ($H$), $r_{hi} = 0.48$, and that household size is associated with cereal consumption, $r_{ch} = 0.56$. These figures seem to indicate that the real predictor of cereal consumption is not income but household size. To test this assertion, the first-order partial correlation between cereal consumption and income is calculated, controlling for the effect of household size. The reader can verify that this partial correlation, $r_{ci.h}$, is 0.02, and the initial correlation between cereal consumption and income vanishes when the household size was controlled. Therefore, the correlation between income and cereal consumption is spurious. The special case when a partial correlation is larger than its respective zero-order correlation involves a suppressor effect (see Chapter 15).[5]

***part correlation coefficient***
A measure of the correlation between $Y$ and $X$ when the linear effects of the other independent variables have been removed from $X$ but not from $Y$.

Another correlation coefficient of interest is the ***part correlation coefficient.*** This coefficient represents the correlation between $Y$ and $X$ when the linear effects of the other independent variables have been removed from $X$ but not from $Y$. The part correlation coefficient, $r_{y(x.z)}$, is calculated as follows:

$$r_{y(x.z)} = \frac{r_{xy} - r_{yz}r_{xz}}{\sqrt{1 - r_{xz}^2}}$$

The part correlation between attitude toward the city and the duration of residence, when the linear effects of the importance attached to weather have been removed from the duration of residence, can be calculated as:

$$r_{y(x_1 \cdot x_2)} = \frac{0.9361 - (0.5495)(0.7334)}{\sqrt{1 - (0.5495)^2}}$$
$$= 0.63806$$

## REAL RESEARCH

### *Selling Ads to Home Shoppers*

Advertisements play a very important role in forming attitudes/preferences for brands. Often advertisers use celebrity spokespersons as a credible source to influence consumers' attitudes and purchase intentions. Another type of source credibility is corporate

credibility, which can also influence consumer reactions to advertisements and shape brand attitudes. In general, it has been found that for low-involvement products, attitude toward the advertisement mediates brand cognition (beliefs about the brand) and attitude toward the brand. What would happen to the effect of this mediating variable when products are purchased through a home shopping network? Home Shopping Budapest in Hungary conducted research to assess the impact of advertisements toward purchase. A survey was conducted where several measures were taken, such as attitude toward the product, attitude toward the brand, attitude toward the ad characteristics, brand cognitions, and so on. It was hypothesized that in a home shopping network, advertisements largely determined attitude toward the brand. In order to find the degree of association of attitude toward the ad with both attitude toward the brand and brand cognition, a partial correlation coefficient could be computed. The partial correlation would be calculated between attitude toward the brand and brand cognitions after controlling for the effects of attitude toward the ad on the two variables. If attitude toward the ad is significantly high, then the partial correlation coefficient should be significantly less than the product moment correlation between brand cognition and attitude toward the brand. Research was conducted that supported this hypothesis. Then, Saatchi & Saatchi *(www.saatchi.com)* designed the ads aired on Home Shopping Budapest to generate positive attitude toward the advertising, and this turned out to be a major competitive weapon for the network.[6] ■

The partial correlation coefficient is generally viewed as more important than the part correlation coefficient because it can be used to determine spurious and suppressor effects. The product moment correlation, partial correlation, and the part correlation coefficients all assume that the data are interval or ratio scaled. If the data do not meet these requirements, the researcher should consider the use of nonmetric correlation.

## NONMETRIC CORRELATION

At times, the researcher may have to compute the correlation coefficient between two variables that are nonmetric. It may be recalled that nonmetric variables do not have interval or ratio scale properties and do not assume a normal distribution. If the nonmetric variables are ordinal and numeric, Spearman's rho, $\rho_s$, and Kendall's tau, $\tau$, are two measures of ***nonmetric correlation*** that can be used to examine the correlation between them. Both these measures use rankings rather than the absolute values of the variables and the basic concepts underlying them are quite similar. Both vary from $-1.0$ to $+1.0$ (see Chapter 15).

***nonmetric correlation***
A correlation measure for two nonmetric variables that relies on rankings to compute the correlation.

In the absence of ties, Spearman's $\rho_s$ yields a closer approximation to the Pearson product moment correlation coefficient, $\rho$, than Kendall's $\tau$. In these cases, the absolute magnitude of $\tau$ tends to be smaller than Pearson's $\rho$. On the other hand, when the data contain a large number of tied ranks, Kendall's $\tau$ seems more appropriate. As a rule of thumb, Kendall's $\tau$ is to be preferred when a large number of cases fall into a relatively small number of categories (thereby leading to a large number of ties). Conversely, the use of Spearman's $\rho_s$ is preferable when we have a relatively larger number of categories (thereby having fewer ties).[7]

The product moment as well as the partial and part correlation coefficients provide a conceptual foundation for bivariate as well as multiple regression analysis.

## REGRESSION ANALYSIS

***Regression analysis*** is a powerful and flexible procedure for analyzing associative relationships between a metric dependent variable and one or more independent variables. It can be used in the following ways:

***regression analysis***
A statistical procedure for analyzing associative relationships between a metric dependent variable and one or more independent variables.

1. Determine whether the independent variables explain a significant variation in the dependent variable: whether a relationship exists

2. Determine how much of the variation in the dependent variable can be explained by the independent variables: strength of the relationship
3. Determine the structure or form of the relationship: the mathematical equation relating the independent and dependent variables
4. Predict the values of the dependent variable
5. Control for other independent variables when evaluating the contributions of a specific variable or set of variables

Although the independent variables may explain the variation in the dependent variable, this does not necessarily imply causation. The use of the terms *dependent* or *criterion* variables, and *independent* or *predictor* variables in regression analysis arises from the mathematical relationship between the variables. These terms do not imply that the criterion variable is dependent on the independent variables in a causal sense. Regression analysis is concerned with the nature and degree of association between variables and does not imply or assume any causality.

## BIVARIATE REGRESSION

***bivariate regression***
A procedure for deriving a mathematical relationship, in the form of an equation, between a single metric dependent variable and a single metric independent variable.

***Bivariate regression*** is a procedure for deriving a mathematical relationship, in the form of an equation, between a single metric dependent or criterion variable and a single metric independent or predictor variable. The analysis is similar in many ways to determining the simple correlation between two variables. However, because an equation has to be derived, one variable must be identified as the dependent and the other as the independent variable. The examples given earlier in the context of simple correlation can be translated into the regression context.

- Can variation in sales be explained in terms of variation in advertising expenditures? What is the structure and form of this relationship, and can it be modeled mathematically by an equation describing a straight line?
- Can the variation in market share be accounted for by the size of the sales force?
- Are consumers' perceptions of quality determined by their perceptions of price?

Before discussing the procedure for conducting bivariate regression, we define some important statistics.

## STATISTICS ASSOCIATED WITH BIVARIATE REGRESSION ANALYSIS

The following statistics and statistical terms are associated with bivariate regression analysis.

***Bivariate regression model.*** The basic regression equation is $Y_i = \beta_0 + \beta_1 X_i + e_i$, where $Y$ = dependent or criterion variable, $X$ = independent or predictor variable, $\beta_0$ = intercept of the line, $\beta_1$ = slope of the line, and $e_i$ is the error term associated with the $i$th observation.

***Coefficient of determination.*** The strength of association is measured by the coefficient of determination, $r^2$. It varies between 0 and 1 and signifies the proportion of the total variation in $Y$ that is accounted for by the variation in $X$.

***Estimated or predicted value.*** The estimated or predicted value of $Y_i$ is $\hat{Y}_i = a + bx$, where $\hat{Y}_i$ is the predicted value of $Y_i$, and $a$ and $b$ are estimators of $\beta_0$ and $\beta_1$, respectively.

***Regression coefficient.*** The estimated parameter $b$ is usually referred to as the nonstandardized regression coefficient.

***Scattergram.*** A scatter diagram, or scattergram, is a plot of the values of two variables for all the cases or observations.

***Standard error of estimate.*** This statistic, $SEE$, is the standard deviation of the actual $Y$ values from the predicted $\hat{Y}$ values.

***Standard error.*** The standard deviation of $b$, $SE_b$, is called the standard error.

***Standardized regression coefficient.*** Also termed the *beta coefficient* or *beta weight*, this is the slope obtained by the regression of $Y$ on $X$ when the data are standardized.

***Sum of squared errors.*** The distances of all the points from the regression line are squared and added together to arrive at the sum of squared errors, which is a measure of total error, $\Sigma e_j^2$.

***t statistic.*** A $t$ statistic with $n - 2$ degrees of freedom can be used to test the null hypothesis that no linear relationship exists between $X$ and $Y$, or $H_0$: $\beta_1 = 0$, where $t = \dfrac{b}{SE_b}$.

# CONDUCTING BIVARIATE REGRESSION ANALYSIS

The steps involved in conducting bivariate regression analysis are described in Figure 17.2. Suppose the researcher wants to explain attitudes toward the city of residence in terms of the duration of residence (see Table 17.1). In deriving such relationships, it is often useful to first examine a scatter diagram.

## Plot the Scatter Diagram

A scatter diagram, or scattergram, is a plot of the values of two variables for all the cases or observations. It is customary to plot the dependent variable on the vertical axis and the independent variable on the horizontal axis. A scatter diagram is useful for determining the form of the relationship between the variables. A plot can alert the researcher to

*Figure 17.2*
Conducting Bivariate Regression Analysis

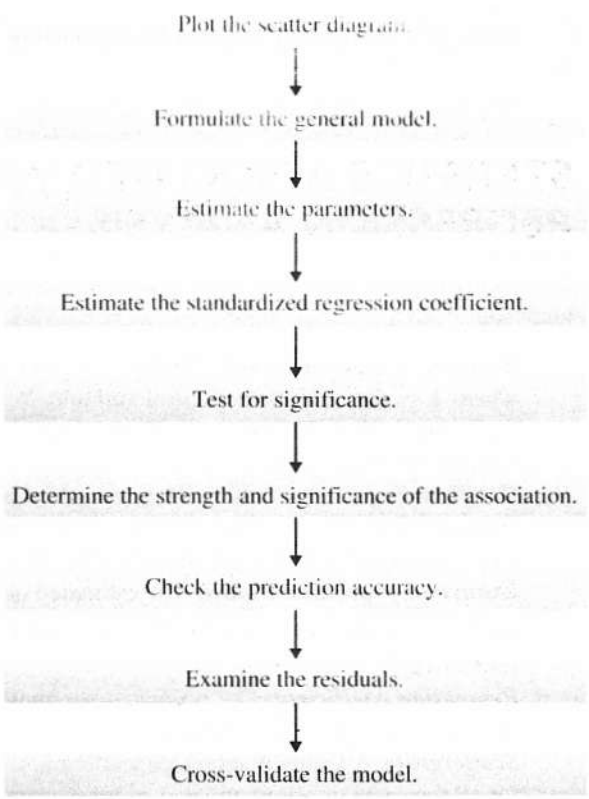
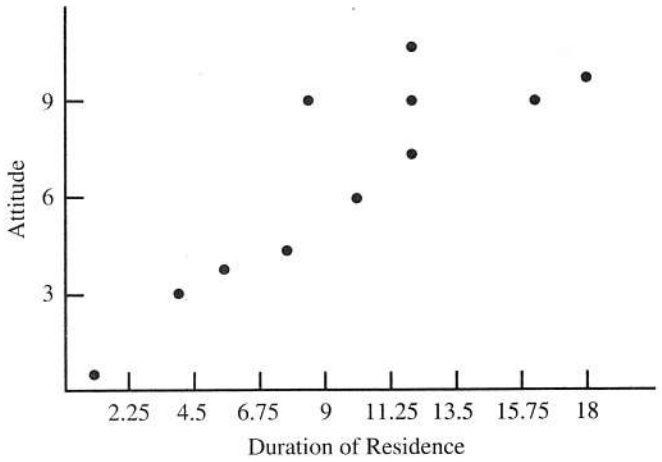


*Figure 17.3*
Plot of Attitude with Duration


**SPSS Output File**




patterns in the data, or to possible problems. Any unusual combinations of the two variables can be easily identified. A plot of $Y$ (attitude toward the city) against $X$ (duration of residence) is given in Figure 17.3. The points seem to be arranged in a band running from the bottom left to the top right. One can see the pattern: as one variable increases, so does the other. It appears from this scattergram that the relationship between $X$ and $Y$ is linear and could be well described by a straight line. However, as seen in Figure 17.4, several straight lines can be drawn through the data. How should the straight line be fitted to best describe the data?

*least-squares procedure*
A technique for fitting a straight line to a scattergram by minimizing the square of the vertical distances of all the points from the line.

The most commonly used technique for fitting a straight line to a scattergram is the ***least-squares procedure.*** This technique determines the best-fitting line by minimizing the square of the vertical distances of all the points from the line. The best-fitting line is called the *regression line.* Any point that does not fall on the regression line is not fu· , accounted for. The vertical distance from the point to the line is the error, $e_j$ (see Figure 1 ' 5). The distances of all the points from the line are squared and added together to arrive at the sum of

*Figure 17.4*
Which Straight Line Is Best?

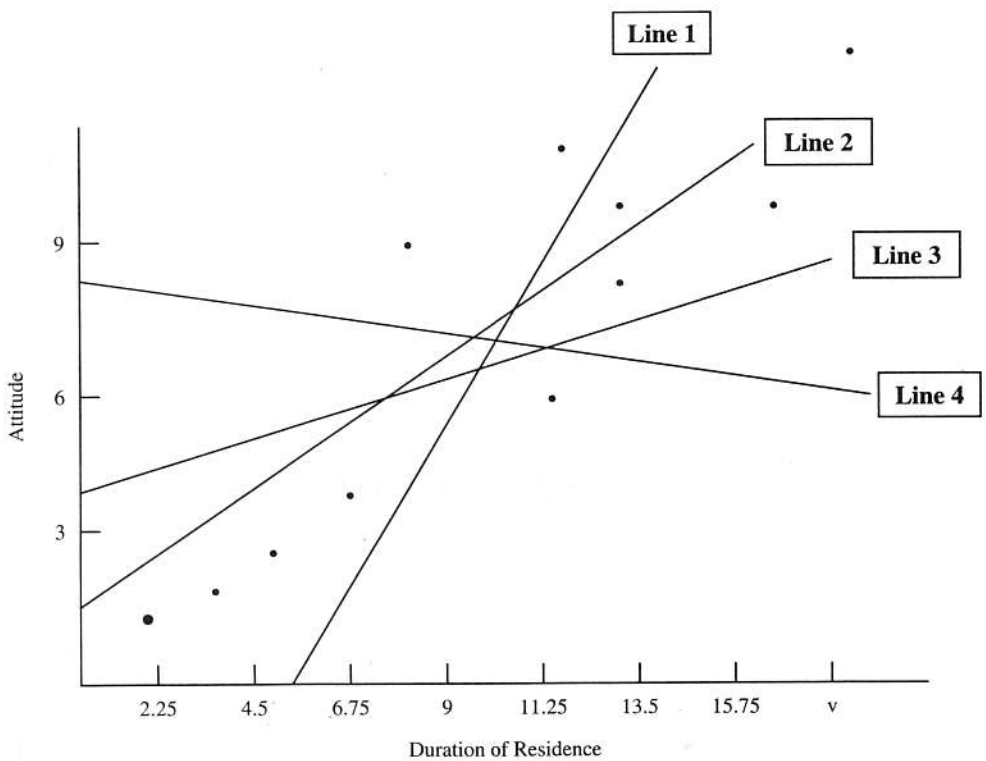
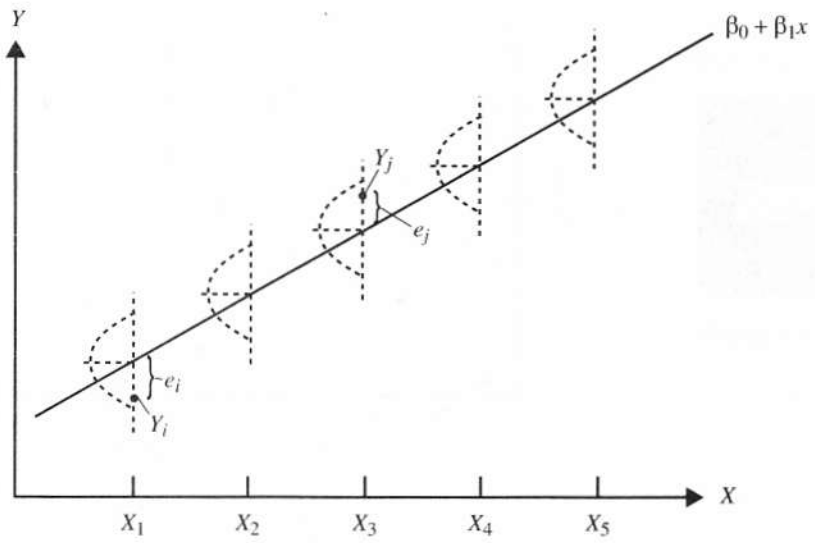


***Figure 17.5***
Bivariate Regression




squared errors, which is a measure of total error, $\Sigma e_j^2$. In fitting the line, the least-squares procedure minimizes the sum of squared errors. If $Y$ is plotted on the vertical axis and $X$ on the horizontal axis, as in Figure 17.5, the best-fitting line is called the regression of $Y$ on $X$, because the vertical distances are minimized. The scatter diagram indicates whether the relationship between $Y$ and $X$ can be modeled as a straight line and, consequently, whether the bivariate regression model is appropriate.

## Formulate the Bivariate Regression Model

In the bivariate regression model, the general form of a straight line is:

$$Y = \beta_0 + \beta_1 X$$

where

$Y$ = dependent or criterion variable
$X$ = independent or predictor variable
$\beta_0$ = intercept of the line
$\beta_1$ = slope of the line

This model implies a deterministic relationship, in that $Y$ is completely determined by $X$. The value of $Y$ can be perfectly predicted if $\beta_0$ and $\beta_1$ are known. In marketing research, however, very few relationships are deterministic. So the regression procedure adds an error term to account for the probabilistic or stochastic nature of the relationship. The basic regression equation becomes:

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

where $e_i$ is the error term associated with the $i$th observation.[8] Estimation of the regression parameters, $\beta_0$ and $\beta_1$, is relatively simple.

## Estimate the Parameters

In most cases, $\beta_0$ and $\beta_1$ are unknown and are estimated from the sample observations using the equation

$$\hat{Y}_i = a + bx_i$$

where $\hat{Y}_i$ is the estimated or predicted value of $Y_i$, and $a$ and $b$ are estimators of $\beta_0$ and $\beta_1$, respectively. The constant $b$ is usually referred to as the nonstandardized regression coefficient. It is the slope of the regression line and it indicates the expected change in $Y$ when $X$ is changed by one unit. The formulas for calculating $a$ and $b$ are simple.[9] The slope, $b$,

may be computed in terms of the covariance between $X$ and $Y$ ($COV_{xy}$) and the variance of $X$ as:

$$b = \frac{COV_{xy}}{S_x^2}$$
$$= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$
$$= \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{XY}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}$$

The intercept, $a$, may then be calculated using:

$$a = \overline{Y} - b\overline{X}$$

For the data in Table 17.1, the estimation of parameters may be illustrated as follows:

$$\sum_{i=1}^{12} X_i Y_i = (10)(6) + (12)(9) + (12)(8) + (4)(3) + (12)(10) + (6)(4) + (8)(5) + (2)(2) + (18)(11) + (9)(9) + (17)(10) + (2)(2)$$
$$= 917$$

$$\sum_{i=1}^{12} X_i^2 = 10^2 + 12^2 + 12^2 + 4^2 + 12^2 + 6^2 + 8^2 + 2^2 + 18^2 + 9^2 + 17^2 + 2^2$$
$$= 1,350$$

It may be recalled from earlier calculations of the simple correlation that

$$\overline{X} = 9.333$$
$$\overline{Y} = 6.583$$

Given $n = 12$, $b$ can be calculated as:

$$b = \frac{917 - (12)(9.333)(6.583)}{1350 - (12)(9.333)^2}$$
$$= 0.5897$$
$$a = \overline{Y} - b\overline{X}$$
$$= 6.583 - (0.5897)(9.333)$$
$$= 1.0793$$

Note that these coefficients have been estimated on the raw (untransformed) data. Should standardization of the data be considered desirable, the calculation of the standardized coefficients is also straightforward.

## Estimate Standardized Regression Coefficient

*Standardization* is the process by which the raw data are transformed into new variables that have a mean of 0 and a variance of 1 (Chapter 14). When the data are standardized, the intercept assumes a value of 0. The term *beta coefficient* or *beta weight* is used to denote the standardized regression coefficient. In this case, the slope obtained by the regression of $Y$ on $X$, $B_{yx}$, is the same as the slope obtained by the regression of $X$ on $Y$, $B_{xy}$. Moreover, each of these regression coefficients is equal to the simple correlation between $X$ and $Y$.

$$B_{yx} = B_{xy} = r_{xy}$$

SPSS Output File

**TABLE 17.2**

**Bivariate Regression**

| | |
|---|---|
| Multiple $R$ | 0.93608 |
| $R^2$ | 0.87624 |
| Adjusted $R^2$ | 0.86387 |
| Standard error | 1.22329 |

ANALYSIS OF VARIANCE

| | DF | SUM OF SQUARES | MEAN SQUARE |
|---|---|---|---|
| Regression | 1 | 105.95222 | 105.95222 |
| Residual | 10 | 14.96444 | 1.49644 |

$F = 70.80266$ Significance of $F = 0.0000$

VARIABLES IN THE EQUATION

| VARIABLE | B | $SE_B$ | BETA (B) | T | SIGNIFICANCE OF T |
|---|---|---|---|---|---|
| DURATION | 0.58972 | 0.07008 | 0.93608 | 8.414 | 0.0000 |
| (Constant) | 1.07932 | 0.74335 | | 1.452 | 0.1772 |

There is a simple relationship between the standardized and nonstandardized regression coefficients:

$$B_{yx} = b_{yx}(S_x/S_y)$$

Once the parameters have been estimated, they can be tested for significance.

## Test for Significance

The statistical significance of the linear relationship between $X$ and $Y$ may be tested by examining the hypotheses:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

The null hypothesis implies that there is no linear relationship between $X$ and $Y$. The alternative hypothesis is that there is a relationship, positive or negative, between $X$ and $Y$. Typically, a two-tailed test is done. A $t$ statistic with $n - 2$ degrees of freedom can be used, where

$$t = \frac{b}{SE_b}$$

$SE_b$ denotes the standard deviation of $b$ and is called the *standard error*.[10] The $t$ distribution was discussed in Chapter 15.

Using a computer program, the regression of attitude on duration of residence, using the data shown in Table 17.1, yielded the results shown in Table 17.2. The intercept, $a$, equals 1.0793, and the slope, $b$, equals 0.5897. Therefore, the estimated equation is:

$$\text{Attitude } (\hat{Y}) = 1.0793 + 0.5897 \text{ (Duration of residence)}$$

The standard error or standard deviation of $b$ is estimated as 0.07008, and the value of the $t$ statistic, $t = 0.5897/0.0700 = 8.414$, with $n - 2 = 10$ degrees of freedom. From Table 4 in the Statistical Appendix, we see that the critical value of $t$ with 10 degrees of freedom and $\alpha = 0.05$ is 2.228 for a two-tailed test. Because the calculated value of $t$ is larger than the critical value, the null hypothesis is rejected. Hence, there is a significant linear relationship between attitude toward the city and duration of residence in the city. The positive sign of the slope coefficient indicates that this relationship is positive. In other words, those who have resided in the city for a longer time have more positive attitudes toward the city. The implication for managers, city officials, and politicians is the same as that discussed for simple correlation, subject to the representativeness of the sample.

For the regression results given in Table 17.2, the value of the beta coefficient is estimated as 0.9361. Note that this is also the value of $r$ calculated earlier in this chapter.

## Determine the Strength and Significance of Association

A related inference involves determining the strength and significance of the association between $Y$ and $X$. The strength of association is measured by the coefficient of determination, $r^2$. In bivariate regression, $r^2$ is the square of the simple correlation coefficient obtained by correlating the two variables. The coefficient $r^2$ varies between 0 and 1. It signifies the proportion of the total variation in $Y$ that is accounted for by the variation in $X$. The decomposition of the total variation in $Y$ is similar to that for analysis of variance (Chapter 16). As shown in Figure 17.6, the total variation, $SS_y$, may be decomposed into the variation accounted for by the regression line, $SS_{reg}$, and the error or residual variation, $SS_{error}$ or $SS_{res}$, as follows:

$$SS_y = SS_{reg} + SS_{res}$$

where

$$SS_y = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$SS_{reg} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

$$SS_{res} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

The strength of association may then be calculated as follows:

$$r^2 = \frac{SS_{reg}}{SS_y}$$

$$= \frac{SS_y - SS_{res}}{SS_y}$$

To illustrate the calculations of $r^2$, let us consider again the regression of attitude toward the city on the duration of residence. It may be recalled from earlier calculations of the simple correlation coefficient that:

$$SS_y = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$= 120.9168$$

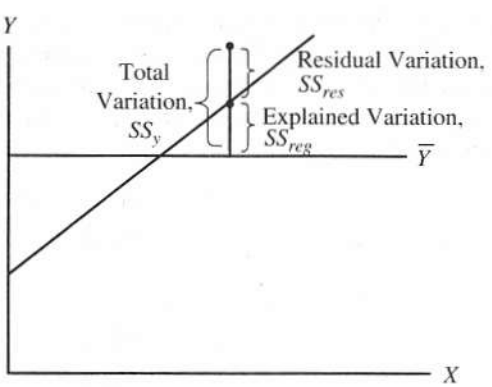
The predicted values ($\hat{Y}$) can be calculated using the regression equation:

$$\text{Attitude } (\hat{Y}) = 1.0793 + 0.5897 \text{ (Duration of residence)}$$

For the first observation in Table 17.1, this value is:

$$(\hat{Y}) = 1.0793 + 0.5897 \times 10 = 6.9763$$

***Figure 17.6***
Decomposition of the Total Variation in Bivariate Regression



For each successive observation, the predicted values are, in order, 8.1557, 8.1557, 3.4381, 8.1557, 4.6175, 5.7969, 2.2587, 11.6939, 6.3866, 11.1042, and 2.2587. Therefore,

$$\begin{aligned} SS_{reg} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 &= (6.9763 - 6.5833)^2 + (8.1557 - 6.5833)^2 \\ &\quad + (8.1557 - 6.5833)^2 + (3.4381 - 6.5833)^2 \\ &\quad + (8.1557 - 6.5833)^2 + (4.6175 - 6.5833)^2 \\ &\quad + (5.7969 - 6.5833)^2 + (2.2587 - 6.5833)^2 \\ &\quad + (11.6939 - 6.5833)^2 + (6.3866 - 6.5833)^2 \\ &\quad + (11.1042 - 6.5833)^2 + (2.2587 - 6.5833)^2 \\ &= 0.1544 + 2.4724 + 2.4724 + 9.8922 + 2.4724 \\ &\quad + 3.8643 + 0.6184 + 18.7021 + 26.1182 \\ &\quad + 0.0387 + 20.4385 + 18.7021 \\ &= 105.9524 \end{aligned}$$

$$\begin{aligned} SS_{res} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 &= (6 - 6.9763)^2 + (9 - 8.1557)^2 + (8 - 8.1557)^2 \\ &\quad + (3 - 3.4381)^2 + (10 - 8.1557)^2 + (4 - 4.6175)^2 \\ &\quad + (5 - 5.7969)^2 + (2 - 2.2587)^2 + (11 - 11.6939)^2 \\ &\quad + (9 - 6.3866)^2 + (10 - 11.1042)^2 + (2 - 2.2587)^2 \\ &= 14.9644 \end{aligned}$$

It can be seen that $SS_y = SS_{reg} + SS_{res}$. Furthermore,

$$\begin{aligned} r^2 &= \frac{SS_{reg}}{SS_y} \\ &= \frac{105.9524}{120.9168} \\ &= 0.8762 \end{aligned}$$

Another equivalent test for examining the significance of the linear relationship between $X$ and $Y$ (significance of $b$) is the test for the significance of the coefficient of determination. The hypotheses in this case are:

$$\begin{aligned} &H_0: R^2_{pop} = 0 \\ &H_1: R^2_{pop} > 0 \end{aligned}$$

The appropriate test statistic is the $F$ statistic:

$$F = \frac{SS_{reg}}{SS_{res}/(n-2)}$$

which has an $F$ distribution with 1 and $n - 2$ degrees of freedom. The $F$ test is a generalized form of the $t$ test (see Chapter 15). If a random variable $t$ is distributed with $n$ degrees of freedom, then $t^2$ is $F$ distributed with 1 and $n$ degrees of freedom. Hence, the $F$ test for testing the significance of the coefficient of determination is equivalent to testing the following hypotheses:

$$\begin{aligned} &H_0: \beta_1 = 0 \\ &H_0: \beta_1 \neq 0 \end{aligned}$$

or

$$\begin{aligned} &H_0: \rho = 0 \\ &H_0: \rho \neq 0 \end{aligned}$$

From Table 17.2, it can be seen that:

$$r^2 = \frac{105.9524}{(105.9524 + 14.9644)}$$
$$= 0.8762$$

which is the same as the value calculated earlier. The value of the $F$ statistic is:

$$F = \frac{105.9524}{(14.9644/10)}$$
$$= 70.8027$$

with 1 and 10 degrees of freedom. The calculated $F$ statistic exceeds the critical value of 4.96 determined from Table 5 in the Statistical Appendix. Therefore, the relationship is significant at $\alpha = 0.05$, corroborating the results of the $t$ test. If the relationship between $X$ and $Y$ is significant, it is meaningful to predict the values of $Y$ based on the values of $X$ and to estimate prediction accuracy.

## Check Prediction Accuracy

To estimate the accuracy of predicted values, $\hat{Y}$, it is useful to calculate the standard error of estimate, *SEE*. This statistic is the standard deviation of the actual $Y$ values from the predicted $\hat{Y}$ values.

$$SEE = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y})^2}{n - 2}}$$

or

$$SEE = \sqrt{\frac{SS_{res}}{n - 2}}$$

or more generally, if there are $k$ independent variables,

$$SEE = \sqrt{\frac{SS_{res}}{n - k - 1}}$$

*SEE* may be interpreted as a kind of average residual or average error in predicting $Y$ from the regression equation.[11]

Two cases of prediction may arise. The researcher may want to predict the mean value of $Y$ for all the cases with a given value of $X$, say $X_0$, or predict the value of $Y$ for a single case. In both situations, the predicted value is the same and is given by $\hat{Y}$, where

$$\hat{Y} = a + bX_0$$

However, the standard error is different in the two situations, although in both situations it is a function of *SEE*. For large samples, the standard error for predicting mean value of $Y$ is $SEE/\sqrt{n}$, and for predicting individual $Y$ values it is *SEE*. Hence, the construction of confidence intervals (see Chapter 12) for the predicted value varies, depending upon whether the mean value or the value for a single observation is being predicted.

For the data given in Table 17.2, the *SEE* is estimated as follows:

$$SEE = \sqrt{\frac{14.9644}{(12 - 2)}}$$
$$= 1.22329$$

The final two steps in conducting bivariate regression, namely examination of residuals and model cross-validation, are considered later.

**ACTIVE RESEARCH**

Visit *www.ford.com* and conduct an Internet search using a search engine and your library's online database to obtain information on the relationship between advertising and sales for automobile manufacturers.

Formulate a bivariate regression model explaining the relationship between advertising and sales in the automobile industry.

As the marketing director for Ford Motor Company, how would you determine your advertising expenditures?

## Assumptions

The regression model makes a number of assumptions in estimating the parameters and in significance testing, as shown in Figure 17.5:

1. The error term is normally distributed. For each fixed value of $X$, the distribution of $Y$ is normal.[12]
2. The means of all these normal distributions of $Y$, given $X$, lie on a straight line with slope $b$.
3. The mean of the error term is 0.
4. The variance of the error term is constant. This variance does not depend on the values assumed by $X$.
5. The error terms are uncorrelated. In other words, the observations have been drawn independently.

Insights into the extent to which these assumptions have been met can be gained by an examination of residuals, which is covered in the next section on multiple regression.[13]

# MULTIPLE REGRESSION

***multiple regression***
A statistical technique that simultaneously develops a mathematical relationship between two or more independent variables and an interval-scaled dependent variable.

***Multiple regression*** involves a single dependent variable and two or more independent variables. The questions raised in the context of bivariate regression can also be answered via multiple regression by considering additional independent variables.

- Can variation in sales be explained in terms of variation in advertising expenditures, prices, and level of distribution?
- Can variation in market shares be accounted for by the size of the sales force, advertising expenditures, and sales promotion budgets?
- Are consumers' perceptions of quality determined by their perceptions of prices, brand image, and brand attributes?

Additional questions can also be answered by multiple regression.

- How much of the variation in sales can be explained by advertising expenditures, prices, and level of distribution?
- What is the contribution of advertising expenditures in explaining the variation in sales when the levels of prices and distribution are controlled?
- What levels of sales may be expected, given the levels of advertising expenditures, prices, and level of distribution?

**REAL RESEARCH**

### *Global Brands—Local Ads*

Europeans welcome brands from other countries, but when it comes to advertising, they prefer the homegrown variety. A survey done by Yankelovich and Partners (*www.yankelovich.com*) and its affiliates finds that most European consumers' favorite commercials are for local brands even though they are more than likely to buy foreign brands. Respondents in France, Germany, and the United Kingdom named Coca-Cola as the most often purchased soft drink. However, the French selected the famous award-winning spot for

France's Perrier bottled water as their favorite commercial. Similarly, in Germany, the favorite advertising was for a German brand of nonalcoholic beer—Clausthaler. However, in the United Kingdom, Coca-Cola was the favorite soft drink and also the favorite advertising. In light of such findings, the important question is—does advertising help? Does it help increase the purchase probability of the brand or does it merely maintain a high brand recognition rate? One way of finding out is by running multiple regressions where the dependent variable is the likelihood of brand purchase and the independent variables are brand attribute evaluations and advertising evaluations. Separate models with and without advertising can be run to assess any significant difference in the contribution. Individual $t$ tests could also be examined to find out the significant contribution of both the brand attributes and advertising. The results will indicate the degree to which advertising plays an important part in brand purchase decisions. In conjunction with these results, a recent study revealed that attempting to build brand loyalty purchases by means of a sales promotion is not a desirable way to achieve such an objective. According to the study, sales promotions only encourage momentary brand switching and merely enhance short-term performance for companies. Furthermore, over the long run, a sales promotion may imply a low quality or unstable brand image to consumers or it may confuse consumers, which could also lead to a decline in brand loyalty. The results of this study show that sacrificing advertising and relying on sales promotions reduce brand associations, which ultimately leads to a decrease in brand loyalty purchases.[14] ■

***multiple regression model***
An equation used to explain the results of multiple regression analysis.

The general form of the ***multiple regression model*** is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + e$$

which is estimated by the following equation:

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_k X_k$$

As before, the coefficient $a$ represents the intercept, but the $b$s are now the partial regression coefficients. The least-squares criterion estimates the parameters in such a way as to minimize the total error, $SS_{res}$. This process also maximizes the correlation between the actual values of $Y$ and the predicted values, $\hat{Y}$. All the assumptions made in bivariate regression also apply in multiple regression. We define some associated statistics and then describe the procedure for multiple regression analysis.[15]

## STATISTICS ASSOCIATED WITH MULTIPLE REGRESSION

Most of the statistics and statistical terms described under bivariate regression also apply to multiple regression. In addition, the following statistics are used:

***Adjusted* $R^2$.** $R^2$, coefficient of multiple determination, is adjusted for the number of independent variables and the sample size to account for diminishing returns. After the first few variables, the additional independent variables do not make much contribution.

***Coefficient of multiple determination.*** The strength of association in multiple regression is measured by the square of the multiple correlation coefficient, $R^2$, which is also called the coefficient of multiple determination.

**F *test.*** The $F$ test is used to test the null hypothesis that the coefficient of multiple determination in the population, $R^2_{pop}$, is zero. This is equivalent to testing the null hypothesis $H_0$: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$. The test statistic has an $F$ distribution with $k$ and $(n - k - 1)$ degrees of freedom.

***Partial* F *test.*** The significance of a partial regression coefficient, $\beta_i$, of $X_i$ may be tested using an incremental $F$ statistic. The incremental $F$ statistic is based on the increment in the explained sum of squares resulting from the addition of the independent variable $X_i$ to the regression equation after all the other independent variables have been included.

***Partial regression coefficient.*** The partial regression coefficient, $b_1$, denotes the change in the predicted value, $\hat{Y}$, per unit change in $X_1$ when the other independent variables, $X_2$ to $X_k$, are held constant.

# CONDUCTING MULTIPLE REGRESSION ANALYSIS

The steps involved in conducting multiple regression analysis are similar to those for bivariate regression analysis. The discussion focuses on partial regression coefficients, strength of association, significance testing, and examination of residuals.

## Partial Regression Coefficients

To understand the meaning of a partial regression coefficient, let us consider a case in which there are two independent variables, so that:

$$\hat{Y} = a + b_1 X_1 + b_2 X_2$$

First, note that the relative magnitude of the partial regression coefficient of an independent variable is, in general, different from that of its bivariate regression coefficient. In other words, the partial regression coefficient, $b_1$, will be different from the regression coefficient, $b$, obtained by regressing $Y$ on only $X_1$. This happens because $X_1$ and $X_2$ are usually correlated. In bivariate regression, $X_2$ was not considered, and any variation in $Y$ that was shared by $X_1$ and $X_2$ was attributed to $X_1$. However, in the case of multiple independent variables, this is no longer true.

The interpretation of the partial regression coefficient, $b_1$, is that it represents the expected change in $Y$ when $X_1$ is changed by one unit but $X_2$ is held constant or otherwise controlled. Likewise, $b_2$ represents the expected change in $Y$ for a unit change in $X_2$, when $X_1$ is held constant. Thus, calling $b_1$ and $b_2$ partial regression coefficients is appropriate. It can also be seen that the combined effects of $X_1$ and $X_2$ on $Y$ are additive. In other words, if $X_1$ and $X_2$ are each changed by one unit, the expected change in $Y$ would be $(b_1 + b_2)$.

Conceptually, the relationship between the bivariate regression coefficient and the partial regression coefficient can be illustrated as follows. Suppose one were to remove the effect of $X_2$ from $X_1$. This could be done by running a regression of $X_1$ on $X_2$. In other words, one would estimate the equation $\hat{X}_1 = a + bX_2$ and calculate the residual $X_r = (X_1 - \hat{X}_1)$. The partial regression coefficient, $b_1$, is then equal to the bivariate regression coefficient, $b_r$, obtained from the equation $\hat{Y} = a + b_r X_r$. In other words, the partial regression coefficient, $b_1$, is equal to the regression coefficient, $b_r$, between $Y$ and the residuals of $X_1$ from which the effect of $X_2$ has been removed. The partial coefficient, $b_2$, can also be interpreted along similar lines.

Extension to the case of $k$ variables is straightforward. The partial regression coefficient, $b_1$, represents the expected change in $Y$ when $X_1$ is changed by one unit and $X_2$ through $X_k$ are held constant. It can also be interpreted as the bivariate regression coefficient, $b$, for the regression of $Y$ on the residuals of $X_1$, when the effect of $X_2$ through $X_k$ has been removed from $X_1$.

The beta coefficients are the partial regression coefficients obtained when all the variables ($Y$, $X_1$, $X_2$, . . . $X_k$) have been standardized to a mean of 0 and a variance of 1 before estimating the regression equation. The relationship of the standardized to the nonstandardized coefficients remains the same as before:

$$B_1 = b_1 \left( \frac{S_{x1}}{S_y} \right)$$

.

.

$$B_k = b_k \left( \frac{S_{xk}}{S_y} \right)$$

The intercept and the partial regression coefficients are estimated by solving a system of simultaneous equations derived by differentiating and equating the partial derivatives to 0. Because these coefficients are automatically estimated by the various computer programs, we will not present the details. Yet it is worth noting that the equations cannot be solved if: (1) the sample size, $n$, is smaller than or equal to the number of independent variables, $k$; or (2) one independent variable is perfectly correlated with another.


SPSS Output File

**TABLE 17.3**

**Multiple Regression**

| | |
|---|---|
| Multiple $R$ | 0.97210 |
| $R^2$ | 0.94498 |
| Adjusted $R^2$ | 0.93276 |
| Standard error | 0.85974 |

ANALYSIS OF VARIANCE

| | DF | SUM OF SQUARES | MEAN SQUARE |
|---|---|---|---|
| Regression | 2 | 114.26425 | 57.13213 |
| Residual | 9 | 6.65241 | 0.73916 |

$F = 77.29364$ Significance of $F = 0.0000$

VARIABLES IN THE EQUATION

| VARIABLE | B | $SE_B$ | BETA (B) | T | SIGNIFICANCE OF T |
|---|---|---|---|---|---|
| IMPORTANCE | 0.28865 | 0.08608 | 0.31382 | 3.353 | 0.0085 |
| DURATION | 0.48108 | 0.05895 | 0.76363 | 8.160 | 0.0000 |
| (Constant) | 0.33732 | 0.56736 | | 0.595 | 0.5668 |

Suppose that in explaining the attitude toward the city, we now introduce a second variable, importance attached to the weather. The data for the 12 pretest respondents on attitude toward the city, duration of residence, and importance attached to the weather are given in Table 17.1. The results of multiple regression analysis are depicted in Table 17.3. The partial regression coefficient for duration ($X_1$) is now 0.4811, different from what it was in the bivariate case. The corresponding beta coefficient is 0.7636. The partial regression coefficient for importance attached to weather ($X_2$) is 0.2887, with a beta coefficient of 0.3138. The estimated regression equation is:

$$(\hat{Y}) = 0.33732 + 0.48108X_1 + 0.28865X_2$$

or

$$\text{Attitude} = 0.33732 + 0.48108\ (\text{Duration}) + 0.28865\ (\text{Importance})$$

This equation can be used for a variety of purposes, including predicting attitudes toward the city, given a knowledge of the respondents' duration of residence in the city and the importance they attach to weather. Note that both Duration and Importance are significant and useful in this prediction.

## Strength of Association

The strength of the relationship stipulated by the regression equation can be determined by using appropriate measures of association. The total variation is decomposed as in the bivariate case:

$$SS_y = SS_{reg} + SS_{res}$$

where

$$SS_y = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

$$SS_{reg} = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$$

$$SS_{res} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

The strength of association is measured by the square of the multiple correlation coefficient, $R^2$, which is also called the coefficient of multiple determination.

$$R^2 = \frac{SS_{reg}}{SS_y}$$

The multiple correlation coefficient, $R$, can also be viewed as the simple correlation coefficient, $r$, between $Y$ and $\hat{Y}$. Several points about the characteristics of $R^2$ are worth noting. The coefficient of multiple determination, $R^2$, cannot be less than the highest bivariate, $r^2$, of any individual independent variable with the dependent variable. $R^2$ will be larger when the correlations between the independent variables are low. If the independent variables are statistically independent (uncorrelated), then $R^2$ will be the sum of bivariate $r^2$ of each independent variable with the dependent variable. $R^2$ cannot decrease as more independent variables are added to the regression equation. Yet diminishing returns set in, so that after the first few variables, the additional independent variables do not make much of a contribution.[16] For this reason, $R^2$ is adjusted for the number of independent variables and the sample size by using the following formula:

$$\text{Adjusted } R^2 = R^2 - \frac{k(1 - R^2)}{n - k - 1}$$

For the regression results given in Table 17.3, the value of $R^2$ is:

$$R^2 = \frac{114.2643}{(114.2643 + 6.6524)}$$
$$= 0.9450$$

This is higher than the $r^2$ value of 0.8762 obtained in the bivariate case. The $r^2$ in the bivariate case is the square of the simple (product moment) correlation between attitude toward the city and duration of residence. The $R^2$ obtained in multiple regression is also higher than the square of the simple correlation between attitude and importance attached to weather (which can be estimated as 0.5379). The adjusted $R^2$ is estimated as:

$$\text{Adjusted } R^2 = 0.9450 - \frac{2(1.0 - 0.9450)}{(12 - 2 - 1)}$$
$$= 0.9328$$

Note that the value of adjusted $R^2$ is close to $R^2$ and both are higher than $r^2$ for the bivariate case. This suggests that the addition of the second independent variable, importance attached to weather, makes a contribution in explaining the variation in attitude toward the city.

## Significance Testing

Significance testing involves testing the significance of the overall regression equation as well as specific partial regression coefficients. The null hypothesis for the overall test is that the coefficient of multiple determination in the population, $R^2_{pop}$, is zero.

$$H_0: R^2_{pop} = 0$$

This is equivalent to the following null hypothesis:

$$H_0: \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$$

The overall test can be conducted by using an $F$ statistic:

$$F = \frac{SS_{reg}/k}{SS_{res}/(n - k - 1)}$$
$$= \frac{R^2/k}{(1 - R^2)/(n - k - 1)}$$

which has an $F$ distribution with $k$ and $(n - k - 1)$ degrees of freedom.[17] For the multiple regression results given in Table 17.3,

$$F = \frac{114.2643/2}{6.6524/9} = 77.2936$$

which is significant at $\alpha = 0.05$.

If the overall null hypothesis is rejected, one or more population partial regression coefficients have a value different from 0. To determine which specific coefficients ($\beta_i$'s) are nonzero, additional tests are necessary. Testing for the significance of the $\beta_i$'s can be done in a manner similar to that in the bivariate case by using $t$ tests. The significance of the partial coefficient for importance attached to weather may be tested by the following equation:

$$t = \frac{b}{SE_b}$$
$$= \frac{0.2887}{0.08608}$$
$$= 3.353$$

which has a $t$ distribution with $n - k - 1$ degrees of freedom. This coefficient is significant at $\alpha = 0.05$. The significance of the coefficient for duration of residence is tested in a similar way and found to be significant. Therefore, both the duration of residence and importance attached to weather are important in explaining attitude toward the city.

Some computer programs provide an equivalent $F$ test, often called the *partial F test.* This involves a decomposition of the total regression sum of squares, $SS_{reg}$, into components related to each independent variable. In the standard approach, this is done by assuming that each independent variable has been added to the regression equation after all the other independent variables have been included. The increment in the explained sum of squares, resulting from the addition of an independent variable, $X_i$, is the component of the variation attributed to that variable and is denoted by $SS_{x_i}$.[18] The significance of the partial regression coefficient for this variable, $b_i$, is tested using an incremental $F$ statistic:

$$F = \frac{SS_{x_i}/1}{SS_{res}/(n - k - 1)}$$

which has an $F$ distribution with 1 and $(n - k - 1)$ degrees of freedom.

Although high $R^2$ and significant partial regression coefficients are comforting, the efficacy of the regression model should be evaluated further by an examination of the residuals.

## Examination of Residuals

***residual***
The difference between the observed value of $Y_i$ and the value predicted by the regression equation, $\hat{Y}_i$.

A ***residual*** is the difference between the observed value of $Y_i$ and the value predicted by the regression equation, $\hat{Y}_i$. Residuals are used in the calculation of several statistics associated with regression. In addition, scattergrams of the residuals, in which the residuals are plotted against the predicted values, $\hat{Y}$, time, or predictor variables, provide useful insights in examining the appropriateness of the underlying assumptions and regression model fitted.[19]

The assumption of a normally distributed error term can be examined by constructing a histogram of the residuals. A visual check reveals whether the distribution is normal. Additional evidence can be obtained by determining the percentages of residuals falling within ±1 *SE* or ±2 *SE*. These percentages can be compared with what would be expected under the normal distribution (68 percent and 95 percent, respectively). More formal assessment can be made by running the K-S one-sample test.

The assumption of constant variance of the error term can be examined by plotting the residuals against the predicted values of the dependent variable, $\hat{Y}_i$. If the pattern is not random, the variance of the error term is not constant. Figure 17.7 shows a pattern whose variance is dependent upon the $\hat{Y}_i$ values.

*Figure 17.7*
Residual Plot Indicating that Variance Is Not Constant

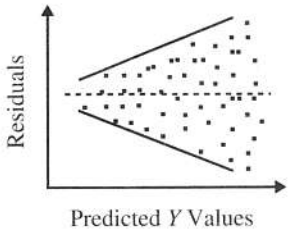
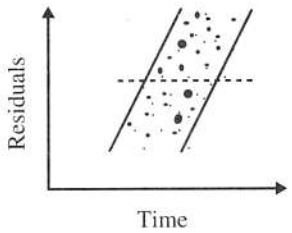



*Figure 17.8*
Plot Indicating a Linear Relationship Between Residuals and Time




A plot of residuals against time, or the sequence of observations, will throw some light on the assumption that the error terms are uncorrelated. A random pattern should be seen if this assumption is true. A plot like the one in Figure 17.8 indicates a linear relationship between residuals and time. A more formal procedure for examining the correlations between the error terms is the Durbin-Watson test.[20]

Plotting the residuals against the independent variables provides evidence of the appropriateness or inappropriateness of using a linear model. Again, the plot should result in a random pattern. The residuals should fall randomly, with relatively equal distribution dispersion about 0. They should not display any tendency to be either positive or negative.

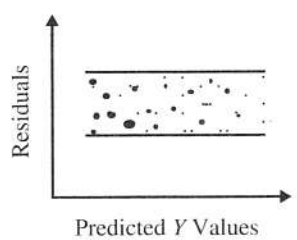
To examine whether any additional variables should be included in the regression equation, one could run a regression of the residuals on the proposed variables. If any variable explains a significant proportion of the residual variation, it should be considered for inclusion. Inclusion of variables in the regression equation should be strongly guided by the researcher's theory. Thus, an examination of the residuals provides valuable insights into the appropriateness of the underlying assumptions and the model that is fitted. Figure 17.9 shows a plot that indicates that the underlying assumptions are met and that the linear model is appropriate. If an examination of the residuals indicates that the assumptions underlying linear regression are not met, the researcher can transform the variables in an attempt to satisfy the assumptions. Transformations, such as taking logs, square roots, or reciprocals, can stabilize the variance, make the distribution normal, or make the relationship linear.

**REAL RESEARCH**

## *What Influences Sports Ticket Prices? A New Stadium!*

A major source of revenue for any professional sports team is through ticket sales, especially sales to season ticket subscribers. A study performed a regression analysis to determine what

*Figure 17.9*
Plot of Residuals Indicating that a Fitted Model Is Appropriate



factors caused ticket prices to vary among teams in the same league within a given year. The regression equation was:

$$\text{LTIX} = a_0 + a_1(\text{HWIN}) + a_2(\text{INCOME}) + a_3(\text{PAY}) + a_4(\text{POPL}) + a_5(\text{TREND}) + a_6(\text{CAP}) + a_7(\text{STAD})$$

where

LTIX = natural log of average ticket price
TIX = average ticket price
HWIN = average number of wins by the team in the previous three seasons
INCOME = average income level of city population
PAY = team payroll
POPL = population size of city
TREND = trends in the industry
CAP = attendance as a percentage of capacity
STAD = if the team is playing in a new stadium

The research gathered data covering a span of seven years (1996–2002). The financial data were gathered through Team Marketing Reports and the rest of the data were collected using publicly available sources such as sports reports. The results of the regression analyses can be seen in the accompanying table.

The results suggest that several factors influenced ticket prices, and the largest factor was that the team was playing in a new stadium.[21] ■

Regression Results

| | *MLB* | | | *NBA* | | | *NFL* | | | *NHL* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Variable* | *Coefficient* | *t-Statistic* | *P-Value* | *Coefficient* | *t-Statistic* | *P-Value* | *Coefficient* | *t-Statistic* | *P-Value* | *Coefficient* | *t-Statistic* | *P-Value* |
| Constant | 1.521 | 12.012 | 0.000 | 2.965 | 20.749 | 0.000 | 2.886 | 18.890 | 0.000 | 3.172 | 16.410 | 0.000 |
| POPL | 0.000 | 5.404 | 0.000 | 0.000 | 5.036 | 0.000 | 0.000 | -2.287 | 0.023 | 0.000 | 2.246 | 0.026 |
| INCOME | 0.000 | 3.991 | 0.000 | 0.000 | 0.208 | 0.836 | 0.000 | 3.645 | 0.000 | 0.000 | 0.669 | 0.504 |
| STAD | 0.337 | 5.356 | 0.000 | 0.108 | 3.180 | 0.002 | 0.226 | 3.357 | 0.001 | 0.321 | 4.087 | 0.000 |
| HWIN | 0.000 | 0.091 | 0.927 | 0.004 | 3.459 | 0.001 | 0.013 | 2.190 | 0.030 | 0.001 | 0.369 | 0.713 |
| CAP | 0.006 | 8.210 | 0.000 | 0.000 | 2.968 | 0.003 | 0.002 | 1.325 | 0.187 | 0.005 | 3.951 | 0.000 |
| PAY | 0.004 | 4.192 | 0.000 | 0.008 | 5.341 | 0.000 | 0.001 | 0.607 | 0.545 | 0.002 | 1.099 | 0.273 |
| TREND | 0.047 | 6.803 | 0.000 | 0.016 | 1.616 | 0.100 | 0.058 | 6.735 | 0.000 | 0.009 | 0.718 | 0.474 |
| CAN (Canada) | | | | | | | | | | −0.146 | −3.167 | 0.002 |
| Adjusted *R*-squared | | | 0.778 | | | 0.488 | | | 0.443 | | | 0.292 |
| *F* Statistic | | | 98.366 | | | 28.227 | | | 24.763 | | | 9.545 |
| *F* Significance | | | 0.000 | | | 0.000 | | | 0.000 | | | |

As in the preceding example, some independent variables considered in a study often turn out not to be significant. When there are a large number of independent variables and the researcher suspects that not all of them are significant, stepwise regression should be used.

## STEPWISE REGRESSION

***stepwise regression***
A regression procedure in which the predictor variables enter or leave the regression equation one at a time.

The purpose of ***stepwise regression*** is to select, from a large number of predictor variables, a small subset of variables that account for most of the variation in the dependent or criterion variable. In this procedure, the predictor variables enter or are removed from the regression equation one at a time.[22] There are several approaches to stepwise regression.

1. ***Forward inclusion.*** Initially, there are no predictor variables in the regression equation. Predictor variables are entered one at a time, only if they meet certain criteria

specified in terms of the $F$ ratio. The order in which the variables are included is based on the contribution to the explained variance.

2. ***Backward elimination***. Initially, all the predictor variables are included in the regression equation. Predictors are then removed one at a time based on the $F$ ratio.
3. ***Stepwise solution***. Forward inclusion is combined with the removal of predictors that no longer meet the specified criterion at each step.

Stepwise procedures do not result in regression equations that are optimal, in the sense of producing the largest $R^2$, for a given number of predictors. Because of the correlations between predictors, an important variable may never be included, or less important variables may enter the equation. To identify an optimal regression equation, one would have to compute combinatorial solutions in which all possible combinations are examined. Nevertheless, stepwise regression can be useful when the sample size is large in relation to the number of predictors, as shown in the following example.

## REAL RESEARCH

### *Stepping Out . . . to the Mall*

Even in the 21st century, browsing is a fundamental part of shopping—whether it is online or in the mall. Customers like to consider their purchase decisions before actually carrying them out. Many consider store-based retailers to have an advantage over Web-based retailers when it comes to browsing because store-based retailers are larger in size and product offerings. Although the Web appeals to younger shoppers, the mall will remain ahead of the game, especially with so many entertainment factors now being built inside malls. A profile of browsers in regional shopping malls was constructed using three sets of independent variables: demographics, shopping behavior, and psychological and attitudinal variables. The dependent variable consisted of a browsing index. In a stepwise regression including all three sets of variables, demographics were found to be the most powerful predictors of browsing behavior. The final regression equation, which contained 20 of the possible 36 variables, included all of the demographics. The accompanying table presents the regression coefficients, standard errors of the coefficients, and their significance levels.

Regression of Browsing Index on Descriptive and Attitudinal Variables by Order of Entry into Stepwise Regression

| *Variable Description* | *Coefficient* | *SE* | *Significance* |
|---|---|---|---|
| Sex (0 = Male, 1 = Female) | −0.485 | 0.164 | 0.001 |
| Employment status (0 = Employed) | 0.391 | 0.182 | 0.003 |
| Self-confidence | −0.152 | 0.128 | 0.234 |
| Education | 0.079 | 0.072 | 0.271 |
| Brand intention | −0.063 | 0.028 | 0.024 |
| Watch daytime TV? (0 = Yes) | 0.232 | 0.144 | 0.107 |
| Tension | −0.182 | 0.069 | 0.008 |
| Income | 0.089 | 0.061 | 0.144 |
| Frequency of mall visits | −0.130 | 0.059 | 0.028 |
| Fewer friends than most | 0.162 | 0.084 | 0.054 |
| Good shopper | −0.122 | 0.090 | 0.174 |
| Others' opinions important | −0.147 | 0.065 | 0.024 |
| Control over life | −0.069 | 0.069 | 0.317 |
| Family size | −0.086 | 0.062 | 0.165 |
| Enthusiastic person | −0.143 | 0.099 | 0.150 |
| Age | 0.036 | 0.069 | 0.603 |
| Number of purchases made | −0.068 | 0.043 | 0.150 |
| Purchases per store | 0.209 | 0.152 | 0.167 |
| Shop on tight budget | −0.055 | 0.067 | 0.412 |
| Excellent judge of quality | −0.070 | 0.089 | 0.435 |
| CONSTANT | 3.250 | | |
| Overall $R^2 = 0.477$ | | | |

In interpreting the coefficients, it should be recalled that the smaller the browsing index (the dependent variable), the greater the tendency to exhibit behaviors associated with browsing. The two predictors with the largest coefficients are sex and employment status. Browsers are more likely to be employed females. They also tend to be somewhat downscale, compared to other mall patrons, exhibiting lower levels of education and income, after accounting for the effects of sex and employment status. Although browsers tend to be somewhat younger than nonbrowsers, they are not necessarily single; those who reported larger family sizes tended to be associated with smaller values of the browsing index.

The downscale profile of browsers relative to other mall patrons indicates that specialty stores in malls should emphasize moderately priced products. This may explain the historically low rate of failure in malls among such stores and the tendency of high-priced specialty shops to be located in only the prestigious malls or upscale nonenclosed shopping centers.[23] ■

## MULTICOLLINEARITY

Stepwise regression and multiple regression are complicated by the presence of multicollinearity. Virtually all multiple regression analyses done in marketing research involve predictors or independent variables that are related. However, ***multicollinearity*** arises when intercorrelations among the predictors are very high. Multicollinearity can result in several problems, including:

***multicollinearity***
A state of very high intercorrelations among independent variables.

1. The partial regression coefficients may not be estimated precisely. The standard errors are likely to be high.
2. The magnitudes as well as the signs of the partial regression coefficients may change from sample to sample.
3. It becomes difficult to assess the relative importance of the independent variables in explaining the variation in the dependent variable.
4. Predictor variables may be incorrectly included or removed in stepwise regression.

What constitutes serious multicollinearity is not always clear, although several rules of thumb and procedures have been suggested in the literature. Procedures of varying complexity have also been suggested to cope with multicollinearity.[24] A simple procedure consists of using only one of the variables in a highly correlated set of variables. Alternatively, the set of independent variables can be transformed into a new set of predictors that are mutually independent by using techniques such as principal components analysis (see Chapter 19). More specialized techniques, such as ridge regression and latent root regression, can also be used.[25]

## RELATIVE IMPORTANCE OF PREDICTORS

When multicollinearity is present, special care is required in assessing the relative importance of independent variables. In applied marketing research, it is valuable to determine the *relative importance of the predictors*. In other words, how important are the independent variables in accounting for the variation in the criterion or dependent variable?[26] Unfortunately, because the predictors are correlated, there is no unambiguous measure of relative importance of the predictors in regression analysis.[27] However, several approaches are commonly used to assess the relative importance of predictor variables.

**ACTIVE RESEARCH**

Visit *www.hp.com* and conduct an Internet search using a search engine and your library's online database to obtain information on the factors consumers use to evaluate competing brands of laptop computers.

As the marketing director for HP computers, how would you improve the image and competitive positioning of your brand?

Formulate a multiple regression model explaining consumer preferences for laptop computer brands as a function of the brand evaluations on the consumer choice criteria factors used to evaluate competing brands.

1. ***Statistical significance.*** If the partial regression coefficient of a variable is not significant, as determined by an incremental $F$ test, that variable is judged to be unimportant. An exception to this rule is made if there are strong theoretical reasons for believing that the variable is important.
2. ***Square of the simple correlation coefficient.*** This measure, $r^2$, represents the proportion of the variation in the dependent variable explained by the independent variable in a bivariate relationship.
3. ***Square of the partial correlation coefficient.*** This measure, $R^2_{yxi \cdot x_j x_k}$, is the coefficient of determination between the dependent variable and the independent variable, controlling for the effects of the other independent variables.
4. ***Square of the part correlation coefficient.*** This coefficient represents an increase in $R^2$ when a variable is entered into a regression equation that already contains the other independent variables.
5. ***Measures based on standardized coefficients or beta weights.*** The most commonly used measures are the absolute values of the beta weights, $|B_i|$, or the squared values, $B_i^2$. Because they are partial coefficients, beta weights take into account the effect of the other independent variables. These measures become increasingly unreliable as the correlations among the predictor variables increase (multicollinearity increases).
6. ***Stepwise regression.*** The order in which the predictors enter or are removed from the regression equation is used to infer their relative importance.

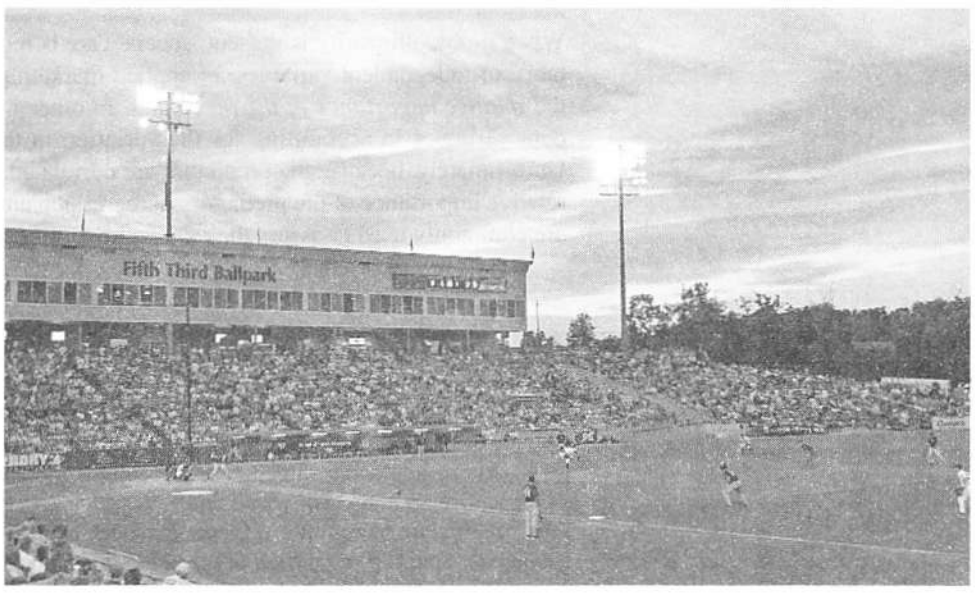
Given that the predictors are correlated, at least to some extent, in virtually all regression situations, none of these measures is satisfactory. It is also possible that the different measures may indicate a different order of importance of the predictors.[28] Yet, if all the measures are examined collectively, useful insights may be obtained into the relative importance of the predictors.

## DECISION RESEARCH

### *The West Michigan Whitecaps: Fanning Fan Loyalty*

#### The Situation

The West Michigan Whitecaps *(www.whitecaps-baseball.com)*, a minor league baseball team in Grand Rapids, wondered what they should do to develop fan loyalty. How could they best keep it, make it grow, and take advantage of it? General Manager Scott Lane got Message Factors *(www.messagefactors.com)*, a Memphis, TN–based research firm, to help them determine how to effectively maintain fan loyalty on a limited budget. Message Factors developed a study that used a proprietary value analysis technique that would examine the relationship between the overall perceived value and specific satisfaction




Regression analysis can help the West Michigan Whitecaps determine the value drivers and enhance the value of Whitecaps games to the fans.

attributes in order to determine loyalty drivers. It helps determine the four things your customers want to tell you, which are: the basics—what customers expect of the company; value issues—what customers value about the company; irritations—what customers do not like about the company; and unimportants—what customers do not care about.

Qualitative research was conducted to identify a set of 71 attributes that influenced fan loyalty. Next, a questionnaire designed to incorporate the 71 attributes was administered to fans at Whitecaps games. The questionnaire was administered to 1,010 respondents. From this, the marketing research company was able to determine the information they were looking for. The basics were determined to be values such as stadium safety, restroom cleanliness, and variety in the food items available. The Whitecaps not only want to meet these basic expectations, but surpass them to guarantee fans will return and be loyal. The value issues are the ones that can really help the team build loyalty. These included things like helpful box office personnel, convenience of purchasing tickets, convenience of parking, and providing the opportunity for autographs. Irritations were determined to involve souvenir price, quality, and lack of variety. However, the research also showed that fans don't really expect to be pleased with this area of sports attendance. It was also determined that there were no unimportant aspects in this survey.

### The Marketing Research Decision

1. In order to determine the relative importance of value drivers, what type of data analysis should Message Factors conduct?
2. Discuss the role of the type of data analysis you recommend in enabling Scott Lane to determine the relative importance of the four value drivers.

### The Marketing Management Decision

1. In order to enhance the value of Whitecaps games to the fans, what should Scott Lane do?
2. Discuss how the marketing management decision action that you recommend to Scott Lane is influenced by the type of data analysis that you suggested earlier and by the findings of that analysis.[29] ■

## CROSS-VALIDATION

Before assessing the relative importance of the predictors or drawing any other inferences, it is necessary to cross-validate the regression model. Regression and other multivariate procedures tend to capitalize on chance variations in the data. This could result in a regression model or equation that is unduly sensitive to the specific data used to estimate the model. One approach for evaluating the model for this, and other problems associated with regression, is cross-validation. ***Cross-validation*** examines whether the regression model continues to hold on comparable data not used in the estimation. The typical cross-validation procedure used in marketing research is as follows:

***cross-validation***
A test of validity that examines whether a model holds on comparable data not used in the original estimation.

1. The regression model is estimated using the entire data set.
2. The available data are split into two parts, the *estimation sample* and the *validation sample*. The estimation sample generally contains 50 to 90 percent of the total sample.
3. The regression model is estimated using the data from the estimation sample only. This model is compared to the model estimated on the entire sample to determine the agreement in terms of the signs and magnitudes of the partial regression coefficients.
4. The estimated model is applied to the data in the validation sample to predict the values of the dependent variable, $\hat{Y}_i$, for the observations in the validation sample.
5. The observed values, $Y_i$, and the predicted values, $\hat{Y}_i$, in the validation sample are correlated to determine the simple $r^2$. This measure, $r^2$, is compared to $R^2$ for the total sample and to $R^2$ for the estimation sample to assess the degree of shrinkage.

***double cross-validation***
A special form of validation in which the sample is split into halves. One half serves as the estimation sample and the other as a validation sample. The roles of the estimation and validation halves are then reversed and the cross-validation process is repeated.

A special form of validation is called double cross-validation. In ***double cross-validation,*** the sample is split into halves. One half serves as the estimation sample, and the other is used as a validation sample in conducting cross-validation. The roles of the estimation and validation halves are then reversed, and the cross-validation is repeated.

# REGRESSION WITH DUMMY VARIABLES

Cross-validation is a general procedure that can be applied even in some special applications of regression, such as regression with dummy variables. Nominal or categorical variables may be used as predictors or independent variables by coding them as dummy variables. The concept of dummy variables was introduced in Chapter 14. In that chapter, we explained how a categorical variable with four categories (heavy, medium, light, and nonusers) can be coded in terms of three dummy variables, $D_1$, $D_2$, and $D_3$, as shown.

| *Product Usage Category* | *Original Variable Code* | *Dummy Variable Code* | | |
|---|---|---|---|---|
| | | $D_1$ | $D_2$ | $D_3$ |
| Nonusers | 1 | 1 | 0 | 0 |
| Light users | 2 | 0 | 1 | 0 |
| Medium users | 3 | 0 | 0 | 1 |
| Heavy users | 4 | 0 | 0 | 0 |

Suppose the researcher were interested in running a regression analysis of the effect of attitude toward the brand on product use. The dummy variables $D_1$, $D_2$, and $D_3$ would be used as predictors. *Regression with dummy variables* would be modeled as:

$$\hat{Y}_i = a + b_1 D_1 + b_2 D_2 + b_3 D_3$$

In this case, "heavy users" has been selected as a reference category and has not been directly included in the regression equation. Note that for heavy users, $D_1$, $D_2$, and $D_3$ assume a value of 0, and the regression equation becomes:

$$\hat{Y}_i = a$$

For nonusers, $D_1 = 1$, and $D_2 = D_3 = 0$, and the regression equation becomes:

$$\hat{Y}_i = a + b_1$$

Thus the coefficient $b_1$ is the difference in predicted $\hat{Y}_i$ for nonusers, as compared to heavy users. The coefficients $b_2$ and $b_3$ have similar interpretations. Although "heavy users" was selected as a reference category, any of the other three categories could have been selected for this purpose.[30]

# ANALYSIS OF VARIANCE AND COVARIANCE WITH REGRESSION

Regression with dummy variables provides a framework for understanding the analysis of variance and covariance. Although multiple regression with dummy variables provides a general procedure for the analysis of variance and covariance, we show only the equivalence of regression with dummy variables to one-way analysis of variance. In regression with dummy variables, the predicted $\hat{Y}$ for each category is the mean of $Y$ for each category. To illustrate using the dummy variable coding of product use we just considered, the predicted $\hat{Y}$ and mean values for each category are as follows:

| *Product Usage Category* | *Predicted Value* $\hat{Y}$ | *Mean Value* $\overline{Y}$ |
|---|---|---|
| Nonusers | $a + b_1$ | $a + b_1$ |
| Light users | $a + b_2$ | $a + b_2$ |
| Medium users | $a + b_3$ | $a + b_3$ |
| Heavy users | $a$ | $a$ |

Given this equivalence, it is easy to see further relationships between dummy variable regression and one-way ANOVA.[31]

| *Dummy Variable Regression* | *One-Way ANOVA* |
|---|---|
| $SS_{res} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ | $= SS_{within} = SS_{error}$ |
| $SS_{reg} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$ | $= SS_{between} = SS_x$ |
| $R^2$ | $= \eta^2$ |
| Overall $F$ test | $= F$ test |

Thus we see that regression in which the single independent variable with $c$ categories has been recoded into $c - 1$ dummy variables is equivalent to one-way analysis of variance. Using similar correspondences, one can also illustrate how $n$-way analysis of variance and analysis of covariance can be performed using regression with dummy variables.

Regression analysis, in its various forms, is a widely used technique. The next example illustrates an application in the context of international marketing research, and the example after that shows how regression can be used in investigating ethics in marketing research.

**REAL RESEARCH**

## *Frequent Fliers—Fly from the Clouds to the Clear*

Airline companies in Asia had been facing uncertainty and tough competition from U.S. carriers for a long time. Asian airlines, hit by high fuel costs, global recession, and preemptive competitive deals, realized they could band together to increase air patronage. Secondary data revealed that among the important factors leading to airline selection by consumers were price, on-time schedules, destinations, deals available, kitchen and food service, on-flight service, and so on. Asian airlines offered these services at par if not better. In fact, research showed that in-flight and kitchen services might have been even better. So why were they feeling the competitive pressure? Qualitative research in the form of focus groups revealed that the frequent flier program was a critical factor for a broad segment in general and the business segment in particular. A survey of international passengers was conducted and multiple regression analysis was used to analyze the data. The likelihood of flying and other choice measures served as the dependent variable and the set of service factors, including the frequent flier program, were the independent variables. The results indicated that frequent flier program indeed had a significant effect on the choice of an airline. Based on these findings, Cathay Pacific, Singapore International Airlines, Thai Airways International, and Malaysian Airlines introduced a cooperative frequent flier program called Asia Plus, available to all travelers. The program was the first time the Asian carriers had offered free travel in return for regular patronage. A multimillion-dollar marketing and advertising campaign was started to promote Asia Plus. Frequent fliers, thus, flew from the clouds to the clear and the Asian airlines experienced increased passenger traffic. Although the frequent flier program proved successful for Asian airlines, the uncertain economy of 2001–2005 pushed them into a huge crisis. The Association of Asia Pacific Airlines (AAPA) said at its annual assembly that the current state of the industry was not encouraging. Despite the challenges ahead for Asian airlines in 2006 to 2010, many believe that it will be possible to renew growth and restore profitability in the future. The director of the AAPA, General Richard Stirland, said, "The industry should seize the opportunity, think the unthinkable, and set a new course to establish a less fragmented and healthier industry."[32] ■

**REAL RESEARCH**

## *Reasons for Researchers Regressing to Unethical Behavior*

As of 2006, the Internet is being used more and more to conduct marketing research studies. Therefore it is crucial that the research community create an ethical code of standards to follow when researching in an online environment. Many online researchers are distressed

at the way other researchers are abusing the Internet as a means of collecting data. Those who conduct online research in an ethical manner feel that an accepted code of ethics of online research and online marketing behavior must be established. Without such a code, dishonest marketing tactics will prevail and ultimately make online research an impractical means of collecting important consumer data. Not only does online marketing research raise ethical problems and concerns, but also traditional marketing research has been targeted as a major source of ethical problems within the discipline of marketing. In particular, marketing research has been charged with engaging in deception, conflict of interest, violation of anonymity, invasion of privacy, data falsifications, dissemination of faulty research findings, and the use of research as a guise to sell merchandise. It has been speculated that when a researcher chooses to participate in unethical activities, that decision may be influenced by organizational factors. Therefore, a study using multiple regression analysis was designed to examine organizational factors as determinants of the incidence of unethical research practices. Six organizational variables were used as the independent variables, namely: extent of ethical problems within the organization, top management actions on ethics, code of ethics, organizational rank, industry category, and organizational role. The respondent's evaluation of the incidence of unethical marketing research practices served as the dependent variable. Regression analysis of the data suggested that four of the six organization variables influenced the extent of unethical research practice: extent of ethical problems within the organization, top management actions on ethics, organizational role, and industry category. Thus, to reduce the incidence of unethical research practice, top management should take stern actions, clarify organizational roles and responsibilities for ethical violations, and address the extent of general ethical problems within the organization.[33] ■

## STATISTICAL SOFTWARE

The computer programs available for conducting correlation analysis are described in Exhibit 17.1. In SPSS, CORRELATE can be used for computing Pearson product moment correlations, PARTIAL CORR for partial correlations, and NONPAR CORR for Spearman's $\rho_s$ and Kendall's $\tau$. The SAS program CORR can be used for calculating Pearson, Spearman's, Kendall's, and partial correlations. In MINITAB, correlation can be computed using STAT>BASIC STATISTICS>CORRELATION function. It calculates Pearson's

*Exhibit 17.1*
Computer Programs for Correlations

***SPSS***
The CORRELATIONS program computes Pearson product moment correlations with significance levels. Univariate statistics, covariance, and cross-product deviations may also be requested. PARTIAL CORR computes partial correlations. The effects of one or more confounding variables can be controlled when describing the relationship between two variables. Significance levels are included in the output.

***SAS***
CORR produces metric and nonmetric correlations between variables, including Pearson's product moment correlation. It also computes partial correlations.

***MINITAB***
Correlations can be computed using the STAT>BASIC STATISTICS>CORRELATION function. It calculates Pearson's product moment using all the columns. Spearman's ranks the columns first and then performs the correlation on the ranked columns.

To compute partial correlation, use the menu commands STAT> BASIC STATISTICS>CORRELATION and STAT>REGRESSION>REGRESSION. Partial correlations can also be calculated by using session commands.

***Excel***
Correlations can be determined in Excel by using the TOOLS>DATA ANALYSIS> CORRELATION function. Utilize the Correlation Worksheet function when a correlation coefficient for two cell ranges is needed. There is no separate function for partial correlations.

*Exhibit 17.2*
Computer Programs for Regression

*SPSS*
REGRESSION calculates bivariate and multiple regression equations, associated statistics, and plots. It allows for an easy examination of residuals. Stepwise regression can also be conducted. Regression statistics can be requested with PLOT, which produces simple scattergrams and some other types of plots.

*SAS*
REG is a general-purpose regression procedure that fits bivariate and multiple regression models using the least-squares procedure. All the associated statistics are computed and residuals can be plotted. Stepwise methods can be implemented. RSREG is a more specialized procedure that fits a quadratic response surface model using least-squares regression. It is useful for determining factor levels that optimize a response. The ORTHOREG procedure is recommended for regression when the data are ill conditioned. GLM uses the method of least squares to fit general linear models and can also be used for regression analysis. NLIN computes the parameters of a nonlinear model using least-squares or weighted least-squares procedures.

*MINITAB*
Regression analysis under the STATS>REGRESSION function can perform simple, polynomial, and multiple analysis. The output includes a linear regression equation, table of coefficients, $R$ square, $R$ squared adjusted, analysis of variance table, a table of fits, and residuals that provide unusual observations. Other available features include stepwise, best subsets, fitted line plot, and residual plots.

*Excel*
Regression can be assessed from the TOOLS>DATA ANALYSIS menu. Depending on the features selected, the output can consist of a summary output table, including an ANOVA table, a standard error of $y$ estimate, coefficients, standard error of coefficients, $R^2$ values, and the number of observations. In addition, the function computes a residual output table, a residual plot, a line fit plot, normal probability plot, and a two-column probability data output table.

product moment. Spearman's ranks the columns first and then performs the correlation on the ranked columns. To compute partial correlation, use the menu commands STAT>BASIC STATISTICS>CORRELATION and STAT>REGRESSION>REGRESSION. Correlations can be determined in Excel by using the TOOLS>DATA ANALYSIS>CORRELATION function. Utilize the Correlation Worksheet function when a correlation coefficient for two cell ranges is needed. There is no separate function for partial correlations.

As described in Exhibit 17.2, these packages contain several programs for performing regression analysis, calculating the associated statistics, performing tests for significance, and plotting the residuals. In SPSS, the main program is REGRESSION. In SAS, the most general program is REG. Other specialized programs such as RSREG, ORTHOREG, GLM, and NLIN are also available, but readers not familiar with the intricate aspects of regression analysis are advised to stick to REG when using SAS. In MINITAB, regression analysis under the STATS>REGRESSION function can perform simple, polynomial, and multiple analysis. In Excel, regression can be assessed from the TOOLS>DATA ANALYSIS menu.

## SPSS WINDOWS

The CORRELATE program computes Pearson product moment correlations and partial correlations with significance levels. Univariate statistics, covariance, and cross-product deviations may also be requested. Significance levels are included in the output. To select this procedure using SPSS for Windows, click:

Analyze>Correlate>Bivariate . . .
Analyze>Correlate>Partial . . .

Scatterplots can be obtained by clicking:

Graphs>Scatter . . . >Simple>Define

The following are the detailed steps for running a correlation between attitude toward the city and duration of residence given in Table 17.1. The corresponding screen captures for these steps can be downloaded from the Web site for this book. A positive correlation is to be expected.

1. Select ANALYZE from the SPSS menu bar.
2. Click CORRELATE and then BIVARIATE.
3. Move "Attitude[attitude]" into the VARIABLES box. Then move "Duration[duration]" into the VARIABLES box.
4. Check PEARSON under CORRELATION COEFFICIENTS.
5. Check ONE_TAILED under TEST OF SIGNIFICANCE.
6. Check FLAG SIGNIFICANT CORRELATIONS.
7. Click OK.

REGRESSION calculates bivariate and multiple regression equations, associated statistics, and plots. It allows for an easy examination of residuals. This procedure can be run by clicking:

Analyze>Regression>Linear . . .

The following are the detailed steps for running a bivariate regression with attitude toward the city as the dependent variable and duration of residence as the independent variable using the data of Table 17.1. The corresponding screen captures for these steps can be downloaded from the Web site for this book.

1. Select ANALYZE from the SPSS menu bar.
2. Click REGRESSION and then LINEAR.
3. Move "Attitude[attitude]" into the DEPENDENT box.
4. Move "Duration[duration]" into the INDEPENDENT(S) box.
5. Select ENTER in the METHOD box.
6. Click on STATISTICS and check ESTIMATES under REGRESSION COEFFICIENTS.
7. Check MODEL FIT.
8. Click CONTINUE.
9. Click OK.

The steps for running multiple regression are similar except for step 4. In step 4, move "Duration[duration]" and "Importance[importance]" into the INDEPENDENT(S) box. The corresponding screen captures for these steps can be downloaded from the Web site for this book.

## PROJECT RESEARCH

### *Multiple Regression*

In the department store project, multiple regression analysis was used to develop a model that explained store preference in terms of respondents' evaluations of the store on the eight choice criteria. The dependent variable was preference for each store. The independent variables were the evaluations of each store on quality of merchandise, variety and assortment of merchandise, returns and adjustment policy, service of store personnel, prices, convenience of location, layout of store, and credit and billing policies. The results indicated that all the factors of the choice criteria, except service of store personnel, were significant in explaining store preference. The coefficients of all the variables were positive, indicating that higher evaluations on each of the significant factors led to higher preference for that store. The model had a good fit and good ability to predict store preference.


SPSS Data File

#### Project Activities

Download the SPSS data file Sears Data 17 from the Web site for this book. This file contains the evaluation of Sears on the eight factors of the choice criteria (quality, variety and assortment, return policy, service of store personnel, fair prices, convenience of location, store lay-

out, and credit and billing policies), the preference for Sears, importance attached to eight factors of the choice criteria, and agreement with the 21 lifestyle statements. The measurement of these variables is described in Chapter 1. The remaining variables have not been included to keep the number of variables below 50 so that you can use the student SPSS software.

1. Run product moment correlations between the evaluation of Sears on the eight factors of the choice criteria and the preference for Sears.
2. Run a multiple regressions, with preference for Sears as the dependent variable and evaluations of Sears on the eight factors of the choice criteria as the independent variables. Interpret the results. ■

## EXPERIENTIAL RESEARCH


**SPSS Data File**

Download the Dell case and questionnaire from the Web site for this book. This information is also given at the end of the book. Download the Dell SPSS data file.

1. Can the overall satisfaction (q4) be explained in terms of all 13 evaluations of Dell (q8_1 to q8_13) when the independent variables are considered simultaneously? Interpret the results.
2. Can the likelihood of choosing Dell (q6) be explained in terms of all 13 evaluations of Dell (q8_1 to q8_13) when the independent variables are considered simultaneously? Interpret the results.
3. Can price sensitivity ratings of q9_5per be explained in terms of all 13 evaluations of Dell (q8_1 to q8_13) when the independent variables are considered simultaneously? Interpret the results.
4. Can price sensitivity ratings of q9_10per be explained in terms of all 13 evaluations of Dell (q8_1 to q8_13) when the independent variables are considered simultaneously? Interpret the results. ■

# REGRESSION ANALYSIS FOR IDENTIFICATION AND ANALYSIS OF MODERATOR–MEDIATOR VARIABLE

## Regression Analysis for Identification and Analysis of Moderator Variable

***Moderator Variable: What is it?*** To comprehend moderator variable the seminal study of Baron and Kenny[34] has been referred. According to Baron and Kenny[34] "a moderator variable is a qualitative (e.g., sex, race, and class) or quantitative (e.g., level of reward) variable that affects the direction or strength of the relation between an independent or predictor variable and a dependent or criterion variable" (p. 1174). For example, cultural values have been empirically verified as moderating variables by several buyer–seller relationship models.[35-37] In the similar vein, several other variables such as product, price,[38] length of the buyer–salesperson relationship,[39] market turbulence and technological change[40] have been discussed as moderator variables in prior marketing studies.

In the moderator model (Figure 17.10), there are three different variables that help predict the dependent variable. First, there is the main effect of predictor (path a; Figure 17.10). Second, there is a direct effect of the moderator (path b; Figure 17.10). This

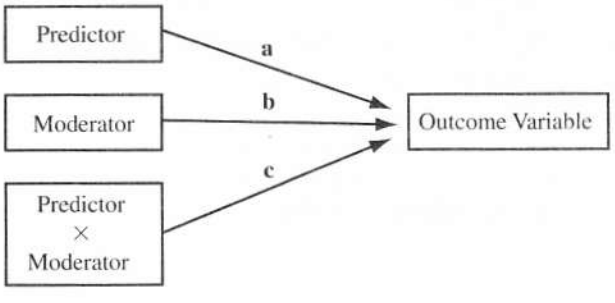



*Figure 17.10*
Moderator Model (Source: Baron and Kenny, 1986)

is similar to the main effect of a predictor and the moderator itself can help predict the dependent variable in its own right. Third, and this is most complex variable, there is an interaction effect between the predictor variable and a moderator variable. This interaction effect is simply calculated by multiplying the value of the predictor by the value of moderator (path c; Figure 17.10). If a moderator effect is positive, then it means that the interaction between the predictor and the moderator is positively related to dependent variable. It implies that the effect of predictor variable on dependent variable will be higher for high values of the moderator. If a moderator effect is negative, then it means that the interaction between the predictor and the moderator is negatively related to dependent variable. It implies that the effect of predictor variable on dependent variable will be lower for higher values of the moderator. A pure moderator variable interacts with the predictor variables while it in itself is not a predictor variable. A quasi-moderator is a variable that not only interacts with the predictor variables but itself is also a predictor variable.[41]

***Identification and Analysis of Moderator Variable.*** To investigate the presence of moderator effect on the form of relationship, hierarchical moderated regression analysis procedure proposed by Sharma et al.[41] has been followed. The primary objective of the analysis is to investigate the presence of pure or quasi-moderating effects on the form of relationship between a predictor and criterion variable. The analysis specifies three different equations, which are as follows:

$$Y = a + b_1 x \quad (1)$$
$$Y = a + b_1 x + b_2 z \quad (2)$$
$$Y = a + b_1 x + b_2 z + b_3 xz \quad (3)$$

Where, "$x$" is the predictor variable, "$z$" is the potential moderator, and "$xz$" is the multiplicative interaction term and "$Y$" is the criterion variable.

In Eq. (1), the hypothesized predictor variable(s) are entered simultaneously in order to investigate the main effects (path a in the moderator model depicted in Figure 17.10) on the criterion variable.

In Eq. (2), the hypothesized moderator variable(s) are entered stepwise, in order to investigate whether or not these variables feature main effects (path b in the moderator model depicted in Figure 17.10).

In Eq. (3), the interaction term(s) (product of predictor and mediator) are entered step wise, in order to investigate hypothesized moderator effects (path c in the moderator model depicted in Figure 17.10).

At each stage of the analysis the significance of the regression coefficients ($b_1$, $b_2$, and $b_3$ in above three equations) are examined, along with the change in $R^2$ from Eqs (1) to (2), and from (2) to (3).

A pure moderator effect is said to been identified when $b_2$ is not significant, but $b_3$ is significant, so that $\Delta R^2$ is accounted for entirely by the interaction term. This fulfils the conditions for a pure moderator, which interacts with a predictor to produce a moderator effect on the strength of relationship between the predictor and criterion variable, but does not feature in the main effect on criterion variable, and therefore is not another predictor.

A quasi-moderator effect is identified when both $b_2$ and $b_3$ are significant, indicating a main effect from the moderator (i.e., itself a predictor variable) as well as has moderating effect on the predictor's relationship with criterion variable.[41]

## Regression for Identification and Analysis of Mediator Variable

***Mediator Variable: What is it?*** Baron and Kenny[34] have defined a given variable may function as a mediator to the extent that it accounts for the relationship between the predictor and the criterion variable. According to Baron and Kenny,[34] mediator explains how external physical events take on internal psychological significance. Where as moderator variables specify when certain effects will hold, mediators speak to how and why such

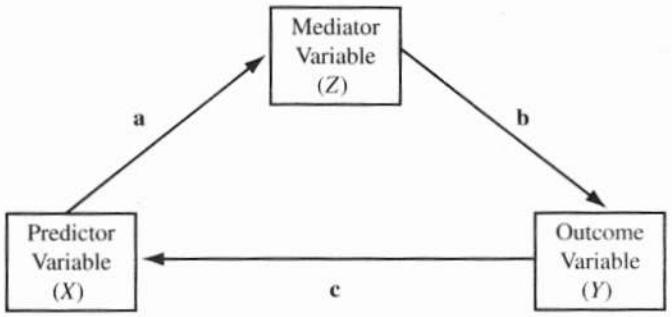



***Figure 17.11***
Mediator Model (Source: Baron and Kenny, 1986)

effects occur. In simpler terms, a moderator variable is one that influences the strength of a relationship between two other variables and a mediator variable is one that explains the relationship between two variables. Mediator represents the generic mechanism through which the independent variables (also known as antecedent variables) are able to positively influence the dependent variable. The antecedent variables do not have a direct effect on the dependent variable, but only an indirect effect through the mediating variable.

Trust has been empirically verified as a mediating variable by several marketing studies.[38, 42, 43] In the similar vein, several other variables, such as relationship quality,[44] perceived communication ability, likeability, expertise,[39] value,[45] and attitude toward ad[46] have been empirically verified as mediator variables in previous marketing literature.

***Testing Mediation with Regression Analysis.*** As shown in Figure 17.11 mediation is a hypothesized causal chain in which predictor variable ($X$) affects mediator variable ($Z$) that, in turn, affects outcome variable ($Y$). The mediator variable ($Z$) also termed as intervening variable mediates the relationship between predictor ($X$) and outcome variable ($Y$). Baron and Kenny[34] proposed a four-step approach in which following series of regression models are to be estimated in order to provide the tests of the linkages of the mediational model.

Step 1: Conduct a regression analysis with $X$ predicting $Z$ (path a):

$$Z = a + b_1X \tag{1}$$

Step 2: Conduct a regression analysis with $X$ predicting $Y$ to test significance of path c:

$$Y = a + b_1X \tag{2}$$

Step 3: Conduct a regression analysis with $X$ and $Z$ predicting $Y$ to access whether the effect of $X$ on $Y$ is significantly reduced in presence of $Z$.

$$Y = a + b_1X + b_2Z \tag{3}$$

To establish mediation effect of variable $Z$, the following condition must hold: First, independent variable ($X$) must affect mediator ($Y$) in the first equation; second, the independent variable ($X$) must affect the outcome variable ($Y$) in the second equation and third, the mediator ($Z$) must affect the outcome variable ($Y$) in the third equation. If all these conditions hold in predicted direction then the effect of independent variable on outcome variable will be certainly lessened in the third equation than in second equation. If the effect of independent variable on outcome variable in presence of mediator variable is reduced (in terms of size of regression coefficient but still significant), the model is consistent with partial mediation. Perfect mediation model holds if the independent variable has no effect in presence of mediator in forth regression model.

Figures 17.12 and 17.13 illustrate full and partial mediating models, respectively. In Figure 17.12 (FMM), there is no direct effect of predictor on outcome variable. The mediator variable accounts for the relationship between independent and dependent variable. The predictor variable influences the criterion variable through key mediating variable trust. As shown in Figure 17.13 (DEM), the mediating variable mediates a portion of effect of independent on dependent variable. The predictor variable in Figure 17.13 has some direct effect on dependent variable apart from a portion of effect mediated by mediating variable.


***Figure 17.12***
Fully Mediated Model (FMM): Perfect Mediation

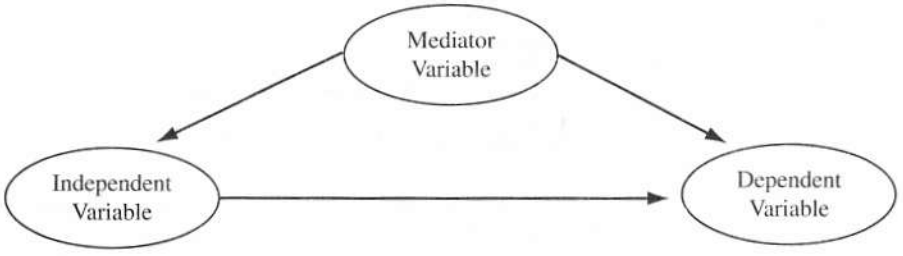



***Figure 17.13***
Direct Effect Model (DEM): Partial Mediation

***Statistical Test of Mediating Effect.*** The indirect effect or mediated effect is calculated by multiplying the two regression unstandardized coefficients **a** and **b** as shown in Figure 17.11. Baron and Kenny[34] present following formula to estimate the standard error ($Se_{ab}$) of indirect effect *ab*:

$$Se_{ab} = \sqrt{b^2Sa^2 + a^2Sb^2 + Sa^2Sb^2}$$

Where, $a$ is the path from independent variable to mediator variable and its standard error is $S_a$; the path from the mediator to the dependent variable is denoted as $b$ and its standard error is $S_b$.

To test statistical significance* of the indirect effect ($ab$), the $z$ score is calculated as:

$$Z_{ab} = \frac{ab}{Se_{ab}}$$

# SUMMARY

The product moment correlation coefficient, $r$, measures the linear association between two metric (interval or ratio scaled) variables. Its square, $r^2$, measures the proportion of variation in one variable explained by the other. The partial correlation coefficient measures the association between two variables after controlling, or adjusting for, the effects of one or more additional variables. The order of a partial correlation indicates how many variables are being adjusted or controlled. Partial correlations can be very helpful for detecting spurious relationships.

Bivariate regression derives a mathematical equation between a single metric criterion variable and a single metric predictor variable. The equation is derived in the form of a straight line by using the least-squares procedure. When the regression is run on standardized data, the intercept assumes a value of 0, and the regression coefficients are called beta weights. The strength of association is measured by the coefficient of determination, $r^2$, which is obtained by computing a ratio of $SS_{reg}$ to $SS_y$. The standard error of estimate is used to assess the accuracy of prediction and may be interpreted as a kind of average error made in predicting $Y$ from the regression equation.

Multiple regression involves a single dependent variable and two or more independent variables. The partial regression coefficient, $b_1$, represents the expected change in $Y$ when $X_1$ is changed by one unit and $X_2$ through $X_k$ are held constant. The strength of association is measured by the coefficient of multiple determination, $R^2$. The significance of the overall regression equation may be tested by the overall $F$ test. Individual partial regression coefficients may be tested for significance using the $t$ test or the incremental $F$ test. Scattergrams of the residuals, in which the residuals are plotted against the predicted values, $\hat{Y}_i$, time, or predictor variables, are useful for examining the appropriateness of the underlying assumptions and the regression model fitted.

In stepwise regression, the predictor variables are entered or removed from the regression equation one at a time for the purpose of selecting a smaller subset of predictors that account for most of the variation in the criterion variable. Multicollinearity, or very high intercorrelations among the predictor variables, can result in several problems. Because the predictors are correlated, regression analysis provides no unambiguous measure of relative importance of the predictors. Cross-validation examines whether the regression model continues to hold true for comparable data not used in estimation. It is a useful procedure for evaluating the regression model.

Nominal or categorical variables may be used as predictors by coding them as dummy variables. Multiple regression with dummy variables provides a general procedure for the analysis of variance and covariance.

*The $z$ score larger than 1.96 in absolute value is significant at the 0.05 level.

# KEY TERMS AND CONCEPTS

product moment correlation (r), *536*
covariance, *537*
partial correlation coefficient, *540*
part correlation coefficient, *541*
nonmetric correlation, *542*
regression analysis, *542*
bivariate regression, *543*
bivariate regression model, *543*
coefficient of determination, *543*
estimated or predicted value, *543*
regression coefficient, *543*
scattergram, *543*
standard error of estimate, *544*
standard error, *544*
standardized regression coefficient, *544*
sum of squared errors, *544*
*t* statistic, *544*
least-squares procedure, *545*
multiple regression, *552*
multiple regression model, *553*
adjusted $R^2$, *553*
coefficient of multiple determination, *553*
*F* test, *553*
partial *F* test, *553*
partial regression coefficient, *553*
residual, *557*
stepwise regression, *559*
multicollinearity, *561*
cross-validation, *563*
double cross-validation, *563*

# SUGGESTED CASES, VIDEO CASES, AND HBS CASES

## *Cases*

Case 3.1 Is Celebrity Advertising Worth Celebrating?
Case 3.3 Matsushita Retargets the U.S.A.
Case 3.4 Pampers Curing Its Rash of Market Share
Case 3.5 DaimlerChrysler Seeks a New Image
Case 3.6 Cingular Wireless: A Singular Focus
Case 3.7 IBM: The World's Top Provider of Computer Hardware, Software, and Services
Case 3.8 Kimberly-Clark: Competing Through Innovation
Case 4.1 Wachovia: "Watch Ovah Ya" Finances
Case 4.2 Wendy's: History and Life After Dave Thomas
Case 4.3 Astec: Continuing to Grow
Case 4.4 Is Marketing Research the Cure for Norton Healthcare Kosair Children's Hospital's Ailments?

## *Video Cases*

Video Case 3.1 The Mayo Clinic: Staying Healthy with Marketing Research
Video Case 4.1 Subaru: "Mr. Survey" Monitors Customer Satisfaction
Video Case 4.2 Procter & Gamble: Using Marketing Research to Build Brands

# LIVE RESEARCH: CONDUCTING A MARKETING RESEARCH PROJECT

1. It is desirable to calculate product moment correlations between all interval scaled variables. This gives an idea of the correlations between variables.
2. Run several bivariate regressions and compare these results with the corresponding product moment correlations.
3. Multiple regressions should be run when examining the association between a single dependent variable and several independent variables.

# ACRONYMS

The main features of regression analysis may be summarized by the acronym REGRESSION:

**R** esidual analysis is useful
**E** stimation of parameters: solution of simultaneous equations
**G** eneral model is linear
**$R^2$** strength of association
**E** rror terms are independent and $N(0, s2)$
**S** tandardized regression coefficients
**S** tandard error of estimate: prediction accuracy
**I** ndividual coefficients and overall *F* tests
**O** ptimal: minimizes total error
**N** onstandardized regression coefficients

# EXERCISES

## *Questions*

1. What is the product moment correlation coefficient? Does a product moment correlation of 0 between two variables imply that the variables are not related to each other?
2. What is a partial correlation coefficient?
3. What are the main uses of regression analysis?
4. What is the least-squares procedure?
5. Explain the meaning of standardized regression coefficients.
6. How is the strength of association measured in bivariate regression? In multiple regression?
7. What is meant by prediction accuracy?
8. What is the standard error of estimate?
9. What assumptions underlie bivariate regression?
10. What is multiple regression? How is it different from bivariate regression?
11. Explain the meaning of a partial regression coefficient. Why is it so called?
12. State the null hypothesis in testing the significance of the overall multiple regression equation. How is this null hypothesis tested?
13. What is gained by an examination of residuals?
14. Explain the stepwise regression approach. What is its purpose?
15. What is multicollinearity? What problems can arise because of multicollinearity?
16. What are some of the measures used to assess the relative importance of predictors in multiple regression?
17. Describe the cross-validation procedure. Describe the double cross-validation procedure.
18. Demonstrate the equivalence of regression with dummy variables to one-way ANOVA.

## *Problems*

1. A major supermarket chain wants to determine the effect of promotion on relative competitiveness. Data were obtained from 15 states on the promotional expenses relative to a major competitor (competitor expenses = 100) and on sales relative to this competitor (competitor sales = 100).

| *State No.* | *Relative Promotional Expense* | *Relative Sales* |
|---|---|---|
| 1 | 95 | 98 |
| 2 | 92 | 94 |
| 3 | 103 | 110 |
| 4 | 115 | 125 |
| 5 | 77 | 82 |
| 6 | 79 | 84 |
| 7 | 105 | 112 |
| 8 | 94 | 99 |
| 9 | 85 | 93 |
| 10 | 101 | 107 |
| 11 | 106 | 114 |
| 12 | 120 | 132 |
| 13 | 118 | 129 |
| 14 | 75 | 79 |
| 15 | 99 | 105 |

You are assigned the task of telling the manager whether there is any relationship between relative promotional expense and relative sales.

a. Plot the relative sales ($Y$-axis) against the relative promotional expense ($X$-axis), and interpret this diagram.
b. Which measure would you use to determine whether there is a relationship between the two variables? Why?
c. Run a bivariate regression analysis of relative sales on relative promotional expense.
d. Interpret the regression coefficients.
e. Is the regression relationship significant?
f. If the company matched the competitor in terms of promotional expense (if the relative promotional expense was 100), what would the company's relative sales be?
g. Interpret the resulting $r^2$.

2. To understand the role of quality and price in influencing the patronage of drugstores, 14 major stores in a large metropolitan area were rated in terms of preference to shop, quality of merchandise, and fair pricing. All the ratings were obtained on an 11-point scale, with higher numbers indicating more positive ratings.

| *Store No.* | *Preference* | *Quality* | *Price* |
|---|---|---|---|
| 1 | 6 | 5 | 3 |
| 2 | 9 | 6 | 11 |
| 3 | 8 | 6 | 4 |
| 4 | 3 | 2 | 1 |
| 5 | 10 | 6 | 11 |
| 6 | 4 | 3 | 1 |
| 7 | 5 | 4 | 7 |
| 8 | 2 | 1 | 4 |
| 9 | 11 | 9 | 8 |
| 10 | 9 | 5 | 10 |
| 11 | 10 | 8 | 8 |
| 12 | 2 | 1 | 5 |
| 13 | 9 | 8 | 5 |
| 14 | 5 | 3 | 2 |

a. Run a multiple regression analysis explaining store preference in terms of quality of merchandise and price.
b. Interpret the partial regression coefficients.
c. Determine the significance of the overall regression.
d. Determine the significance of the partial regression coefficients.
e. Do you think that multicollinearity is a problem in this case? Why or why not?

3. You come across a magazine article reporting the following relationship between annual expenditure on prepared dinners ($PD$) and annual income ($INC$):

$$PD = 23.4 + 0.003\ INC$$

The coefficient of the $INC$ variable is reported as significant.

a. Does this relationship seem plausible? Is it possible to have a coefficient that is small in magnitude and yet significant?
b. From the information given, can you tell how good the estimated model is?
c. What are the expected expenditures on prepared dinners of a family earning $30,000?
d. If a family earning $40,000 spent $130 annually on prepared dinners, what is the residual?
e. What is the meaning of a negative residual?

## INTERNET AND COMPUTER EXERCISES

SPSS Data File

1. Conduct the following analyses for the Nike data given in Internet and Computer Exercises 1 of Chapter 15.
   a. Calculate the simple correlations between awareness, attitude, preference, intention, and loyalty toward Nike and interpret the results.
   b. Run a bivariate regression with loyalty as the dependent variable and intention as the independent variable. Interpret the results.
   c. Run a multiple regression with loyalty as the dependent variable and awareness, attitude, preference, and intention as the independent variables. Interpret the results. Compare the coefficients for intention obtained in bivariate and multiple regressions.
2. Conduct the following analyses for the outdoor lifestyle data given in Internet and Computer Exercises 2 of Chapter 15.
   a. Calculate the simple correlations between $V_1$ to $V_6$ and interpret the results.
   b. Run a bivariate regression with preference for an outdoor lifestyle ($V_1$) as the dependent variable and meeting people ($V_6$) as the independent variable. Interpret the results.
   c. Run a multiple regression with preference for an outdoor lifestyle as the dependent variable and $V_2$ to $V_6$ as the independent variables. Interpret the results. Compare the coefficients for $V_6$ obtained in the bivariate and the multiple regressions.
3. In a pretest, data were obtained from 20 respondents on preferences for sneakers on a 7-point scale, 1 = not preferred, 7 = greatly preferred ($V_1$). The respondents also provided their evaluations of the sneakers on comfort ($V_2$), style ($V_3$), and durability ($V_4$), also on 7-point scales, 1 = poor and 7 = excellent. The resulting data are given in the following table.

| $V_1$ | $V_2$ | $V_3$ | $V_4$ |
|---|---|---|---|
| 6.00 | 6.00 | 3.00 | 5.00 |
| 2.00 | 3.00 | 2.00 | 4.00 |
| 7.00 | 5.00 | 6.00 | 7.00 |
| 4.00 | 6.00 | 4.00 | 5.00 |
| 1.00 | 3.00 | 2.00 | 2.00 |
| 6.00 | 5.00 | 6.00 | 7.00 |
| 5.00 | 6.00 | 7.00 | 5.00 |
| 7.00 | 3.00 | 5.00 | 4.00 |
| 2.00 | 4.00 | 6.00 | 3.00 |
| 3.00 | 5.00 | 3.00 | 6.00 |
| 1.00 | 3.00 | 2.00 | 3.00 |
| 5.00 | 4.00 | 5.00 | 4.00 |
| 2.00 | 2.00 | 1.00 | 5.00 |
| 4.00 | 5.00 | 4.00 | 6.00 |
| 6.00 | 5.00 | 4.00 | 7.00 |
| 3.00 | 3.00 | 4.00 | 2.00 |
| 4.00 | 4.00 | 3.00 | 2.00 |
| 3.00 | 4.00 | 3.00 | 2.00 |
| 4.00 | 4.00 | 3.00 | 2.00 |
| 2.00 | 3.00 | 2.00 | 4.00 |

   a. Calculate the simple correlations between $V_1$ to $V_4$ and interpret the results.
   b. Run a bivariate regression with preference for sneakers ($V_1$) as the dependent variable and evaluation on comfort ($V_2$) as the independent variable. Interpret the results.
   c. Run a bivariate regression with preference for sneakers ($V_1$) as the dependent variable and evaluation on style ($V_3$) as the independent variable. Interpret the results.
   d. Run a bivariate regression with preference for sneakers ($V_1$) as the dependent variable and evaluation on durability ($V_4$) as the independent variable. Interpret the results.
   e. Run a multiple regression with preference for sneakers ($V_1$) as the dependent variable and $V_2$ to $V_4$ as the independent variables. Interpret the results. Compare the coefficients for $V_2$, $V_3$, and $V_4$ obtained in the bivariate and the multiple regressions.
4. Use an appropriate microcomputer or mainframe program (SPSS, SAS, MINITAB, or Excel) to analyze the data for:
   a. Problem 1
   b. Problem 2
   c. Fieldwork exercise

## ACTIVITIES

### *Fieldwork*

1. Visit 10 different drugstores in your area. Evaluate each store in terms of its overall image and quality of in-store service using 11-point rating scales (1 = poor, 11 = excellent). Then analyze the data you have collected as follows:
   a. Plot the overall image (*Y*-axis) against relative in-store service (*X*-axis) and interpret this diagram.
   b. Which measure would you use to determine whether there is a relationship between the two variables? Why?
   c. Run a bivariate regression analysis of overall image on in-store service.
   d. Interpret the regression coefficients.
   e. Is the regression relationship significant?
   f. Interpret the resulting $r^2$.

### *Group Discussion*

1. As a small group, discuss the following statement: "Regression is such a basic technique that it should always be used in analyzing data."
2. As a small group, discuss the relationship among bivariate correlation, bivariate regression, multiple regression, and analysis of variance.

CHAPTER 18

# Discriminant and Logit Analysis

"Often you have measured different groups of respondents on many metric variables. Discriminant analysis is a useful way to answer the questions . . . Are the groups different? . . . On what variables are they most different? . . . Can I predict which group a person belongs to using these variables?"

*Jamie Baker-Prewitt, vice president, decision sciences, Burke, Inc.*

## Objectives

After reading this chapter, the student should be able to:

1. Describe the concept of discriminant analysis, its objectives, and its applications in marketing research.
2. Outline the procedures for conducting discriminant analysis, including the formulation of the problem, estimation of the discriminant function coefficients, determination of significance, interpretation, and validation.
3. Discuss multiple discriminant analysis and the distinction between two-group and multiple discriminant analysis.
4. Explain stepwise discriminant analysis and describe the Mahalanobis procedure.
5. Describe the binary logit model and its relative merits versus discriminant and regression analysis.

## Overview

This chapter discusses the techniques of discriminant analysis and logit analysis. We begin by examining the relationship of discriminant analysis to analysis of variance (Chapter 16) and regression analysis (Chapter 17). We present a model and describe the general procedure for conducting discriminant analysis, with emphasis on formulation, estimation, determination of significance, interpretation, and validation of the results. The procedure is illustrated with an example of two-group discriminant analysis, followed by an example of multiple (three-group) discriminant analysis. The stepwise discriminant analysis procedure is also covered. When the dependent variable is binary, the logit model can also be used instead of two-group discriminant analysis. We explain the logit model and discuss its relative merits versus discriminant and regression analysis.

### REAL RESEARCH

#### *Rebate Redeemers*

A study of 294 consumers was undertaken to determine the correlates of rebate proneness, or the characteristics of consumers who respond favorably to rebate promotions. The predictor variables were four factors related to household shopping attitudes and behaviors, and selected demographic characteristics (sex, age, and income). The dependent variable was the respondent's degree of rebate proneness, of which three levels were identified. Respondents who reported no rebate-triggered purchases during the past 12 months were classified as nonusers; those who reported one or two such purchases as light users; and those with more than two purchases, frequent users of rebates. Multiple discriminant analysis was used to analyze the data.

Two primary findings emerged. First, consumers' perception of the effort/value relationship was the most effective variable in discriminating among frequent, light, and nonusers of rebate offers. Clearly, rebate-sensitive consumers associate less effort with fulfilling the requirements of the rebate purchase, and they are willing to accept a relatively smaller refund than other customers. Second, consumers who are aware of the regular prices of products, so that they recognize bargains, are more likely than others to respond to rebate offers.

These findings were utilized by Toshiba (*www.toshiba.com*) when it offered a $50 cash rebate on its notebook computer Satellite M55 models purchased between June 26, 2005 and September 24, 2005. The company felt that this would encourage the rebate-sensitive customers to choose Toshiba notebooks.[1] ■

The rebate proneness example examined three groups (nonusers, light users, and frequent users of rebates). Significant intergroup differences were found using multiple predictor variables. An examination of differences across groups lies at the heart of the basic concept of discriminant analysis.

Multiple discriminant analysis can help identify the factors that differentiate frequent users, light users, and nonusers of rebates.


## BASIC CONCEPT OF DISCRIMINANT ANALYSIS

***discriminant analysis***
A technique for analyzing marketing research data when the criterion or dependent variable is categorical and the predictor or independent variables are interval in nature.

***discriminant functions***
The linear combination of independent variables developed by discriminant analysis that will best discriminate between the categories of the dependent variable.

***Discriminant analysis*** is a technique for analyzing data when the criterion or dependent variable is categorical and the predictor or independent variables are interval in nature.[2] For example, the dependent variable may be the choice of a brand of personal computer (brand A, B, or C) and the independent variables may be ratings of attributes of PCs on a 7-point Likert scale. The objectives of discriminant analysis are as follows:

1. Development of ***discriminant functions,*** or linear combinations of the predictor or independent variables, which will best discriminate between the categories of the criterion or dependent variable (groups)
2. Examination of whether significant differences exist among the groups, in terms of the predictor variables
3. Determination of which predictor variables contribute to most of the intergroup differences
4. Classification of cases to one of the groups based on the values of the predictor variables
5. Evaluation of the accuracy of classification

Discriminant analysis techniques are described by the number of categories possessed by the criterion variable. When the criterion variable has two categories, the technique is known as ***two-group discriminant analysis.*** When three or more categories are involved, the technique is referred to as ***multiple discriminant analysis.*** The main distinction is that, in the two-group case, it is possible to derive only one discriminant function. In multiple discriminant analysis, more than one function may be computed.[3]

***two-group discriminant analysis***
Discriminant analysis technique where the criterion variable has two categories.

***multiple discriminant analysis***
Discriminant analysis technique where the criterion variable involves three or more categories.

Examples of discriminant analysis abound in marketing research. This technique can be used to answer questions such as:

- In terms of demographic characteristics, how do customers who exhibit store loyalty differ from those who do not?
- Do heavy, medium, and light users of soft drinks differ in terms of their consumption of frozen foods?
- What psychographic characteristics help differentiate between price-sensitive and non-price-sensitive buyers of groceries?
- Do the various market segments differ in their media consumption habits?

- In terms of lifestyles, what are the differences between heavy patrons of regional department store chains and patrons of national chains?
- What are the distinguishing characteristics of consumers who respond to direct mail solicitations?

## RELATIONSHIP OF DISCRIMINANT ANALYSIS TO REGRESSION AND ANOVA

The relationship among discriminant analysis, analysis of variance (ANOVA), and regression analysis is shown in Table 18.1. We explain this relationship with an example in which the researcher is attempting to explain the amount of life insurance purchased in terms of age and income. All three procedures involve a single criterion or dependent variable and multiple predictor or independent variables. However, the nature of these variables differ. In analysis of variance and regression analysis, the dependent variable is metric or interval scaled (amount of life insurance purchased in dollars), whereas in discriminant analysis it is categorical (amount of life insurance purchased classified as high, medium, or low). The independent variables are categorical in the case of analysis of variance (age and income are each classified as high, medium, or low) but metric in the case of regression and discriminant analysis (age in years and income in dollars, i.e., both measured on a ratio scale).

Two-group discriminant analysis, in which the dependent variable has only two categories, is closely related to multiple regression analysis. In this case, multiple regression, in which the dependent variable is coded as a 0 or 1 dummy variable, results in partial regression coefficients that are proportional to discriminant function coefficients (see the following section on the discriminant analysis model). The nature of dependent and independent variables in the binary logit model is similar to that in discriminant analysis.

## DISCRIMINANT ANALYSIS MODEL

***discriminant analysis model***
The statistical model on which discriminant analysis is based.

The ***discriminant analysis model*** involves linear combinations of the following form:

$$D = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_kX_k$$

where

$D$ = discriminant score
$b$'s = discriminant coefficient or weight
$X$'s = predictor or independent variable

The coefficients, or weights ($b$), are estimated so that the groups differ as much as possible on the values of the discriminant function. This occurs when the ratio of between-group sum of squares to within-group sum of squares for the discriminant scores is at a maximum. Any other linear combination of the predictors will result in a smaller ratio.

**TABLE 18.1**
Similarities and Differences Among ANOVA, Regression, and Discriminant Analysis

| | ANOVA | Regression | Discriminant/ Logit Analysis |
|---|---|---|---|
| *Similarities* | | | |
| Number of dependent variables | One | One | One |
| Number of independent variables | Multiple | Multiple | Multiple |
| *Differences* | | | |
| Nature of the dependent variables | Metric | Metric | Categorical/Binary |
| Nature of the independent variables | Categorical | Metric | Metric |

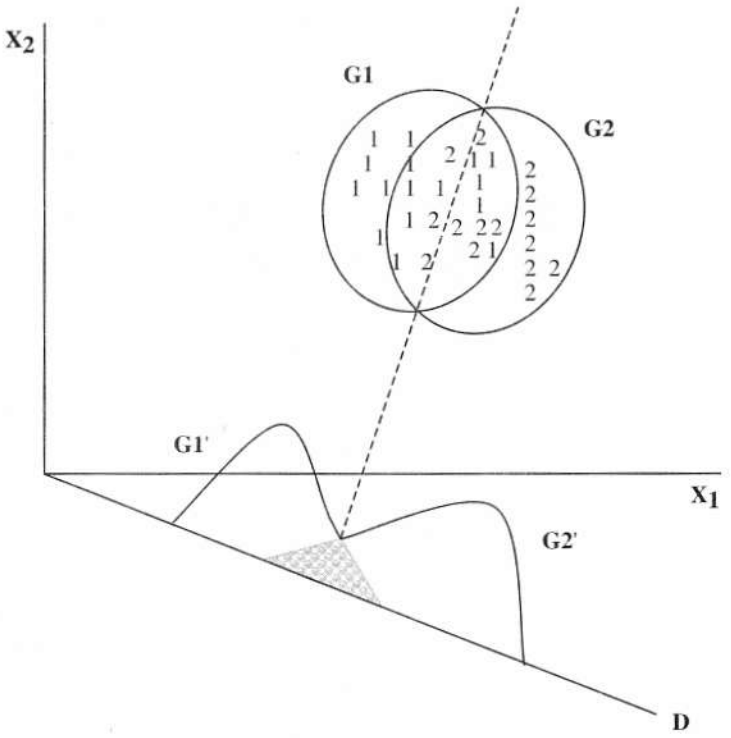



***Figure 18.1***
A Geometric Interpretation of Two-Group Discriminant Analysis

We give a brief geometrical exposition of two-group discriminant analysis. Suppose we had two groups, G1 and G2, and each member of these groups was measured on two variables $X_1$ and $X_2$. A scatter diagram of the two groups is shown in Figure 18.1, where $X_1$ and $X_2$ are the two axes. Members of G1 are denoted by 1 and members of G2 by 2. The resultant ellipses encompass some specified percentage of the points (members), say 93 percent in each group. A straight line is drawn through the two points where the ellipses intersect and then projected to a new axis, *D*. The overlap between the univariate distributions G1' and G2', represented by the shaded area in Figure 18.1, is smaller than would be obtained by any other line drawn through the ellipses representing the scatter plots. Thus, the groups differ as much as possible on the *D* axis. Several statistics are associated with discriminant analysis.

## STATISTICS ASSOCIATED WITH DISCRIMINANT ANALYSIS

The important statistics associated with discriminant analysis include the following.

***Canonical correlation.*** Canonical correlation measures the extent of association between the discriminant scores and the groups. It is a measure of association between the single discriminant function and the set of dummy variables that define the group membership.

***Centroid.*** The centroid is the mean values for the discriminant scores for a particular group. There are as many centroids as there are groups, because there is one for each group. The means for a group on all the functions are the *group centroids.*

***Classification matrix.*** Sometimes also called *confusion* or *prediction matrix*, the classification matrix contains the number of correctly classified and misclassified cases. The correctly classified cases appear on the diagonal, because the predicted and actual groups are the same. The off-diagonal elements represent cases that have been incorrectly classified. The sum of the diagonal elements divided by the total number of cases represents the *hit ratio.*

***Discriminant function coefficients.*** The discriminant function coefficients (unstandardized) are the multipliers of variables, when the variables are in the original units of measurement.

***Discriminant scores.*** The unstandardized coefficients are multiplied by the values of the variables. These products are summed and added to the constant term to obtain the discriminant scores.

***Eigenvalue.*** For each discriminant function, the eigenvalue is the ratio of between-group to within-group sums of squares. Large eigenvalues imply superior functions.

***F values and their significance.*** These are calculated from a one-way ANOVA, with the grouping variable serving as the categorical independent variable. Each predictor, in turn, serves as the metric dependent variable in the ANOVA.

***Group means and group standard deviations.*** These are computed for each predictor for each group.

***Pooled within-group correlation matrix.*** The pooled within-group correlation matrix is computed by averaging the separate covariance matrices for all the groups.

***Standardized discriminant function coefficients.*** The standardized discriminant function coefficients are the discriminant function coefficients and are used as the multipliers when the variables have been standardized to a mean of 0 and a variance of 1.

***Structure correlations.*** Also referred to as *discriminant loadings,* the structure correlations represent the simple correlations between the predictors and the discriminant function.

***Total correlation matrix.*** If the cases are treated as if they were from a single sample and the correlations computed, a total correlation matrix is obtained.

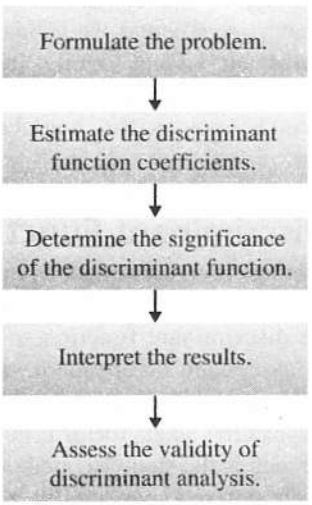
***Wilks' λ.*** Sometimes also called the $U$ statistic, Wilks' $\lambda$ for each predictor is the ratio of the within-group sum of squares to the total sum of squares. Its value varies between 0 and 1. Large values of $\lambda$ (near 1) indicate that group means do not seem to be different. Small values of $\lambda$ (near 0) indicate that the group means seem to be different.

The assumptions in discriminant analysis are that each of the groups is a sample from a multivariate normal population and all of the populations have the same covariance matrix. The role of these assumptions and the statistics just described can be better understood by examining the procedure for conducting discriminant analysis.

## CONDUCTING DISCRIMINANT ANALYSIS

The steps involved in conducting discriminant analysis consist of formulation, estimation, determination of significance, interpretation, and validation (see Figure 18.2). These steps are discussed and illustrated within the context of two-group discriminant analysis. Discriminant analysis with more than two groups is discussed later in this chapter.

*Figure 18.2*
Conducting Discriminant Analysis



## Formulate the Problem

The first step in discriminant analysis is to formulate the problem by identifying the objectives, the criterion variable, and the independent variables. The criterion variable must consist of two or more mutually exclusive and collectively exhaustive categories. When the dependent variable is interval or ratio scaled, it must first be converted into categories. For example, attitude toward the brand, measured on a 7-point scale, could be categorized as unfavorable (1, 2, 3), neutral (4), or favorable (5, 6, 7). Alternatively, one could plot the distribution of the dependent variable and form groups of equal size by determining the appropriate cutoff points for each category. The predictor variables should be selected based on a theoretical model or previous research, or, in the case of exploratory research, the experience of the researcher should guide their selection.

*analysis sample*
Part of the total sample that is used for estimation of the discriminant function.

*validation sample*
That part of the total sample used to check the results of the estimation sample.

The next step is to divide the sample into two parts. One part of the sample, called the estimation or ***analysis sample,*** is used for estimation of the discriminant function. The other part, called the *holdout* or ***validation sample,*** is reserved for validating the discriminant function. When the sample is large enough, it can be split in half. One half serves as the analysis sample and the other is used for validation. The role of the halves is then interchanged and the analysis is repeated. This is called *double cross-validation* and is similar to the procedure discussed in regression analysis (Chapter 17).

Often the distribution of the number of cases in the analysis and validation samples follows the distribution in the total sample. For instance, if the total sample contained 50 percent loyal and 50 percent nonloyal consumers, then the analysis and validation samples would each contain 50 percent loyal and 50 percent nonloyal consumers. On the other hand, if the sample contained 25 percent loyal and 75 percent nonloyal consumers, the analysis and validation samples would be selected to reflect the same distribution (25 percent versus 75 percent).

Finally, it has been suggested that the validation of the discriminant function should be conducted repeatedly. Each time, the sample should be split into different analysis and validation parts. The discriminant function should be estimated and the validation analysis carried out. Thus, the validation assessment is based on a number of trials. More rigorous methods have also been suggested.[4]

To better illustrate two-group discriminant analysis, let us look at an example. For illustrative purposes, we consider only a small number of observations. In actual practice, discriminant analysis is performed on a much larger sample such as that in the Dell Experiential Research considered later. Suppose we want to determine the salient characteristics of families that have visited a vacation resort during the last two years. Data were obtained from a pretest sample of 42 households. Of these, 30 households shown in Table 18.2 were included in the analysis sample and the remaining 12 shown in Table 18.3 were part of the validation sample. The households that visited a resort during the last two years are coded as 1; those that did not, as 2 (VISIT). Both the analysis and validation samples were balanced in terms of VISIT. As can be seen, the analysis sample contains 15 households in each category, whereas the validation sample has six in each category. Data were also obtained on annual family income (INCOME), attitude toward travel (TRAVEL, measured on a 9-point scale), importance attached to family vacation (VACATION, measured on a 9-point scale), household size (HSIZE), and age of the head of the household (AGE).

## Estimate the Discriminant Function Coefficients

*direct method*
An approach to discriminant analysis that involves estimating the discriminant function so that all the predictors are included simultaneously.

*stepwise discriminant analysis*
Discriminant analysis in which the predictors are entered sequentially based on their ability to discriminate between the groups.

Once the analysis sample has been identified, as in Table 18.2, we can estimate the discriminant function coefficients. Two broad approaches are available. The ***direct method*** involves estimating the discriminant function so that all the predictors are included simultaneously. In this case, each independent variable is included, regardless of its discriminating power. This method is appropriate when, based on previous research or a theoretical model, the researcher wants the discrimination to be based on all the predictors. An alternative approach is the stepwise method. In ***stepwise discriminant analysis,*** the predictor variables are entered sequentially, based on their ability to discriminate

SPSS Data File

**TABLE 18.2**

**Information on Resort Visits: Analysis Sample**

| No. | Resort Visit | Annual Family Income ($000) | Attitude Toward Travel | Importance Attached to Family Vacation | Household Size | Age of Head of Household | Amount Spent on Family Vacation |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 50.2 | 5 | 8 | 3 | 43 | M (2) |
| 2 | 1 | 70.3 | 6 | 7 | 4 | 61 | H (3) |
| 3 | 1 | 62.9 | 7 | 5 | 6 | 52 | H (3) |
| 4 | 1 | 48.5 | 7 | 5 | 5 | 36 | L (1) |
| 5 | 1 | 52.7 | 6 | 6 | 4 | 55 | H (3) |
| 6 | 1 | 75.0 | 8 | 7 | 5 | 68 | H (3) |
| 7 | 1 | 46.2 | 5 | 3 | 3 | 62 | M (2) |
| 8 | 1 | 57.0 | 2 | 4 | 6 | 51 | M (2) |
| 9 | 1 | 64.1 | 7 | 5 | 4 | 57 | H (3) |
| 10 | 1 | 68.1 | 7 | 6 | 5 | 45 | H (3) |
| 11 | 1 | 73.4 | 6 | 7 | 5 | 44 | H (3) |
| 12 | 1 | 71.9 | 5 | 8 | 4 | 64 | H (3) |
| 13 | 1 | 56.2 | 1 | 8 | 6 | 54 | M (2) |
| 14 | 1 | 49.3 | 4 | 2 | 3 | 56 | H (3) |
| 15 | 1 | 62.0 | 5 | 6 | 2 | 58 | H (3) |
| 16 | 2 | 32.1 | 5 | 4 | 3 | 58 | L (1) |
| 17 | 2 | 36.2 | 4 | 3 | 2 | 55 | L (1) |
| 18 | 2 | 43.2 | 2 | 5 | 2 | 57 | M (2) |
| 19 | 2 | 50.4 | 5 | 2 | 4 | 37 | M (2) |
| 20 | 2 | 44.1 | 6 | 6 | 3 | 42 | M (2) |
| 21 | 2 | 38.3 | 6 | 6 | 2 | 45 | L (1) |
| 22 | 2 | 55.0 | 1 | 2 | 2 | 57 | M (2) |
| 23 | 2 | 46.1 | 3 | 5 | 3 | 51 | L (1) |
| 24 | 2 | 35.0 | 6 | 4 | 5 | 64 | L (1) |
| 25 | 2 | 37.3 | 2 | 7 | 4 | 54 | L (1) |
| 26 | 2 | 41.8 | 5 | 1 | 3 | 56 | M (2) |
| 27 | 2 | 57.0 | 8 | 3 | 2 | 36 | M (2) |
| 28 | 2 | 33.4 | 6 | 8 | 2 | 50 | L (1) |
| 29 | 2 | 37.5 | 3 | 2 | 3 | 48 | L (1) |
| 30 | 2 | 41.3 | 3 | 3 | 2 | 42 | L (1) |

SPSS Data File

**TABLE 18.3**

**Information on Resort Visits: Holdout Sample**

| No. | Resort Visit | Annual Family Income ($000) | Attitude Toward Travel | Importance Attached to Family Vacation | Household Size | Age of Head of Household | Amount Spent on Family Vacation |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 50.8 | 4 | 7 | 3 | 45 | M (2) |
| 2 | 1 | 63.6 | 7 | 4 | 7 | 55 | H (3) |
| 3 | 1 | 54.0 | 6 | 7 | 4 | 58 | M (2) |
| 4 | 1 | 45.0 | 5 | 4 | 3 | 60 | M (2) |
| 5 | 1 | 68.0 | 6 | 6 | 6 | 46 | H (3) |
| 6 | 1 | 62.1 | 5 | 6 | 3 | 56 | H (3) |
| 7 | 2 | 35.0 | 4 | 3 | 4 | 54 | L (1) |
| 8 | 2 | 49.6 | 5 | 3 | 5 | 39 | L (1) |
| 9 | 2 | 39.4 | 6 | 5 | 3 | 44 | H (3) |
| 10 | 2 | 37.0 | 2 | 6 | 5 | 51 | L (1) |
| 11 | 2 | 54.5 | 7 | 3 | 3 | 37 | M (2) |
| 12 | 2 | 38.2 | 2 | 2 | 3 | 49 | L (1) |

**TABLE 18.4**

**Results of Two-Group Discriminant Analysis**

GROUP MEANS

| VISIT | INCOME | TRAVEL | VACATION | HSIZE | AGE |
|---|---|---|---|---|---|
| 1 | 60.52000 | 5.40000 | 5.80000 | 4.33333 | 53.73333 |
| 2 | 41.91333 | 4.33333 | 4.06667 | 2.80000 | 50.13333 |
| Total | 51.21667 | 4.86667 | 4.93333 | 3.56667 | 51.93333 |

GROUP STANDARD DEVIATIONS

| | | | | | |
|---|---|---|---|---|---|
| 1 | 9.83065 | 1.91982 | 1.82052 | 1.23443 | 8.77062 |
| 2 | 7.55115 | 1.95180 | 2.05171 | 0.94112 | 8.27101 |
| Total | 12.79523 | 1.97804 | 2.09981 | 1.33089 | 8.57395 |

POOLED WITHIN-GROUPS CORRELATION MATRIX

| | INCOME | TRAVEL | VACATION | HSIZE | AGE |
|---|---|---|---|---|---|
| INCOME | 1.00000 | | | | |
| TRAVEL | 0.19745 | 1.00000 | | | |
| VACATION | 0.09148 | 0.08434 | 1.00000 | | |
| HSIZE | 0.08887 | −0.01681 | 0.07046 | 1.00000 | |
| AGE | −0.01431 | −0.19709 | 0.01742 | −0.04301 | 1.00000 |

Wilks' λ (U-statistic) and univariate *F* ratio with 1 and 28 degrees of freedom

| VARIABLE | WILKS' λ | *F* | SIGNIFICANCE |
|---|---|---|---|
| INCOME | 0.45310 | 33.80 | 0.0000 |
| TRAVEL | 0.92479 | 2.277 | 0.1425 |
| VACATION | 0.82377 | 5.990 | 0.0209 |
| HSIZE | 0.65672 | 14.64 | 0.0007 |
| AGE | 0.95441 | 1.338 | 0.2572 |

CANONICAL DISCRIMINANT FUNCTIONS

| FUNCTION | EIGENVALUE | PERCENT OF VARIANCE | CUMULATIVE PERCENT | CANONICAL CORRELATION | AFTER FUNCTION | WILKS' λ | CHI-SQUARE | DF | SIG. |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | 0 | 0.3589 | 26.130 | 5 | 0.0001 |
| 1* | 1.7862 | 100.00 | 100.00 | 0.8007 | | | | | |

*Marks the 1 canonical discriminant functions remaining in the analysis.

STANDARD CANONICAL DISCRIMINANT FUNCTION COEFFICIENTS

| | FUNC 1 |
|---|---|
| INCOME | 0.74301 |
| TRAVEL | 0.09611 |
| VACATION | 0.23329 |
| HSIZE | 0.46911 |
| AGE | 0.20922 |

STRUCTURE MATRIX

Pooled within-groups correlations between discriminating variables and canonical discriminant functions (variables ordered by size of correlation within function).

| | FUNC 1 |
|---|---|
| INCOME | 0.82202 |
| HSIZE | 0.54096 |
| VACATION | 0.34607 |
| TRAVEL | 0.21337 |
| AGE | 0.16354 |

*(Continued)*

**TABLE 18.4**

**Results of Two-Group Discriminant Analysis *(Continued)***

UNSTANDARDIZED CANONICAL DISCRIMINANT FUNCTION COEFFICIENTS

| | FUNC 1 |
|---|---|
| INCOME | 0.8476710E-01 |
| TRAVEL | 0.4964455E-01 |
| VACATION | 0.1202813 |
| HSIZE | 0.4273893 |
| AGE | 0.2454380E-01 |
| (constant) | −7.975476 |

CANONICAL DISCRIMINANT FUNCTIONS EVALUATED AT GROUP MEANS (GROUP CENTROIDS)

| GROUP | FUNC 1 |
|---|---|
| 1 | 1.29118 |
| 2 | −1.29118 |

CLASSIFICATION RESULTS

| | | | PREDICTED GROUP MEMBERSHIP | | TOTAL |
|---|---|---|---|---|---|
| | | VISIT | 1 | 2 | |
| Original | Count | 1 | 12 | 3 | 15 |
| | | 2 | 0 | 15 | 15 |
| | % | 1 | 80.0 | 20.0 | 100.0 |
| | | 2 | 0.0 | 100.0 | 100.0 |
| Cross-validated | Count | 1 | 11 | 4 | 15 |
| | | 2 | 2 | 13 | 15 |
| | % | 1 | 73.3 | 26.7 | 100.0 |
| | | 2 | 13.3 | 86.7 | 100.0 |

[a] Cross-validation is done only for those cases in the analysis. In cross-validation, each case is classified by the functions derived from all cases other than that case.
[b] 90.0% of original grouped cases correctly classified.
[c] 80.0% of cross-validated grouped cases correctly classified.

CLASSIFICATION RESULTS FOR CASES NOT SELECTED FOR USE IN THE ANALYSIS (HOLDOUT SAMPLE)

| | ACTUAL GROUP | NO. OF CASES | PREDICTED GROUP MEMBERSHIP | |
|---|---|---|---|---|
| | | | 1 | 2 |
| Group | 1 | 6 | 4 | 2 |
| | | | 66.7% | 33.3% |
| Group | 2 | 6 | 0 | 6 |
| | | | 0.0% | 100.0% |

Percent of grouped cases correctly classified: 83.33%.


**SPSS Output File**

among groups. This method, described in more detail later, is appropriate when the researcher wants to select a subset of the predictors for inclusion in the discriminant function.

The results of running two-group discriminant analysis on the data of Table 18.2 using a popular computer program are presented in Table 18.4. Some intuitive feel for the results may be obtained by examining the group means and standard deviations. It appears that the two groups are more widely separated in terms of income than other variables. There appears to be more of a separation on the importance attached to family vacation than on attitude toward travel. The difference between the two groups on age of the head of the household is small, and the standard deviation of this variable is large.

The pooled within-groups correlation matrix indicates low correlations between the predictors. Multicollinearity is unlikely to be a problem. The significance of the univariate

*F* ratios indicates that when the predictors are considered individually, only income, importance of vacation, and household size significantly differentiate between those who visited a resort and those who did not.

Because there are two groups, only one discriminant function is estimated. The eigenvalue associated with this function is 1.7862 and it accounts for 100 percent of the explained variance. The canonical correlation associated with this function is 0.8007. The square of this correlation, $(0.8007)^2 = 0.64$, indicates that 64 percent of the variance in the dependent variable (VISIT) is explained or accounted for by this model.

## Determine the Significance of Discriminant Function

It would not be meaningful to interpret the analysis if the discriminant functions estimated were not statistically significant. The null hypothesis that, in the population, the means of all discriminant functions in all groups are equal can be statistically tested. In SPSS, this test is based on Wilks' $\lambda$. If several functions are tested simultaneously (as in the case of multiple discriminant analysis), the Wilks' $\lambda$ statistic is the product of the univariate $\lambda$ for each function. The significance level is estimated based on a chi-square transformation of the statistic. In testing for significance in the vacation resort example (see Table 18.4), it may be noted that the Wilks' $\lambda$ associated with the function is 0.3589, which transforms to a chi-square of 26.13 with 5 degrees of freedom. This is significant beyond the 0.05 level. In SAS, an approximate *F* statistic, based on an approximation to the distribution of the likelihood ratio, is calculated. A test of significance is not available in MINITAB. If the null hypothesis is rejected, indicating significant discrimination, one can proceed to interpret the results.[5]

## Interpret the Results

The interpretation of the discriminant weights, or coefficients, is similar to that in multiple regression analysis. The value of the coefficient for a particular predictor depends on the other predictors included in the discriminant function. The signs of the coefficients are arbitrary, but they indicate which variable values result in large and small function values and associate them with particular groups.

Given the multicollinearity in the predictor variables, there is no unambiguous measure of the relative importance of the predictors in discriminating between the groups.[6] With this caveat in mind, we can obtain some idea of the relative importance of the variables by examining the absolute magnitude of the standardized discriminant function coefficients. Generally, predictors with relatively large standardized coefficients contribute more to the discriminating power of the function, as compared with predictors with smaller coefficients, and are, therefore, more important.

Some idea of the relative importance of the predictors can also be obtained by examining the structure correlations, also called *canonical loadings* or *discriminant loadings*. These simple correlations between each predictor and the discriminant function represent the variance that the predictor shares with the function. The greater the magnitude of a structure correlation, the more important the corresponding predictor. Like the standardized coefficients, these correlations must also be interpreted with caution.

An examination of the standardized discriminant function coefficients for the vacation resort example is instructive. Given the low intercorrelations between the predictors, one might cautiously use the magnitudes of the standardized coefficients to suggest that income is the most important predictor in discriminating between the groups, followed by household size and importance attached to family vacation. The same observation is obtained from examination of the structure correlations. These simple correlations between the predictors and the discriminant function are listed in order of magnitude.

The unstandardized discriminant function coefficients are also given. These can be applied to the raw values of the variables in the holdout set for classification purposes. The group centroids, giving the value of the discriminant function evaluated at the group means, are also shown. Group 1, those who have visited a resort, has a positive value (1.29118), whereas group 2 has an equal negative value. The signs of the coefficients associated with all the predictors are positive. This suggests that higher family income, household size, importance attached to family vacation, attitude toward travel, and age are more likely to result in the family visiting the resort. It would be reasonable to develop a profile of the two groups in terms of the three predictors that seem to be the most important: income, household size, and importance of vacation. The values of these three variables for the two groups are given at the beginning of Table 18.4.

The determination of relative importance of the predictors is further illustrated by the following example.

## REAL RESEARCH

### *Satisfied Salespeople Stay*

A recent survey asked businesspeople about the concern of hiring and maintaining employees during the current harsh economic climate. It was reported that 85 percent of respondents were concerned about recruiting employees and 81 percent said they were concerned about retaining employees. When the economy is uncertain, as in 2005–2006, turnover is rapid. Generally speaking, if an organization wants to retain its employees, it must learn why people leave their jobs and why others stay and are satisfied with their jobs. Discriminant analysis was used to determine what factors explained the differences between salespeople who left a large computer manufacturing company and those who stayed. The independent variables were company rating, job security, seven job-satisfaction dimensions, four role-conflict dimensions, four role-ambiguity dimensions, and nine measures of sales performance. The dependent variable was the dichotomy between those who stayed and those who left. The canonical correlation, an index of discrimination ($R = 0.4572$), was significant (Wilks' $\lambda = 0.7909$, $F(26,173) = 1.7588$, $p = 0.0180$). This result indicated that the variables discriminated between those who left and those who stayed.

The results from simultaneously entering all variables in discriminant analysis are presented in the accompanying table. The rank order of importance, as determined by the relative magnitude of the structure correlations, is presented in the first column. Satisfaction with the job and promotional opportunities were the two most important discriminators, followed by job security. Those who stayed in the company found the job to be more exciting, satisfying, challenging, and interesting than those who left.[7]

Discriminant Analysis Results

| *Variable* | *Coefficients* | *Standardized Coefficients* | *Structure Correlations* |
|---|---|---|---|
| 1. Work[a] | 0.0903 | 0.3910 | 0.5446 |
| 2. Promotion[a] | 0.0288 | 0.1515 | 0.5044 |
| 3. Job security | 0.1567 | 0.1384 | 0.4958 |
| 4. Customer relations[b] | 0.0086 | 0.1751 | 0.4906 |
| 5. Company rating | 0.4059 | 0.3240 | 0.4824 |
| 6. Working with others[b] | 0.0018 | 0.0365 | 0.4651 |
| 7. Overall performance[b] | −0.0148 | −0.3252 | 0.4518 |
| 8. Time–territory management[b] | 0.0126 | 0.2899 | 0.4496 |
| 9. Sales produced[b] | 0.0059 | 0.1404 | 0.4484 |
| 10. Presentation skill[b] | 0.0118 | 0.2526 | 0.4387 |
| 11. Technical information[b] | 0.0003 | 0.0065 | 0.4173 |

*(Continued)*

Discriminant Analysis Results *(Continued)*

| *Variable* | *Coefficients* | *Standardized Coefficients* | *Structure Correlations* |
|---|---|---|---|
| 12. Pay–benefits[a] | 0.0600 | 0.1843 | 0.3788 |
| 13. Quota achieved[b] | 0.0035 | 0.2915 | 0.3780 |
| 14. Management[a] | 0.0014 | 0.0138 | 0.3571 |
| 15. Information collection[b] | −0.0146 | −0.3327 | 0.3326 |
| 16. Family[c] | −0.0684 | −0.3408 | −0.3221 |
| 17. Sales manager[a] | −0.0121 | −0.1102 | 0.2909 |
| 18. Coworker[a] | 0.0225 | 0.0893 | 0.2671 |
| 19. Customer[c] | −0.0625 | −0.2797 | −0.2602 |
| 20. Family[d] | 0.0473 | 0.1970 | 0.2180 |
| 21. Job[d] | 0.1378 | 0.5312 | 0.2119 |
| 22. Job[c] | 0.0410 | 0.5475 | −0.1029 |
| 23. Customer[d] | −0.0060 | −0.0255 | 0.1004 |
| 24. Sales manager[c] | −0.0365 | −0.2406 | −0.0499 |
| 25. Sales manager[d] | −0.0606 | −0.3333 | 0.0467 |
| 26. Customer[a] | −0.0338 | −0.1488 | 0.0192 |

*Note:* Rank order of importance is based on the magnitude of the structure correlations.
[a] Satisfaction
[b] Performance
[c] Ambiguity
[d] Conflict ■

Note that in this example, promotion was identified as the second most important variable based on the structure correlations. However, it is not the second most important variable based on the absolute magnitude of the standardized discriminant function coefficients. This anomaly results from multicollinearity.

*characteristic profile*
An aid to interpreting discriminant analysis results by describing each group in terms of the group means for the predictor variables.

Another aid to interpreting discriminant analysis results is to develop a ***characteristic profile*** for each group by describing each group in terms of the group means for the predictor variables. If the important predictors have been identified, then a comparison of the group means on these variables can assist in understanding the intergroup differences. However, before any findings can be interpreted with confidence, it is necessary to validate the results.

## Assess Validity of Discriminant Analysis

Many computer programs, such as SPSS, offer a leave-one-out cross-validation option. In this option, the discriminant model is reestimated as many times as there are respondents in the sample. Each reestimated model leaves out one respondent and the model is used to predict for that respondent. When a large holdout sample is not possible, this gives a sense of the robustness of the estimate using each respondent, in turn, as a holdout.

As explained earlier, where possible, the data should be randomly divided into two subsamples: analysis and validation. The analysis sample is used for estimating the discriminant function; the validation sample is used for developing the classification matrix. The discriminant weights, estimated by using the analysis sample, are multiplied by the values of the predictor variables in the holdout sample to generate discriminant scores for the cases in the holdout sample. The cases are then assigned to groups based on their discriminant scores and an appropriate decision rule. For example, in two-group discriminant analysis, a case will be assigned to the group whose centroid is the closest. The ***hit ratio,*** or the percentage of cases correctly classified, can then be determined by summing the diagonal elements and dividing by the total number of cases.[8]

*hit ratio*
The percentage of cases correctly classified by the discriminant analysis.

It is helpful to compare the percentage of cases correctly classified by discriminant analysis to the percentage that would be obtained by chance. When the groups are equal

in size, the percentage of chance classification is 1 divided by the number of groups. How much improvement should be expected over chance? No general guidelines are available, although some authors have suggested that classification accuracy achieved by discriminant analysis should be at least 25 percent greater than that obtained by chance.[9]

Most discriminant analysis programs also estimate a classification matrix based on the analysis sample. Because they capitalize on chance variation in the data, such results are invariably better than leave-one-out classification or the classification obtained on the holdout sample.

Table 18.4, of the vacation resort example, also shows the classification results based on the analysis sample. The hit ratio, or the percentage of cases correctly classified, is $(12 + 15)/30 = 0.90$, or 90 percent. One might suspect that this hit ratio is artificially inflated, as the data used for estimation was also used for validation. Leave-one-out cross-validation correctly classifies only $(11 + 13)/30 = 0.80$ or 80 percent of the cases. Conducting classification analysis on an independent holdout set of data results in the classification matrix with a hit ratio of $(4 + 6)/12 = 0.833$, or 83.3 percent (see Table 18.4). Given two groups of equal size, by chance one would expect a hit ratio of $1/2 = 0.50$, or 50 percent. Hence, the improvement over chance is more than 25 percent, and the validity of the discriminant analysis is judged as satisfactory.

## REAL RESEARCH

### *Home Bodies and Couch Potatoes*

Two-group discriminant analysis was used to assess the strength of each of five dimensions used in classifying individuals as TV users or nonusers. The procedure was appropriate for this use because of the nature of the predefined categorical groups (users and nonusers) and the interval scales used to generate individual factor scores.

Two equal groups of 185 elderly consumers, users and nonusers (total $n = 370$), were created. The discriminant equation for the analysis was estimated by using a subsample of 142 respondents from the sample of 370. Of the remaining respondents, 198 were used as a validation subsample in a cross-validation of the equation. Thirty respondents were excluded from the analysis because of missing values.

The canonical correlation for the discriminant function was 0.4291, significant at the $p < 0.0001$ level. The eigenvalue was 0.2257. The accompanying table summarizes the standardized canonical discriminant coefficients. A substantial portion of the variance is explained by the discriminant function. In addition, as the table shows, the home orientation dimension made a fairly strong contribution to classifying individuals as users or nonusers of television. Morale, security and health, and respect also contributed significantly. The social factor appeared to make little contribution.

The cross-validation procedure using the discriminant function from the analysis sample gave support to the contention that the dimensions aided researchers in discriminating between users and nonusers of television. As the table shows, the discriminant function was successful in classifying 75.76 percent of the cases. This suggests that consideration of the identified dimensions will help marketers understand the elderly market. Although it is very important for marketers to know and understand the elderly market, the Generation Xers (those born between 1961 and 1981) are also a group that should not be overlooked by marketers. Due to technological advances with the Internet and television, a revolutionary form of interactive TV (ITV) has been created. As of 2006, ITV services were fully deployed and operational and combined Internet and broadcasting with software programs and hardware components to give consumers Internet access, online shopping, music downloads, and an interactive broadcast program, all through their television. With a prosperous-looking forecast for ITV, who better to target this revolutionary form of television than Generation Xers? Discriminant analysis can again be used to determine who among Generation Xers are users or nonusers of ITV and to market ITV services successfully.[10]

**ACTIVE RESEARCH**

Visit *www.timberland.com* and conduct an Internet search using a search engine and your library's online database to obtain information on Timberland's marketing program for outdoor shoes.

As the marketing manager for Timberland, how would your understanding of the consumers' decision-making process affect your decision to sell outdoor shoes on the Internet?

What type of data would you collect and what analysis would you conduct to determine the differentiating characteristics of users and nonusers of rugged outdoor shoes?

Summary of Discriminant Analysis

***Standard Canonical Discriminant Function Coefficients***

| | |
|---|---|
| Morale | 0.27798 |
| Security and health | 0.39850 |
| Home orientation | 0.77496 |
| Respect | 0.32069 |
| Social | −0.01996 |

***Classification Results for Cases Selected for Use in the Analysis***

| | *Number of* | *Predicted Group Membership* | |
|---|---|---|---|
| *Actual Group* | *Cases* | *Nonusers* | *Users* |
| TV nonusers | 77 | 56 | 21 |
| | | 72.7% | 27.3% |
| TV users | 65 | 24 | 41 |
| | | 36.9% | 63.1% |

Percent of grouped cases correctly classified: 68.31%

***Classification Results for Cases Used for Cross-Validation***

| | *Number of* | *Predicted Group Membership* | |
|---|---|---|---|
| *Actual Group* | *Cases* | *Nonusers* | *Users* |
| TV nonusers | 108 | 85 | 23 |
| | | 78.7% | 21.3% |
| TV users | 90 | 25 | 65 |
| | | 27.8% | 72.2% |

Percent of grouped cases correctly classified: 75.76% ■

The extension from two-group discriminant analysis to multiple discriminant analysis involves similar steps.

# MULTIPLE DISCRIMINANT ANALYSIS

## Formulate the Problem

The data presented in Tables 18.2 and 18.3 can also be used to illustrate three-group discriminant analysis. In the last column of these tables, the households are classified into three categories, based on the amount spent on family vacation (high, medium, or low). Ten households fall in each category. The question of interest is whether the households that spend high, medium, or low amounts on their vacations (AMOUNT) can be differentiated in terms of annual family income (INCOME), attitude toward travel (TRAVEL), importance attached to family vacation (VACATION), household size (HSIZE), and age of the head of household (AGE).[11]

## Estimate the Discriminant Function Coefficients

Table 18.5 presents the results of estimating three-group discriminant analysis. An examination of group means indicates that income appears to separate the groups more widely

**TABLE 18.5**

**Results of Three-Group Discriminant Analysis**

GROUP MEANS

| AMOUNT | INCOME | TRAVEL | VACATION | HSIZE | AGE |
|---|---|---|---|---|---|
| 1 | 38.57000 | 4.50000 | 4.70000 | 3.10000 | 50.30000 |
| 2 | 50.11000 | 4.00000 | 4.20000 | 3.40000 | 49.50000 |
| 3 | 64.97000 | 6.10000 | 5.90000 | 4.20000 | 56.00000 |
| Total | 51.21667 | 4.86667 | 4.93333 | 3.56667 | 51.93333 |

GROUP STANDARD DEVIATIONS

| | | | | | |
|---|---|---|---|---|---|
| 1 | 5.29718 | 1.71594 | 1.88856 | 1.19722 | 8.09732 |
| 2 | 6.00231 | 2.35702 | 2.48551 | 1.50555 | 9.25263 |
| 3 | 8.61434 | 1.19722 | 1.66333 | 1.13529 | 7.60117 |
| Total | 12.79523 | 1.97804 | 2.09981 | 1.33089 | 8.57395 |

POOLED WITHIN-GROUPS CORRELATION MATRIX

| | INCOME | TRAVEL | VACATION | HSIZE | AGE |
|---|---|---|---|---|---|
| INCOME | 1.00000 | | | | |
| TRAVEL | 0.05120 | 1.00000 | | | |
| VACATION | 0.30681 | 0.03588 | 1.00000 | | |
| HSIZE | 0.38050 | 0.00474 | 0.22080 | 1.00000 | |
| AGE | −0.20939 | −0.34022 | −0.01326 | −0.02512 | 1.00000 |

Wilks' λ (U-statistic) and univariate *F* ratio with 2 and 27 degrees of freedom.

| VARIABLE | WILKS' LAMBDA | *F* | SIGNIFICANCE |
|---|---|---|---|
| INCOME | 0.26215 | 38.000 | 0.0000 |
| TRAVEL | 0.78790 | 3.634 | 0.0400 |
| VACATION | 0.88060 | 1.830 | 0.1797 |
| HSIZE | 0.87411 | 1.944 | 0.1626 |
| AGE | 0.88214 | 1.804 | 0.1840 |

CANONICAL DISCRIMINANT FUNCTIONS

| FCN | EIGENVALUE | % OF VARIANCE | CUM PCT | CANONICAL CORR | | AFTER FCN | WILKS' λ | CHI-SQUARE | DF | SIG. |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | : | 0 | 0.1664 | 44.831 | 10 | 0.00 |
| 1* | 3.8190 | 93.93 | 93.93 | 0.8902 | : | 1 | 0.8020 | 5.517 | 4 | 0.24 |
| 2* | 0.2469 | 6.07 | 100.00 | 0.4450 | | | | | | |

*Marks the two canonical discriminant functions remaining in the analysis.

STANDARDIZED CANONICAL DISCRIMINANT FUNCTION COEFFICIENTS

| | FUNC 1 | FUNC 2 |
|---|---|---|
| INCOME | 1.04740 | −0.42076 |
| TRAVEL | 0.33991 | 0.76851 |
| VACATION | −0.14198 | 0.53354 |
| HSIZE | −0.16317 | 0.12932 |
| AGE | 0.49474 | 0.52447 |

STRUCTURE MATRIX

Pooled within-groups correlations between discriminating variables and canonical discriminant functions (variables ordered by size of correlation within function).

| | FUNC 1 | FUNC 2 |
|---|---|---|
| INCOME | 0.85556* | −0.27833 |
| HSIZE | 0.19319* | 0.07749 |
| VACATION | 0.21935 | 0.58829* |
| TRAVEL | 0.14899 | 0.45362* |
| AGE | 0.16576 | 0.34079* |

*(Continued)*

**TABLE 18.5**

**Results of Three-Group Discriminant Analysis *(Continued)***

UNSTANDARDIZED CANONICAL DISCRIMINANT FUNCTION COEFFICIENTS

| | FUNC 1 | FUNC 2 |
|---|---|---|
| INCOME | 0.1542658 | −0.6197148E-01 |
| TRAVEL | 0.1867977 | 0.4223430 |
| VACATION | −0.6952264E-01 | 0.2612652 |
| HSIZE | −0.1265334 | 0.1002796 |
| AGE | 0.5928055E-01 | 0.6284206E-01 |
| (constant) | −11.09442 | −3.791600 |

CANONICAL DISCRIMINANT FUNCTIONS EVALUATED AT GROUP MEANS (GROUP CENTROIDS)

| GROUP | FUNC 1 | FUNC 2 |
|---|---|---|
| 1 | −2.04100 | 0.41847 |
| 2 | −0.40479 | −0.65867 |
| 3 | 2.44578 | 0.24020 |

CLASSIFICATION RESULTS

| | | | PREDICTED GROUP MEMBERSHIP | | | |
|---|---|---|---|---|---|---|
| | | AMOUNT | 1 | 2 | 3 | TOTAL |
| Original | Count | 1 | 9 | 1 | 0 | 10 |
| | | 2 | 1 | 9 | 0 | 10 |
| | | 3 | 0 | 2 | 8 | 10 |
| | % | 1 | 90.0 | 10.0 | 0.0 | 100.0 |
| | | 2 | 10.0 | 90.0 | 0.0 | 100.0 |
| | | 3 | 0.0 | 20.0 | 80.0 | 100.0 |
| Cross-validated | Count | 1 | 7 | 3 | 0 | 10 |
| | | 2 | 4 | 5 | 1 | 10 |
| | | 3 | 0 | 2 | 8 | 10 |
| | % | 1 | 70.0 | 30.0 | 0.0 | 100.0 |
| | | 2 | 40.0 | 50.0 | 10.0 | 100.0 |
| | | 3 | 0.0 | 20.0 | 80.0 | 100.0 |

[a] Cross-validation is done only for those cases in the analysis. In cross-validation, each case is classified by the functions derived from all cases other than that case.
[b] 86.7% of original grouped cases correctly classified.
[c] 66.7% of cross-validated grouped cases correctly classified.

CLASSIFICATION RESULTS FOR CASES NOT SELECTED FOR USE IN THE ANALYSIS

| | | | PREDICTED GROUP MEMBERSHIP | | |
|---|---|---|---|---|---|
| | Actual Group | No.of Cases | 1 | 2 | 3 |
| Group | 1 | 4 | 3 | 1 | 0 |
| | | | 75.0% | 25.0% | 0.0% |
| Group | 2 | 4 | 0 | 3 | 1 |
| | | | 0.0% | 75.0% | 25.0% |
| Group | 3 | 4 | 1 | 0 | 3 |
| | | | 25.0% | 0.0% | 75.0% |

Percent of grouped cases correctly classified: 75.0%


**SPSS Output File**

than any other variable. There is some separation on travel and vacation. Groups 1 and 2 are very close in terms of household size and age. Age has a large standard deviation relative to the separation between the groups. The pooled within-groups correlation matrix indicates some correlation of vacation and household size with income. Age has some negative correlation with travel. Yet these correlations are on the lower side, indicating that although multicollinearity may be of some concern, it is not likely to be a serious problem. The significance attached to the univariate $F$ ratios indicates that when the predictors are

considered individually, only income and travel are significant in differentiating between the two groups.

In multiple discriminant analysis, if there are $G$ groups, $G - 1$ discriminant functions can be estimated if the number of predictors is larger than this quantity. In general, with $G$ groups and $k$ predictors, it is possible to estimate up to the smaller of $G - 1$ or $k$ discriminant functions. The first function has the highest ratio of between-groups to within-groups sum of squares. The second function, uncorrelated with the first, has the second highest ratio, and so on. However, not all the functions may be statistically significant.

Because there are three groups, a maximum of two functions can be extracted. The eigenvalue associated with the first function is 3.8190, and this function accounts for 93.93 percent of the explained variance. Because the eigenvalue is large, the first function is likely to be superior. The second function has a small eigenvalue of 0.2469 and accounts for only 6.07 percent of the explained variance.

## Determine the Significance of the Discriminant Function

To test the null hypothesis of equal group centroids, both the functions must be considered simultaneously. It is possible to test the means of the functions successively by first testing all means simultaneously. Then one function is excluded at a time, and the means of the remaining functions are tested at each step. In Table 18.5, the 0 below "After Function" indicates that no functions have been removed. The value of Wilks' $\lambda$ is 0.1644. This transforms to a chi-square of 44.831, with 10 degrees of freedom, which is significant beyond the 0.05 level. Thus, the two functions together significantly discriminate among the three groups. However, when the first function is removed, the Wilks' $\lambda$ associated with the second function is 0.8020, which is not significant at the 0.05 level. Therefore, the second function does not contribute significantly to group differences.

## Interpret the Results

The interpretation of the results is aided by an examination of the standardized discriminant function coefficients, the structure correlations, and certain plots. The standardized coefficients indicate a large coefficient for income on function 1, whereas function 2 has relatively larger coefficients for travel, vacation, and age. A similar conclusion is reached by an examination of the structure matrix (see Table 18.5). To help interpret the functions, variables with large coefficients for a particular function are grouped together. These groupings are shown with asterisks. Thus, income and household size have asterisks for function 1 because these variables have coefficients that are larger for function 1 than for function 2. These variables are associated primarily with function 1. On the other hand, travel, vacation, and age are predominantly associated with function 2, as indicated by the asterisks.

Figure 18.3 is a scattergram plot of all the groups on function 1 and function 2. It can be seen that group 3 has the highest value on function 1, and group 1 the lowest. Because function 1 is primarily associated with income and household size, one would expect the three groups to be ordered on these two variables. Those with higher incomes and higher household size are likely to spend large amounts of money on vacations. Conversely, those with low incomes and smaller household size are likely to spend small amounts on vacations. This interpretation is further strengthened by an examination of group means on income and household size.

Figure 18.3 further indicates that function 2 tends to separate group 1 (highest value) and group 2 (lowest value). This function is primarily associated with travel, vacation, and age. Given the positive correlations of these variables with function 2 in the structure matrix, we expect to find group 1 to be higher than group 2 in terms of travel, vacation, and age. This is indeed true for travel and vacation, as indicated by the

***Figure 18.3***
All-Groups Scattergram


**SPSS Output File**

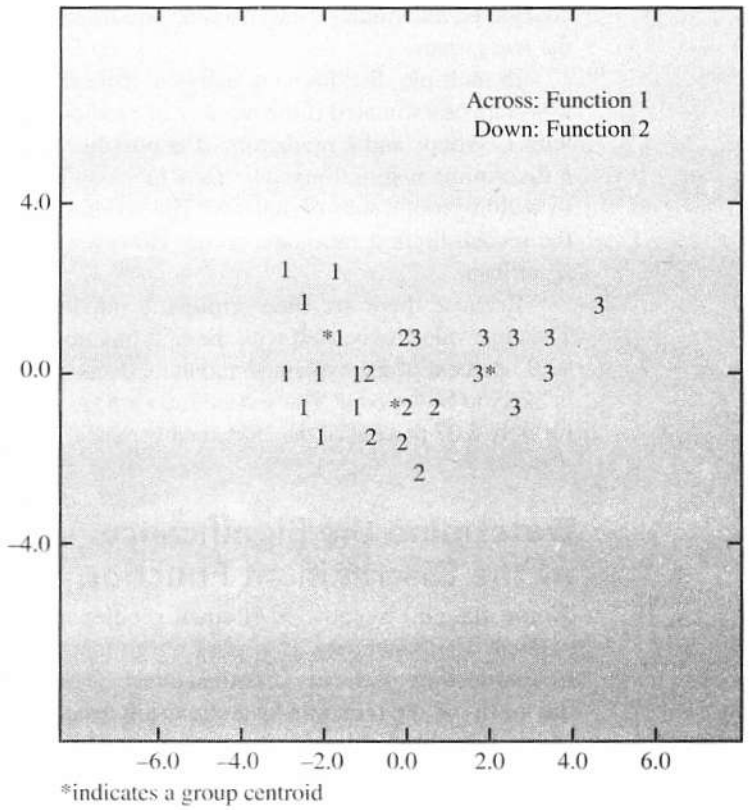
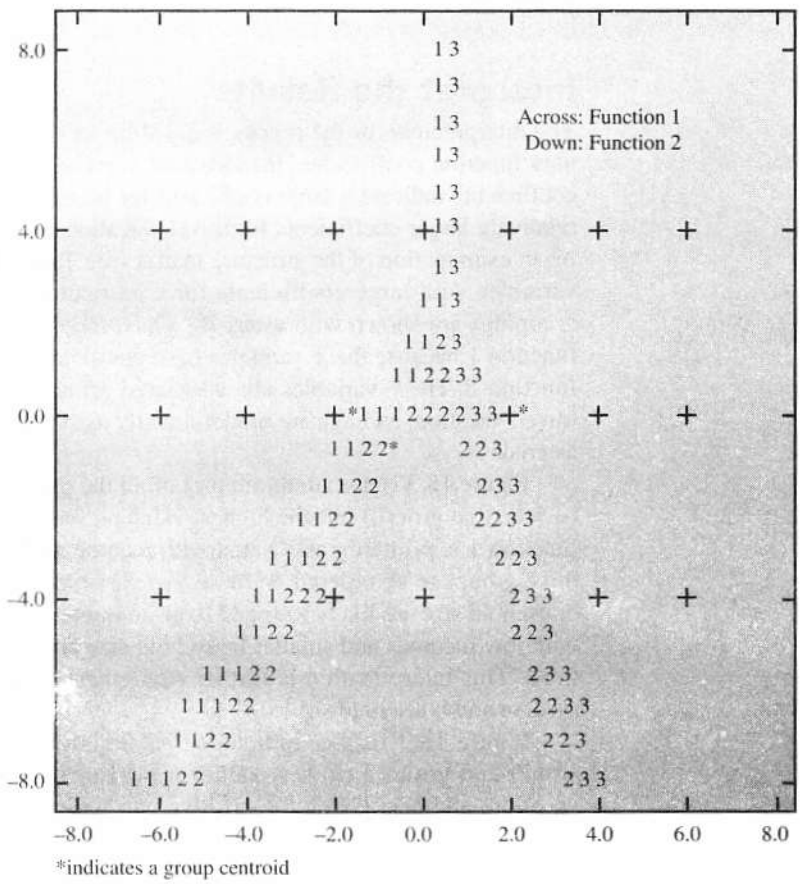



***Figure 18.4***
Territorial Map


**SPSS Output File**



group means of these variables. If families in group 1 have more favorable attitudes toward travel and attach more importance to family vacation than group 2, why do they spend less? Perhaps they would like to spend more on vacations but cannot afford it because they have low incomes.

***territorial map***
A tool for assessing discriminant analysis results that plots the group membership of each case on a graph.

A similar interpretation is obtained by examining a ***territorial map***, as shown in Figure 18.4. In a territorial map, each group centroid is indicated by an asterisk. The group boundaries are shown by numbers corresponding to the groups. Thus, group 1 centroid is bounded by 1s, group 2 centroid by 2s, and group 3 centroid by 3s.

# ASSESS VALIDITY OF DISCRIMINANT ANALYSIS

The classification results based on the analysis sample indicate that $(9 + 9 + 8)/30 = 86.7$ percent of the cases are correctly classified. Leave-one-out cross-validation correctly classifies only $(7 + 5 + 8)/30 = 0.667$ or 66.7 percent of the cases. When the classification analysis is conducted on the independent holdout sample of Table 18.3, a hit ratio of $(3 + 3 + 3)/12 = 75$ percent is obtained. Given three groups of equal size, by chance alone one would expect a hit ratio of $1/3 = 0.333$ or 33.3 percent. Thus, the improvement over chance is greater than 25 percent, indicating at least satisfactory validity.

## REAL RESEARCH

### *The Home Is Where the Patient's Heart Is*

As of 2006, the largest industry sector in the U.S. economy was the health services industry. Through 2010, it is expected that spending on health care services will grow significantly faster than the economy. Contributing to the positive outlook for this industry are the current demographics, especially with demand for long-term care increasing as the population ages. It is expected that the number of Americans who are 85 and older will triple by 2020, and with such a large increase, it is crucial that the health care system be portrayed positively to this segment of the population. Consumers were surveyed to determine their attitudes toward four systems of health care delivery (home health care, hospitals, nursing homes, and outpatient clinics) along 10 attributes. A total of 102 responses were obtained, and the results were analyzed using multiple discriminant analysis (Table 1). Three discriminant functions were identified. Chi-square tests performed on the results indicated that all three discriminant functions were significant at the 0.01 level. The first function accounted for 63 percent of the total discriminative power, and the remaining two functions contributed 29.4 percent and 7.6 percent, respectively.

TABLE 1 Standardized Discriminant Function Coefficients

| | *Discriminant Function* | | |
|---|---|---|---|
| ***Variable*** | ***1*** | ***2*** | ***3*** |
| Safe | −0.20 | −0.04 | 0.15 |
| Convenient | 0.08 | 0.08 | 0.07 |
| Chance of medical complications[a] | −0.27 | 0.10 | 0.16 |
| Expensive[a] | 0.30 | −0.28 | 0.52 |
| Comfortable | 0.53 | 0.27 | −0.19 |
| Sanitary | −0.27 | −0.14 | −0.70 |
| Best medical care | −0.25 | 0.67 | −0.10 |
| Privacy | 0.40 | 0.08 | 0.49 |
| Faster recovery | 0.30 | 0.32 | −0.15 |
| Staffed with best medical personnel | −0.17 | −0.03 | 0.18 |
| Percentage of explained variance | 63.0 | 29.4 | 7.6 |
| Chi-square | 663.3[b] | 289.2[b] | 70.1[b] |

[a] These two items were worded negatively on the questionnaire. They were reverse coded for purposes of data analysis.
[b] $p < 0.01$.

TABLE 2 Centroids of Health Care Systems in Discriminant Space

| | *Discriminant Function* | | |
|---|---|---|---|
| ***System*** | ***1*** | ***2*** | ***3*** |
| Hospital | −1.66 | 0.97 | −0.08 |
| Nursing home | −0.60 | −1.36 | −0.27 |
| Outpatient clinic | 0.54 | −0.13 | 0.77 |
| Home health care | 1.77 | 0.50 | −0.39 |

TABLE 3 Classification Table

| | *Classification (%)* | | | |
|---|---|---|---|---|
| ***System*** | ***Hospital*** | ***Nursing Home*** | ***Outpatient Clinic*** | ***Home Health Care*** |
| Hospital | 86 | 6 | 6 | 2 |
| Nursing home | 9 | 78 | 10 | 3 |
| Outpatient clinic | 9 | 13 | 68 | 10 |
| Home health care | 5 | 4 | 13 | 78 |

Table 1 gives the standardized discriminant function coefficients of the 10 variables in the discriminant equations. Coefficients ranged in value from −1 to +1. In determining the ability of each attribute to classify the delivery system, absolute values were used. In the first discriminant function, the two variables with the largest coefficients were comfort (0.53) and privacy (0.40). Because both related to personal attention and care, the first dimension was labeled "personalized care." In the second function, the two variables with the largest coefficients were quality of medical care (0.67) and likelihood of faster recovery (0.32). Hence, this dimension was labeled "quality of medical care." In the third discriminant function, the most significant attributes were sanitation (−0.70) and expense (0.52). Because these two attributes represent value and price, the third discriminant function was labeled "value."

The four group centroids are shown in Table 2. This table shows that home health care was evaluated most favorably along the dimension of personalized care, and hospitals least favorably. Along the dimension of quality of medical care, there was a substantial separation between nursing homes and the other three systems. Also, home health care received higher evaluations on the quality of medical care than did outpatient clinics. Outpatient clinics, on the other hand, were judged to offer the best value.

Classification analysis of the 102 responses, reported in Table 3, showed correct classifications ranging from 86 percent for hospitals to 68 percent for outpatient clinics. The misclassifications for hospitals were 6 percent each to nursing homes and outpatient clinics, and 2 percent to home health care. Nursing homes showed misclassifications of 9 percent to hospitals, 10 percent to outpatient clinics, and 3 percent to home health care.

**ACTIVE RESEARCH**

Visit *www.tennis.com* and write a report about the current editorial content and features of *Tennis* magazine.

*Tennis* magazine would like to determine what editorial content and feature preferences differentiate its readers, who vary in their tennis activity level, characterized as high, medium, or low. What data should be obtained and what analysis should be conducted to arrive at an answer?

As the editor of *Tennis*, how would you change the editorial content of the magazine if the formulated hypotheses were supported by data collected in a survey of readers?

For outpatient clinics, 9 percent misclassifications were made to hospitals, 13 percent to nursing homes, and 10 percent to home health care. For home health care, the misclassifications were 5 percent to hospitals, 4 percent to nursing homes, and 13 percent to outpatient clinics. The results demonstrated that the discriminant functions were fairly accurate in predicting group membership.[12] ■

## STEPWISE DISCRIMINANT ANALYSIS

Stepwise discriminant analysis is analogous to stepwise multiple regression (see Chapter 17) in that the predictors are entered sequentially based on their ability to discriminate between the groups. An $F$ ratio is calculated for each predictor by conducting a univariate analysis of variance in which the groups are treated as the categorical variable and the predictor as the criterion variable. The predictor with the highest $F$ ratio is the first to be selected for inclusion in the discriminant function, if it meets certain significance and tolerance criteria. A second predictor is added based on the highest adjusted or partial $F$ ratio, taking into account the predictor already selected.

Each predictor selected is tested for retention based on its association with the other predictors selected. The process of selection and retention is continued until all predictors meeting the significance criteria for inclusion and retention have been entered in the discriminant function. Several statistics are computed at each stage. In addition, at the conclusion, a summary of the predictors entered or removed is provided. The standard output associated with the direct method is also available from the stepwise procedure.

***Mahalanobis procedure***
A stepwise procedure used in discriminant analysis to maximize a generalized measure of the distance between the two closest groups.

The selection of the stepwise procedure is based on the optimizing criterion adopted. The ***Mahalanobis procedure*** is based on maximizing a generalized measure of the distance between the two closest groups. This procedure allows marketing researchers to make maximal use of the available information.[13]

The Mahalanobis method was used to conduct a two-group stepwise discriminant analysis on the data pertaining to the visit variable in Tables 18.2 and 18.3. The first predictor variable to be selected was income, followed by household size and then vacation. The order in which the variables were selected also indicates their importance in discriminating between the groups. This was further corroborated by an examination of the standardized discriminant function coefficients and the structure correlation coefficients. Note that the findings of the stepwise analysis agree with the conclusions reported earlier by the direct method.

## THE LOGIT MODEL

When the dependent variable is binary and there are several independent variables that are metric, in addition to two-group discriminant analysis one can also use OLS regression, the logit, and the probit models for estimation. The data preparation for running OLS regression, logit, and probit models is similar in that the dependent variable is coded as 0 or 1. OLS regression was discussed in Chapter 17. The probit model is less commonly used and will not be discussed, but we give an explanation of the logit model.[14]

***binary logit model***
The binary logit model commonly deals with the issue of how likely an observation is to belong to each group. It estimates the probability of an observation belonging to a particular group.

As discussed earlier under the basic concept of discriminant analysis, there are several instances in marketing where we want to explain a binary dependent variable in terms of metric independent variables. (Note that logit analysis can also handle categorical independent variables when these are recoded using dummy variables, as discussed in Chapter 14.) Discriminant analysis deals with the issue of which group an observation is likely to belong to. On the other hand, the ***binary logit model*** commonly deals with the issue of how likely an observation is to belong to each group.

It estimates the probability of an observation belonging to a particular group. Thus, the logit model falls somewhere between regression and discriminant analysis in application. We can estimate the probability of a binary event taking place using the binary logit model, also called *logistic regression*. Consider an event that has two outcomes: success and failure. The probability of success may be modeled using the logit model as:

$$\log_e\left(\frac{P}{1-P}\right) = a_0 + a_1X_1 + a_2X_2 + \cdots + a_kX$$

or

$$\log_e\left(\frac{P}{1-P}\right) = \sum_{i=0}^{k} a_iX_i$$

or

$$P = \frac{\exp\left(\sum_{i=0}^{k} a_iX_i\right)}{1 + \exp\left(\sum_{i=0}^{k} a_iX_i\right)}$$

where

$P$ = probability of success
$X_i$ = independent variable $i$
$a_i$ = parameter to be estimated

It can be seen from the third equation that although $X_i$ may vary from $-\infty$ to $+\infty$, $P$ is constrained to lie between 0 and 1. When $X_i$ approaches $-\infty$, $P$ approaches 0, and when $X_i$ approaches $+\infty$, $P$ approaches 1. This is desirable because $P$ is a probability and must lie between 0 and 1. On the other hand, when OLS regression is used the estimation model is

$$P = \sum_{i=0}^{k} a_iX_i$$

Thus, when OLS regression is used, $P$ is not constrained to lie between 0 and 1; it is possible to obtain estimated values of $P$ that are less than 0 or greater than 1. These values are, of course, conceptually and intuitively unappealing. We demonstrate this phenomenon in our illustrative application.

## Estimating the Binary Logit Model

As discussed in Chapter 17, the linear regression model is fit by the ordinary least squares (OLS) procedure. In OLS regression, the parameters are estimated so as to minimize the sum of squared errors of prediction. The error terms in regression can take on any values and are assumed to follow a normal distribution when conducting statistical tests. In contrast, in the binary logit model, each error can assume only two values. If $Y = 0$, the error is $P$, and if $Y = 1$, the error is $1 - P$. Therefore, we would like to estimate the parameters in such a way that the estimated values of $P$ would be close to 0 when $Y = 0$ and close to 1 when $Y = 1$. The procedure that is used to achieve this and estimate the parameters of the binary logit model is called the *maximum likelihood method*. This method is so called

because it estimates the parameters so as to maximize the likelihood or probability of observing the actual data.

## Model Fit

In multiple regression the model fit is measured by the square of the multiple correlation coefficient, $R^2$, which is also called the *coefficient of multiple determination* (see Chapter 17). In logistic regression (binary logit), commonly used measures of model fit are based on the likelihood function and are Cox & Snell $R$ square and Nagelkerke $R$ square. Both these measures are similar to $R^2$ in multiple regression. The Cox & Snell $R$ square is constrained in such a way that it cannot equal 1.0, even if the model perfectly fits the data. This limitation is overcome by the Nagelkerke $R$ square.

As discussed earlier in this chapter, in discriminant analysis, the model fit is assessed by determining the proportion of correct prediction. A similar procedure can also be used for the binary logit model. If the estimated probability is greater than 0.5, then the predicted value of $Y = 1$. On the other hand, if the estimated probability is less than 0.5, then the predicted value of $Y$ is set to 0. The predicted values of $Y$ can then be compared to the corresponding actual values to determine the percentage of correct predictions.

## Significance Testing

The testing of individual estimated parameters or coefficients for significance is similar to that in multiple regression. In this case, the significance of the estimated coefficients is based on Wald's statistic. This statistic is a test of significance of the logistic regression coefficient based on the asymptotic normality property of maximum likelihood estimates and is estimated as:

$$\text{Wald} = (a_i/SE_{a_i})^2$$

where

$a_i$ = logistical coefficient for that predictor variable
$SE_{a_i}$ = standard error of the logistical coefficient

The Wald statistic is chi-square distributed with 1 degree of freedom if the variable is metric and the number of categories minus 1 if the variable is nonmetric.

The associated significance has the usual interpretation. For practical purposes, the significance of the null hypothesis that $a_{i=0}$ can also be tested using a $t$ test where the degrees of freedom equals the number of observations minus the number of estimated parameters. The ratio of the coefficient to its standard error is compared to the critical $t$ value. For large numbers of observations, the $z$ test can be used.

## Interpretation of the Coefficients

The interpretation of the coefficients or estimated parameters is similar to that in multiple regression, of course taking into account that the nature of the dependent variable is different. In logistic regression, the log odds, that is, $\log_e\left(\frac{P}{1-P}\right)$, is a linear function of the estimated parameters. Thus, if $X_i$ is increased by one unit, the log odds will change by $a_i$ units, when the effect of other independent variables is held constant. Thus $a_i$ is the size of the change in the log odds of the dependent variable event when the corresponding independent variable $X_i$ is increased by one unit and the effect of the other independent variables is held constant. The sign of $a_i$ will determine whether the probability increases (if the sign is positive) or decreases (if the sign is negative) by this amount.

## An Illustrative Application of Logistic Regression

We illustrate the logit model by analyzing the data of Table 18.6. This table gives the data for 30 respondents, 15 of whom are brand loyal (indicated by 1) and 15 of whom are not (indicated by 0). We also measure attitude toward the brand (Brand), attitude toward the product category (Product), and attitude toward shopping (Shopping), all on a 1 (unfavorable) to 7 (favorable) scale. The objective is to estimate the probability of a consumer being brand loyal as a function of attitude toward the brand, the product category, and shopping.

First we run an OLS regression on the data of Table 18.6 to illustrate the limitations of this procedure for analyzing binary data. The estimated equation is given by

$$P = -0.684 + 0.183 \text{ Brand} + 0.020 \text{ Product} + 0.074 \text{ Shopping}$$

where

$P$ = probability of a consumer being brand loyal

Only the constant term and Brand are significant at the 0.05 level. It can be seen from the estimated regression equation that the estimated values of $P$ are negative for low values


**SPSS Data File**

**TABLE 18.6**
**Explaining Brand Loyalty**

| No. | Loyalty | Brand | Product | Shopping |
|---|---|---|---|---|
| 1 | 1 | 4 | 3 | 5 |
| 2 | 1 | 6 | 4 | 4 |
| 3 | 1 | 5 | 2 | 4 |
| 4 | 1 | 7 | 5 | 5 |
| 5 | 1 | 6 | 3 | 4 |
| 6 | 1 | 3 | 4 | 5 |
| 7 | 1 | 5 | 5 | 5 |
| 8 | 1 | 5 | 4 | 2 |
| 9 | 1 | 7 | 5 | 4 |
| 10 | 1 | 7 | 6 | 4 |
| 11 | 1 | 6 | 7 | 2 |
| 12 | 1 | 5 | 6 | 4 |
| 13 | 1 | 7 | 3 | 3 |
| 14 | 1 | 5 | 1 | 4 |
| 15 | 1 | 7 | 5 | 5 |
| 16 | 0 | 3 | 1 | 3 |
| 17 | 0 | 4 | 6 | 2 |
| 18 | 0 | 2 | 5 | 2 |
| 19 | 0 | 5 | 2 | 4 |
| 20 | 0 | 4 | 1 | 3 |
| 21 | 0 | 3 | 3 | 4 |
| 22 | 0 | 3 | 4 | 5 |
| 23 | 0 | 3 | 6 | 3 |
| 24 | 0 | 4 | 4 | 2 |
| 25 | 0 | 6 | 3 | 6 |
| 26 | 0 | 3 | 6 | 3 |
| 27 | 0 | 4 | 3 | 2 |
| 28 | 0 | 3 | 5 | 2 |
| 29 | 0 | 5 | 5 | 3 |
| 30 | 0 | 1 | 3 | 2 |

of the independent variables (e.g., when Brand = 1, Product = 1, and Shopping = 1, and for many other values of Brand = 1, 2, or 3). Likewise, the estimated values of $P$ are greater than 1 for high values of the independent variables (e.g., when Brand = 7, Product = 7, and Shopping = 7). This is intuitively and conceptually unappealing because $P$ is a probability and must lie between 0 and 1.

This limitation of OLS regression is overcome by logistic regression. The output for logistic regression when analyzing the data for Table 18.6 is shown in Table 18.7. The Cox & Snell $R$ square and Nagelkerke $R$ square measures indicate a reasonable fit of the model to the data. This is further verified by the classification table that reveals that 24 of the 30, that is, 80 percent of the cases, are correctly classified. The significance of the estimated coefficients is based on Wald's statistic. We note that only attitude toward the brand is significant in explaining brand loyalty. Unlike discriminant analysis, logistic regression results in standard error estimates for the estimated coefficients and hence their significance can be assessed. The positive sign for the coefficient indicates that positive attitude toward the brand results in higher loyalty toward the brand. Attitude toward the product category and attitude toward shopping do not influence brand loyalty. Thus, a manager seeking to increase brand loyalty should focus on fostering more positive attitude toward the brand and not worry about attitude toward the product category or attitude toward shopping.

The logit model can also be used when the dependent variable has more than two categories. In this case, the model is termed the *multinomial logit*. This procedure is discussed elsewhere by the author.[15]


**SPSS Output File**

**TABLE 18.7**

**Results of Binary Logit Model or Logistic Regression**

DEPENDENT VARIABLE ENCODING

| ORIGINAL VALUE | INTERNAL VALUE |
|---|---|
| Not Loyal | 0 |
| Loyal | 1 |

MODEL SUMMARY

| STEP | −2 LOG LIKELIHOOD | COX & SNELL R SQUARE | NAGELKERKE R SQUARE |
|---|---|---|---|
| 1 | 23.471[a] | .453 | .604 |

[a] Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

CLASSIFICATION TABLE[a]

| | | | PREDICTED | | |
|---|---|---|---|---|---|
| | | | LOYALTY TO THE BRAND | | PERCENTAGE |
| | OBSERVED | | NOT LOYAL | LOYAL | CORRECT |
| Step 1 | Loyalty to the brand | Not Loyal | 12 | 3 | 80.0 |
| | | Loyal | 3 | 12 | 80.0 |
| | Overall Percentage | | | | 80.0 |

[a] The cut Value is .500

VARIABLES IN THE EQUATION

| | | B | S.E. | WALD | DF | SIG. | EXP (B) |
|---|---|---|---|---|---|---|---|
| Step 1[a] | Brand | 1.274 | .479 | 7.075 | 1 | .008 | 3.575 |
| | Product | .186 | .322 | .335 | 1 | .563 | 1.205 |
| | Shopping | .590 | .491 | 1.442 | 1 | .230 | 1.804 |
| | Constant | −8.642 | 3.346 | 6.672 | 1 | .010 | .000 |

[a] Variable(s) entered on step 1: Brand, Product, Shopping.

## DECISION RESEARCH

### *Boston Market: Sizing the Market*

#### The Situation

Michel Andres, president and CEO of Boston Market, is well aware of the fact that according to syndicated data, home meal replacement (HMR) will be the family dining business of this century. HMR is portable, high-quality food that's meant for take-out, and it is the fastest-growing and most significant opportunity in the food industry today. According to ACNielsen's consumer panel data (*acnielsen.com*), 55 percent of respondents purchased a meal for at-home consumption several times a month. Convenience and type of food were the two most influential factors when purchasing HMR. Also, 77 percent of the respondents preferred their meals ready to eat.

Another recent study by the NPD Group (*www.npd.com*) projected that between 2005 and 2010, virtually all growth in food sales will come from food service, defined as food prepared at least partially away from home. Estimates of total HMR market size, as well as future potential, vary widely. Sara Lee's research shows HMR accounting for as much as 80 percent of food industry growth in 2006. Findings by McKinsey & Co. support that premise from two perspectives: first, the fact that virtually all foods sales growth by the year 2006 will come from food service; second, that by 2006 many Americans will never have cooked a meal from scratch. It is the most important trend to hit the food industry since the advent of frozen food.

Boston Market is now the HMR leader. As of 2006, Boston Market had more than 650 restaurants and operated in 30 states. The company is a wholly owned subsidiary of McDonald's Corporation (*www.mcdonalds.com*). Michel Andres wants to capitalize upon this HMR trend, and to do so would like to determine who the heavy users of HMR are and how they differ from the light users and nonusers of HMR.

#### The Marketing Research Decision

1. What data analysis should be conducted to determine a profile of the heavy users of HMR and to identify the differences between the heavy users, the light users, and the nonusers of HMR?
2. Discuss the role of the type of research you recommend in enabling Michel Andres to size the HMR market and determine what new products and services Boston Market should introduce.


Discriminant analysis can help Boston Market identify the differences between the heavy users, light users, and the nonusers of home meal replacement.

### The Marketing Management Decision

1. What new products and services should Michel Andres introduce to target the heavy users of HMR?
2. Discuss how the marketing management decision action that you recommend to Michel Andres is influenced by the syndicated sources of data that you suggested earlier and by the content of information they provide. ■

The next example gives an application of discriminant analysis in international marketing research; the example after that presents an application in ethics.

**REAL RESEARCH**

## *Satisfactory Results of Satisfaction Programs in Europe*

These days, more and more computer companies are emphasizing customer service programs rather than their erstwhile emphasis on computer features and capabilities. Hewlett-Packard (*www.hp.com*) learned this lesson while doing business in Europe. Research conducted on the European market revealed that there was a difference in emphasis on service requirements across age segments. Focus groups revealed that customers above 40 years of age had a hard time with the technical aspects of the computer and greatly required the customer service programs. On the other hand, younger customers appreciated the technical aspects of the product, which added to their satisfaction. Further research in the form of a large single cross-sectional survey was done to uncover the factors leading to differences in the two segments. A two-group discriminant analysis was conducted with satisfied and dissatisfied customers as the two groups and several independent variables such as technical information, ease of operation, variety and scope of customer service programs, and so on. Results confirmed the fact that the variety and scope of customer satisfaction programs was indeed a strong differentiating factor. This was a crucial finding because HP could better handle dissatisfied customers by focusing more on customer services than technical details. Consequently, HP successfully started three programs on customer satisfaction—customer feedback, customer satisfaction surveys, and total quality control. This effort resulted in increased customer satisfaction. After seeing the successful results of these programs in Europe, HP developed a goal to earn and keep customers' satisfaction, trust, and loyalty and to enable them to successfully apply technology to meet their business and personal needs. To achieve this goal, HP established and implemented a total customer experience and quality (TCE&Q) leadership framework in 2005. The details of this framework were documented in HP's 2005 Global Citizenship report.[16] ■

**REAL RESEARCH**

## *Discriminant Analysis Discriminates Ethical and Unethical Firms*

In order to identify the important variables that predict ethical and unethical behavior, discriminant analysis was used. Prior research suggested that the variables that affect ethical decisions are attitudes, leadership, the presence or absence of ethical codes of conduct, and the organization's size.

To determine which of these variables are the best predictors of ethical behavior, 149 firms were surveyed and asked to indicate how their firm operates in 18 different ethical situations. Of these 18 situations, nine related to marketing activities. These activities included using misleading sales presentations, accepting gifts for preferential treatment, pricing below out-of-pocket expenses, and so forth. Based on these nine issues, the respondent firms were classified into two groups: "never practice" and "practice."

An examination of the variables that influenced classification via two-group discriminant analysis indicated that attitudes and a company's size were the best predictors of

ethical behavior. Evidently, smaller firms tend to demonstrate more ethical behavior on marketing issues. One particular company aimed at conducting ethical business practices is the Smile Internet Bank in the United Kingdom (*www.smile.co.uk*). In early 2002, Smile's marketing group launched six cartoon characters that focused on the bank's ethical position. Each cartoon character symbolized one of six bad banking traits, and ultimately positioned Smile as offering the opposite of these traits. In 2006, Smile was marketing a series of ethical mutual funds that invested in ethically sound companies. This marketing strategy has been successful.[17] ■

## STATISTICAL SOFTWARE

In the mainframe version of SPSS, the DISCRIMINANT procedure is used for conducting discriminant analysis. This is a general program that can be used for two-group or multiple discriminant analysis. Furthermore, the direct or the stepwise method can be adopted.

In SAS, the DISCRIM procedure can be used for performing two-group or multiple discriminant analysis. If the assumption of a multivariate normal distribution cannot be met, the NEIGHBOR procedure can be used. In this procedure, a nonparametric nearest neighbor rule is used for classifying the observations. CANDISC performs canonical discriminant analysis and is related to principal component analysis and canonical correlation. The STEPDISC procedure can be used for performing stepwise discriminant analysis. The mainframe and microcomputer versions are similar, except that the program NEIGHBOR is not available on the microcomputer version.

In MINITAB, discriminant analysis can be conducted using the Stats>Multivariate> Discriminate Analysis function. It computes both linear and quadratic discriminant analysis in the classification of observations into two or more groups. Discriminant analysis is not available in Excel.

## SPSS WINDOWS

The DISCRIMINANT program performs both two-group and multiple discriminant analysis. To select this procedure using SPSS for Windows, click:

Analyze>Classify>Discriminant . . .

The following are the detailed steps for running a two-group discriminant analysis with Resort Visit (visit) as the dependent variable and annual family income (income), attitude toward travel (attitude), importance attached to family vacation (vacation), household size (hsize), and age of the household (age) as the independent variables, using the data of Table 18.2. The corresponding screen captures for these steps can be downloaded from the Web site for this book.

1. Select ANALYZE from the SPSS menu bar.
2. Click CLASSIFY and then DISCRIMINANT.
3. Move "visit" into the GROUPING VARIABLE box.
4. Click DEFINE RANGE. Enter 1 for MINIMUM and 2 for MAXIMUM. Click CONTINUE.
5. Move "income," "travel," "vacation," "hsize," and "age" into the INDEPENDENTS box.
6. Select ENTER INDEPENDENTS TOGETHER (default option).
7. Click on STATISTICS. In the pop-up window, in the DESCRIPTIVES box check MEANS and UNIVARIATE ANOVAS. In the MATRICES box check WITHIN-GROUP CORRELATIONS. Click CONTINUE.

8. Click CLASSIFY. . . . In the pop-up window in the PRIOR PROBABILITIES box check ALL GROUPS EQUAL (default). In the DISPLAY box check SUMMARY TABLE and LEAVE-ONE-OUT CLASSIFICATION. In the USE COVARIANCE MATRIX box check WITHIN-GROUPS. Click CONTINUE.
9. Click OK.

The steps for running three-group discriminant analysis are similar. Select the appropriate dependent and independent variables. In step 4 above, click DEFINE RANGE. Enter 1 for MINIMUM and 3 for MAXIMUM. Click CONTINUE. For running stepwise discriminant analysis, in step 6 above select USE STEPWISE METHOD.

To run logit analysis or logistic regression using SPSS for Windows, click:

Analyze > Regression>Binary Logistic . . .

The following are the detailed steps for running logit analysis with brand loyalty as the dependent variable and attitude toward the brand (brand), attitude toward the product category (product), and attitude toward shopping (shopping) as the independent variables using the data of Table 18.6. The corresponding screen captures for these steps can be downloaded from the Web site for this book.

1. Select ANALYZE from the SPSS menu bar.
2. Click REGRESSION and then BINARY LOGISTIC.
3. Move "Loyalty to the Brand [Loyalty]" into the DEPENDENT VARIABLE box.
4. Move "Attitude toward the Brand [Brand]," "Attitude toward the Product category [Product]," and "Attitude toward Shopping [Shopping]," into the COVARIATES (S box).
5. Select ENTER for METHOD (default option).
6. Click OK.

## PROJECT RESEARCH

### *Two-Group Discriminant Analysis*

In the department store project, two-group discriminant analysis was used to examine whether those respondents who were familiar with the stores, versus those who were unfamiliar, attached different relative importance to the eight factors of the choice criteria. The dependent variable was the two familiarity groups, and the independent variables were the importance attached to the eight factors of the choice criteria. The overall discriminant function was significant, indicating significant differences between the two groups. The results indicated that, as compared to the unfamiliar respondents, the familiar respondents attached greater relative importance to quality of merchandise, return and adjustment policy, service of store personnel, and credit and billing policies.


**SPSS Data File**

#### Project Activities

Download the SPSS data file Sears Data 17 from the Web site for this book. See Chapter 17 for a description of this file.

1. Recode preference for Sears into two groups: 1 to 4 = 1; 5 to 6 = 2. Can these two groups be explained in terms of the evaluations of Sears on the eight factors of the choice criteria? Compare these results to the regression results in Chapter 17.
2. Recode preference for Sears into three groups: 1 to 3 = 1; 4 = 2; 5 to 6 = 3. Can these three groups be explained in terms of the evaluations of Sears on the eight factors of the choice criteria? Compare these results to the regression results in Chapter 17. ■


**SPSS Data File**

## EXPERIENTIAL RESEARCH

Download the Dell case and questionnaire from the Web site for this book. This information is also given at the end of the book. Download the Dell SPSS data file.

1. Do a two-group discriminant analysis with the two overall satisfaction groups derived based on the recoding of q4 (as specified in Chapter 14) as the dependent variables and all the 13 evaluations of Dell (q8_1 to q8_13) as the independent variables. Interpret the results.
2. Do a two-group discriminant analysis with the two likelihood of choosing Dell groups derived based on the recoding of q6 (as specified in Chapter 14) as the dependent variables and all the 13 evaluations of Dell (q8_1 to q8_13) as the independent variables. Interpret the results.
3. Do a three-group discriminant analysis with the three price sensitivity groups derived based on the recoding of q9_5per (as specified in Chapter 14) as the dependent variables and all the 13 evaluations of Dell (q8_1 to q8_13) as the independent variables. Interpret the results.
4. Do a three-group discriminant analysis with the three price sensitivity groups derived based on the recoding of q9_10per (as specified in Chapter 14) as the dependent variables and all the 13 evaluations of Dell (q8_1 to q8_13) as the independent variables. Interpret the results. ■

# SUMMARY

Discriminant analysis is useful for analyzing data when the criterion or dependent variable is categorical and the predictor or independent variables are interval scaled. When the criterion variable has two categories, the technique is known as two-group discriminant analysis. Multiple discriminant analysis refers to the case when three or more categories are involved.

Conducting discriminant analysis is a five-step procedure. First, formulating the discriminant problem requires identification of the objectives and the criterion and predictor variables. The sample is divided into two parts. One part, the analysis sample, is used to estimate the discriminant function. The other part, the holdout sample, is reserved for validation. Estimation, the second step, involves developing a linear combination of the predictors, called discriminant functions, so that the groups differ as much as possible on the predictor values.

Determination of statistical significance is the third step. It involves testing the null hypothesis that, in the population, the means of all discriminant functions in all groups are equal. If the null hypothesis is rejected, it is meaningful to interpret the results.

The fourth step, the interpretation of discriminant weights or coefficients, is similar to that in multiple regression analysis. Given the multicollinearity in the predictor variables, there is no unambiguous measure of the relative importance of the predictors in discriminating between the groups. However, some idea of the relative importance of the variables may be obtained by examining the absolute magnitude of the standardized discriminant function coefficients and by examining the structure correlations or discriminant loadings. These simple correlations between each predictor and the discriminant function represent the variance that the predictor shares with the function. Another aid to interpreting discriminant analysis results is to develop a characteristic profile for each group, based on the group means for the predictor variables.

Validation, the fifth step, involves developing the classification matrix. The discriminant weights estimated by using the analysis sample are multiplied by the values of the predictor variables in the holdout sample to generate discriminant scores for the cases in the holdout sample. The cases are then assigned to groups based on their discriminant scores and an appropriate decision rule. The percentage of cases correctly classified is determined and compared to the rate that would be expected by chance classification.

Two broad approaches are available for estimating the coefficients. The direct method involves estimating the discriminant function so that all the predictors are included simultaneously. An alternative is the stepwise method, in which the predictor variables are entered sequentially, based on their ability to discriminate among groups.

In multiple discriminant analysis, if there are $G$ groups and $k$ predictors, it is possible to estimate up to the smaller of $G - 1$ or $k$ discriminant functions. The first function has the highest ratio of between-group to within-group sums of squares. The second function, uncorrelated with the first, has the second highest ratio, and so on.

Logit analysis, also called logistic regression, is an alternative to two-group discriminant analysis when the dependent variable is binary. The logit model estimates the probability of a binary event. Unlike OLS regression, the logit model constraints the probability to lie between 0 and 1. Unlike discriminant analysis, logistic regression results in standard error estimates for the estimated coefficients and hence their significance can be assessed.

## KEY TERMS AND CONCEPTS

discriminant analysis, *578*
discriminant functions, *578*
two-group discriminant analysis, *578*
multiple discriminant analysis, *578*
discriminant analysis model, *579*
canonical correlation, *580*
centroid, *580*
classification matrix, *580*
discriminant function coefficients, *581*
discriminant scores, *581*
eigenvalue, *581*
*F* values and their significance, *581*
group means and group standard deviations, *581*
pooled within-group correlation matrix, *581*
standardized discriminant function coefficients, *581*
structure correlations, *581*
total correlation matrix, *581*
Wilks' λ, *581*
analysis sample, *582*
validation sample, *582*
direct method, *582*
stepwise discriminant analysis, *582*
characteristic profile, *588*
hit ratio, *588*
territorial map, *595*
Mahalanobis procedure, *597*
binary logit model, *597*

## SUGGESTED CASES, VIDEO CASES, AND HBS CASES

### *Cases*

Case 3.2 The Demographic Discovery of the New Millennium
Case 3.3 Matsushita Retargets the U.S.A.
Case 3.4 Pampers Curing Its Rash of Market Share
Case 3.6 Cingular Wireless: A Singular Focus
Case 3.7 IBM: The World's Top Provider of Computer Hardware, Software, and Services
Case 3.8 Kimberly-Clark: Competing Through Innovation
Case 4.1 Wachovia: "Watch Ovah Ya" Finances
Case 4.2 Wendy's: History and Life After Dave Thomas
Case 4.3 Astec: Continuing to Grow
Case 4.4 Is Marketing Research the Cure for Norton Healthcare Kosair Children's Hospital's Ailments?

### *Video Cases*

Video Case 3.1 The Mayo Clinic: Staying Healthy with Marketing Research
Video Case 4.1 Subaru: "Mr. Surv~y" Monitors Customer Satisfaction
Video Case 4.2 Procter & Gamble: Using Marketing Research to Build Brands

## LIVE RESEARCH: CONDUCTING A MARKETING RESEARCH PROJECT

1. Differences between groups (e.g., loyalty groups, usage groups, lifestyle groups, etc.) are of interest in most projects. These differences in terms of multiple variables can be examined using discriminant analysis.
2. If market segmentation has been conducted, then differences between segments can be examined using discriminant analysis.

# ACRONYMS

The steps involved and some key concepts in discriminant analysis may be summarized by the acronym DISCRIMINANT:

**D** ependent variable: categorical
**I** ndpendent variable: metric
**S** tructure correlations or discriminant loadings
**C** alculation of the discriminant function
**R** elative importance of predictors: ambiguous
**I** nterpretation: scattergram and territorial map
**M** eans and standard deviations for groups
**I** nference: determination of significance
**N** umber of functions possible : Minimum ($G$-1, $k$)
**A** ssociation: canonical correlation
**N** umber 1 function has highest eigenvalue
**T** esting for validity: classification analysis

# EXERCISES

## *Questions*

1. What are the objectives of discriminant analysis?
2. What is the main distinction between two-group and multiple discriminant analysis?
3. Describe the relationship of discriminant analysis to regression and ANOVA.
4. What are the steps involved in conducting discriminant analysis?
5. How should the total sample be split for estimation and validation purposes?
6. What is Wilks' $\lambda$? For what purpose is it used?
7. Define discriminant scores.
8. Explain what is meant by an eigenvalue.
9. What is a classification matrix?
10. Explain the concept of structure correlations.
11. How is the statistical significance of discriminant analysis determined?
12. Describe a common procedure for determining the validity of discriminant analysis.
13. When the groups are of equal size, how is the accuracy of chance classification determined?
14. How does the stepwise discriminant procedure differ from the direct method?

## *Problems*

1. In investigating the differences between heavy and light or nonusers of frozen foods, it was found that the two largest standardized discriminant function coefficients were 0.97 for convenience orientation and 0.61 for income. Is it correct to conclude that convenience orientation is more important than income when each variable is considered by itself?
2. Given the following information, calculate the discriminant score for each respondent. The value of the constant is 2.04.

| ***Unstandardized Discriminant Function Coefficients*** | |
|---|---|
| Age | 0.38 |
| Income | 0.44 |
| Risk taking | −0.39 |
| Optimistic | 1.26 |

| ***Respondent ID*** | ***Age*** | ***Income*** | ***Risk Taking*** | ***Optimistic*** |
|---|---|---|---|---|
| 0246 | 36 | 43.7 | 21 | 65 |
| 1337 | 44 | 62.5 | 28 | 56 |
| 2375 | 57 | 33.5 | 25 | 40 |
| 2454 | 63 | 38.7 | 16 | 36 |

# INTERNET AND COMPUTER EXERCISES

1. Conduct a two-group discriminant analysis on the data given in Tables 18.2 and 18.3 using the SPSS, SAS, and MINITAB packages. Compare the output from all the packages. Discuss the similarities and differences.
2. Conduct a three-group stepwise discriminant analysis on the data given in Tables 18.2 and 18.3 using the SPSS, SAS, or MINITAB package. Compare the results to those given in Table 18.5 for three-group discriminant analysis.
3. Analyze the Nike data given in Internet and Computer Exercises 1 of Chapter 15. Do the three usage groups differ in terms of awareness, attitude, preference, intention, and loyalty toward Nike when these variables are considered simultaneously?
4. Analyze the outdoor lifestyle data given in Internet and Computer Exercises 2 of Chapter 15. Do the three groups based on location of residence differ on the importance attached to enjoying nature, relating to the weather, living in harmony with the environment, exercising regularly, and meeting other people (V2 to V6) when these variables are considered simultaneously?
5. Conduct a two-group discriminant analysis on the data you obtained in Fieldwork exercise 1, using the SPSS, SAS, or MINITAB package. Is it possible to differentiate between graduate and undergraduate students using the four attitudinal measures?

# ACTIVITIES

## *Fieldwork*

1. Interview 15 graduate and 15 undergraduate students. Measure their attitudes toward college education (It is worthwhile getting a college degree), enjoyment in life (It is important to have fun in life), your university (I am not very happy that I chose to go to school here), and work ethic (In general, there is a lack of work ethic on the college campus). For each attitude, measure the degree of disagreement/agreement using a 7-point rating scale (1 = disagree, 7 = agree).

## *Group Discussion*

1. Is it meaningful to determine the relative importance of predictors in discriminating between the groups? Why or why not? Discuss as a small group.