

11

BASIC DATA ANALYSIS

SCOTT M. SMITH

Brigham Young University and Qualtrics.com, SurveyZ.com, and SurveyPro.com

GERALD S. ALBAUM

University of New Mexico

From a managerial perspective, data can be viewed as recorded information useful for making decisions. Completed questionnaires or other measurement instruments must be edited, coded, entered into a data set for processing by computer, and carefully analyzed before their complete meanings and implications can be understood.

Analysis can be viewed as the categorization, the aggregation into constituent parts, and the manipulation of data to obtain answers to the research question or questions underlying the research project. A special aspect of analysis is interpretation. The process of interpretation involves taking the results of analysis, making inferences relevant to the research relationships studied, and drawing managerially useful conclusions about these relationships. The analysis of obtained data represents the end of the research process except for the preparation of the report, and everything done prior to this stage has been done so that the proper analysis

can be completed and the research questions can be answered.

AN OVERVIEW OF THE ANALYSIS PROCESS

The overall process of analyzing and making inferences from sample data can be viewed as a process of refinement that involves a number of separate and sequential steps that may be identified as part of three broad stages:

Tabulation: identifying appropriate categories for the information desired, sorting the data into them, making the initial counts of responses, and using summarizing measures to provide economy of description and so facilitate understanding.

Formulating additional hypotheses: using the inductions derived from the data concerning the relevant variables, their parameters, their differences, and their relationships to suggest working hypotheses not originally considered.

AUTHOR'S NOTE: This chapter is reprinted from Smith and Albaum (2005) with permission from Sage Publications Inc.

196 • ANALYSIS AND MODELING

Making inferences: reaching conclusions about the variables that are important, their parameters, their differences, and the relationships among them. In the tabulation process, appropriate categories are defined for coding the information desired, making the initial counts of responses, and preparing a descriptive summary of the data.

General Comments on Data Tabulation

Seven steps are involved in the process of data tabulation:

Categorize. Define appropriate categories for coding the information collected.

Code. Assign codes to the respondent's answers.

Create data file. Enter the data into the computer and create a data file.

Error checking. Check the data file for errors by performing a simple tabulation analysis to identify errors in coding or data entry. Once errors are identified, data may be recoded to collapse categories or combine or delete responses.

Generate new variables. New variables may be computed by data manipulations that multiply, sum, or otherwise transform variables.

Weight data subclasses. Weights are often used to adjust the representation of sample subgroups so that they match the proportions found in the population.

Tabulate. Summarize the responses to each variable included in the analysis.

Our discussion here will be comparatively brief. We consider only the core aspects of the task:

1. Defining of response categories
2. Editing and coding
3. Handling missing data
4. Identifying outliers
5. Creating indices
6. Tabulation

As simple as these steps are from a technical standpoint, they are most important in assuming

a quality analysis and thereby merit introductory discussion. A discussion of survey-based data management, which is another name for this process, is provided by Fink (2003, chap. 1).

Defining Categories

Analysis of any sizable array of data often requires that responses be grouped into categories or classes. The identification of response categories early in the study has several advantages. Ideally, it forces the analyst to consider all possible interpretations and responses to the questionnaire. It often leads to improvements in the questionnaire or observation forms. It permits more detailed instruction of interviewers and results in higher consistency in interpreting responses. Editing problems are also reduced.

The definition of categories allows for identification of the database columns assigned to each question or variable and to indicate the values assigned to each response alternative. Depending on the data collection method, data code sheets can be prepared and precoded. Data files are often formatted as comma-separated (CSV) free-format files, meaning that each variable appears in the same relative position for each respondent with some delimiter (comma, tab, space) between variables. The major data analysis software programs read data files and then display them in a spreadsheet-like database (see Table 11.1). Often, the data are entered directly into the spreadsheet-format data file of the statistical program to be used for analysis. Where data are not collected and formatted electronically, precoding of printed questionnaires will eliminate transcription, thereby decreasing both processing errors and costs. Most of today's computer-based software for telephone (computer-assisted telephone interviewing [CATI]) or online (Surveyz.com, Qualtrics.com) surveys automate this entire process. They not only define the question categories in the database but also automatically build the database and record the completed responses as they are submitted. The data may then be analyzed online, exported to Microsoft Excel™, or imported into a dedicated statistical analysis program such as SPSS.

As desirable as the early definition of categories is, it can sometimes only be done after the

data have been collected. This is usually the case when open-ended text questions, unstructured interviews, and projective techniques are used.

The selection of categories is controlled by both the purposes of the study and the nature of the responses. Useful classifications meet the following conditions:

Similarity of response within the category. Each category should contain responses that, for purposes of the study, are sufficiently similar that they can be considered homogenous.

Differences of responses between categories. Differences in category descriptions should be great enough to disclose any important distinctions in the characteristic being examined.

Mutually exclusive categories. There should be an unambiguous description of categories, defined so that any response can be placed in only one category.

Categories should be exhaustive. The classification schema should provide categories for all responses.

The use of extensive open-ended questions often provides rich contextual and anecdotal information but is a practice often associated with fledgling researchers. Open-ended questions, of course, have their place in marketing research. However, the researcher should be aware of the inherent difficulties in questionnaire coding and tabulation, not to mention their tendency to be more burdensome to the respondent. All of this is by way of saying that any open-ended question should be carefully checked to see if a closed-ended question (i.e., check the appropriate box) can be substituted without doing violence to the intent of the question. Obviously, sometimes this substitution should not be made.

Editing and Coding

Editing

Editing is the process of reviewing the data to ensure maximum accuracy and clarity. Editing should be conducted as the data are being collected. This applies to the editing of the collection forms used for pretesting as well as those

for the full-scale project. Careful editing early in the collection process will often catch misunderstandings of instructions, errors in recording, and other problems at a stage when it is still possible to eliminate them for the later stages of the study. Early editing has the additional advantage of permitting the questioning of interviewers while the material is still relatively fresh in their minds. Obviously, this has limited application for printed questionnaires, though online or CATI surveys can be edited even when data are being collected.

Editing is normally centralized so as to ensure consistency and uniformity in the treatment of the data. If the sample is not large, a single editor usually edits all the data to reduce variation in interpretation. In those cases where the size of the project makes the use of more than one editor mandatory, it is usually best to assign each editor a different portion of the data collection form to edit. In this way, the same editor edits the same items on all forms, an arrangement that tends to improve both consistency and productivity.

Typically, interviewer and respondent data are monitored to ensure that data requirements are fulfilled. Each collection form should be edited to ensure that data quality requirements are fulfilled. Regarding data obtained by an interviewer (and, to an extent, self-report data), the following should be specifically evaluated:

1. Legibility of entries. Obviously the data must be legible in order to be used. Where not legible, although it may be possible to infer the response from other data collected, where any real doubt exists about the meaning of data, it should not be used.

2. Completeness of entries. On a fully structured collection form, the absence of an entry is ambiguous. It may mean that the respondent could not or would not provide the answer, that the interviewer failed to ask the question, or that there was a failure to record collected data.

3. Consistency of entries. Inconsistencies raise the question of which response is correct. (If a respondent family is indicated as being a nonwatcher of game shows, for example, and a later entry indicates that they watched *Wheel of Fortune* twice during the past week, an

198 • ANALYSIS AND MODELING

obvious question arises as to which is correct.) Discrepancies may be cleared up by questioning the interviewer or making callbacks to the respondent. When discrepancies cannot be resolved, discarding both entries is usually the wisest course of action.

4. Accuracy of entries. An editor should keep an eye out for any indication of inaccuracy in the data. Of particular importance is the detection of any repetitive response patterns in the reports of individual interviews. Such patterns may well be indicative of systematic interviewer bias or interviewer/respondent dishonesty.

Coding

Coding is the process of assigning responses to data categories, and numbers are assigned to identify them with the categories. Precoding refers to the practice of assigning codes to categories and sometimes printing this information on structured questionnaires and observation forms before the data are collected. The interviewer is able to code the responses when interpreting the response and marking the category into which it should be placed.

Postcoding refers to the assignment of codes to responses after the data are collected. Postcoding is most often required when responses are reported in an unstructured format (open-ended text or numeric input). Careful interpretation and good judgment are required to ensure that the meaning of the response and the meaning of the category are consistently and uniformly matched.

Once a complete code has been established, after postcoding, a formal coding manual or codebook is often created and made available to those who will be entering or analyzing the data. The codebook used for a study of supermarkets in the United States is shown in Figure 11.1 as an illustration.

Like good questionnaire construction, good coding requires training and supervision. The editor-coder should be provided with written instructions, including examples. He or she should be exposed to the interviewing of respondents and become acquainted with the process and problems of collecting the data, thus providing aid in its interpretation. The coder

also should be aware of the computer routines that are expected to be applied, insofar as they may require certain kinds of data formats.

Whenever possible (and when cost allows), more than one person should do the coding, specifically the postcoding. By comparing the results of the various coders, a process known as determining intercoder reliability, any inconsistencies can be brought out. In addition to the obvious objective of eliminating data coding inconsistencies, the need for recoding sometimes points to the need for additional categories for data classification and may sometimes mean that there is a need to combine some of the categories. Coding is an activity that should not be taken lightly. Improper coding leads to poor analyses and may even constrain the types of analysis that can be completed.

Tabulation for Purposes of Cleaning the Data

The raw input to most data analyses consists of the basic data matrix, as shown in Table 11.1. In most data matrices, each row contains the data for a respondent's records, and the columns identify the variables or data fields collected for the respondents. This rectangular array of entries contains information that is to be summarized and portrayed in some way. For example, the analyses of a column of data might include a tabulation of data counts in each of the categories or the computation of the mean and standard deviation. This summary analysis is often done simply because we are unable to comprehend the meaning of the entire column of values. In so doing, we often (willingly) forgo the full information provided by the data to understand some of its basic characteristics, such as central tendency, dispersion, or categories of responses. Because we summarize the data and make inferences from them, it is doubly important that the data be accurate.

The purpose of the initial data cleaning tabulation is to identify outliers, missing data and other indications of data, coding, transcription, or entry errors. The tabulation of responses will invariably reveal codes that are out of range or otherwise invalid. For example, one tabulation might reveal 46 males (Category 1), 54 females (Category 2), and one Category 5 response,

Question		
Number	Variable	
	01	Sample group 1 = Group A 2 = Group B
	02	Respondent ID number xxx = actual number
1	03	Residential district 1 = SE 2 = SW 3 = NW 4 = NE
2	04	How often shop at Albertson's xx = actual number
2	05	How often shop at Raley's xx = actual number
2	06	How often shop at Wal-Mart xx = actual number
2	07	How often shop at Smith's xx = actual number
3	08	Primary shopper 1 = self 2 = spouse 3 = parent(s) 4 = housekeeper 5 = other
4A	09	Store most likely to shop at 1 = Albertson's 2 = Raley's 3 = Wal-Mart 4 = Smith's 5 = Other
4B	10	Time to get to store xxx = actual number
4C	11	How to get there 1 = car/taxi 2 = bus 3 = walk
5	12	Amount spent at Albertson's xxxxx = actual number
5	13	Amount spent at Raley's xxxxx = actual amount
5	14	Amount spent at Wal-Mart xxxxx = actual number
5	15	Amount spent at Smith's xxxxx = actual number
	16	BLANK
6	17	Supermarket evaluated 1 = Albertson's 2 = Smith's
6	18–35	Semantic scales for Albertson's x = 1 to 7, starting from the left side of scale location (18) layout (27) prices (19) shopping experience (28) atmosphere (20) reputation (29) quality of products (21) service (30) modern (22) helpfulness of clerks (31) friendliness of clerks (23) dull (32) customers (24) selection of products (33) cluttered (25) dirty (34) check-out (26) like (35)
6	36–53	Semantic scales for Smith's X = 1 to 7, starting from the left side of the scale location (36) layout (45) prices (37) shopping experience (46) atmosphere (38) reputation (47) quality of products (39) service (48) modern (40) helpfulness of clerks (49) friendliness of clerks (41) dull (50) customers (42) selection of products (51) cluttered (43) dirty (52) check-out (44) like (53)
7	54	Gender 1 = female 2 = male
8	55	Marital status 1 = single 2 = married 3 = divorced/separated/widowed 4 = other
9	56	Age xx = actual number
10	57	Employment status 1 = full-time 2 = part-time 3 = not employed

Figure 11.1 Codebook for Comparative Supermarket Study

200 • ANALYSIS AND MODELING

Table 11.1 Illustration of Data Matrix

<i>Object</i>	<i>Variable</i>				
	<i>1</i>	<i>2</i>	<i>3 ...</i>	<i>j ...</i>	<i>m</i>
1	X_{11}	X_{12}	$X_{13} \dots$	$X_{1j} \dots$	X_{1m}
2	X_{21}	X_{22}	$X_{23} \dots$	$X_{2j} \dots$	X_{2m}
3	X_{31}	X_{32}	$X_{33} \dots$	$X_{3j} \dots$	X_{3m}
<i>i</i>	X_{i1}	X_{i2}	$X_{i3} \dots$	$X_{ij} \dots$	X_{im}
<i>n</i>	X_{n1}	X_{n2}	$X_{n3} \dots$	$X_{nj} \dots$	X_{nm}

which is obviously an error. Some errors, such as the preceding one, represent entry of values that are out of range or wild codes (Lewis-Beck, 1995, p. 7). That is, the value is not one that has been assigned to a possible response to a question. A miscoding error that is more difficult to detect is one where this is an erroneous recording of a response category using a number that is assigned to a response. That is, in the coding shown in Figure 11.1, a response of self to question number 3 (code = 1) might have been coded as spouse (code = 2). It is hoped that not too many errors of this type occur.

An aspect of cleaning the data is dealing with missing data. That is, some respondents may not provide responses for all the questions. One way of handling this is to use statistical imputation. This involves estimating how respondents who did not answer particular questions would have answered if they had chosen to. Researchers are mixed in their views about this process. A much safer way to handle a nonresponse situation is to treat the nonresponse as missing data in the analysis. Statistical packages such as SPSS can handle this either question by question or by deleting the respondent with a missing value from all analyses. Also, the researcher can choose to eliminate a respondent from the data set if there is too much missing data. Yet another way is to simply assign the group's mean value to the missing items. Or, when a missing item is for an item that is part of a multi-item measure, a respondent's mean value for the rest of the items can be used for the missing value. Chapter 10 of this book discusses in detail missing data.

Another issue that can arise is how to deal with outliers (Fink, 2003, pp. 22–23). Outliers

are respondents whose answers appear to be inconsistent with the rest of the data set. An easy way to check for outliers is by running frequency analyses, or counts, of responses to questions. Regression analysis also can be used to detect outliers. This is discussed briefly later in this chapter. Outliers can be discarded from the analysis, but one must be careful to not throw out important and useful information. If an outlier is retained, then it may be best to use the median rather than the mean as the measure of central tendency when such a measure is part of the analysis.

Short of having two or more coders create the data file independently of each other and then assessing intercoder reliability, there is not much that can be done to prevent coder error except to impress upon coders the necessity of accurate data entry. Multiple coders can be very time-consuming and costly, particularly for large data files. Each error that is identified should be traced back to the questionnaire to determine the proper code. The cleaning process is complete when either the data file has been edited to correct the errors or the corrections have been made in the analysis program.

Basic Tabulation Analysis

Tabulation may be thought of as the final step in the data collection process and the first step in the analytical process. Tabulation is simply the counting of the number of responses in each data category (often a single column of the data matrix contains the responses to all categories).

The most basic is the simple tabulation, often called the marginal tabulation and familiar to all

Table 11.2 Cross-Tabulation: Olive Oil Purchased in Past 3 Months by Income Classes of Respondents (Hypothetical Data)

<i>Income Class</i>	<i>Number of Liters Purchased</i>				<i>Total</i>
	<i>Zero</i>	<i>One</i>	<i>Two</i>	<i>Three or More</i>	
Less than \$15,000	160	25	15	0	200
\$15,000–\$24,999	120	15	10	5	150
\$25,000–\$34,999	60	20	15	5	100
\$35,000–\$49,999	5	10	5	5	25
\$50,000 and over	5	5	5	10	25
Total	250	75	50	25	500

students of elementary statistics as the frequency distribution. A simple tabulation or distribution consists of a count of the number of responses that occur in each of the data categories that comprise a variable. A cross-tabulation is one of the more commonly employed and useful forms of tabulation for analytical purposes. A cross-tabulation involves the simultaneous counting of the number of observations that occur in each of the data categories of two or more variables. An example is given in Table 11.2. We shall examine the use of cross-tabulations in detail later in the chapter.

Although computer analysis provides the advantages of flexibility and ease when manipulating data, these very advantages increase the importance of planning the tabulation analysis. There is a common tendency for the researcher to decide that, because cross-tabulations (and correlations) are so easily obtained, large numbers of tabulations should be run. Not only is this methodologically unsound, but in commercial applications, it is often costly in computer time and the time of the analyst as well. For 50 variables, for example, there are 1,225 different two-variable cross-tabulations that can be made. Only a few of these are potentially of interest in a typical study.

BASIC CONCEPTS OF ANALYZING ASSOCIATIVE DATA

Our brief discussion of cross-tabulations marked the beginning of a major topic of this chapter—the analysis of associative data. Although we shall continue to be interested in

the study of variation in a single variable (or a composite of variables), a large part of the rest of the chapter will focus on methods for analyzing how the variation of one variable is associated with variation in other variables.

The computation of row or column percentages in the presentation of cross-tabulations is taken up first. We then show how various insights can be obtained as one goes beyond two variables in a cross-tabulation to three (or more) variables. In particular, examples are presented of how the introduction of a third variable can often refine or explain the observed association between the first two variables.

Bivariate Cross-Tabulation

Cross-tabulation represents the simplest form of associative data analysis. At the minimum, we can start out with only two variables, such as occupation and education, each of which has a discrete set of exclusive and exhaustive categories. Data of this type are called qualitative or categorical since each variable is assumed to be nominal scaled. This cross-tabulation analysis is known as bivariate cross-tabulation. Bivariate cross-tabulation is widely used in marketing applications to analyze variables at all levels of measurement. In fact, it is the single most widely used bivariate technique in applied settings. Reasons for the continued popularity of bivariate cross-tabulation include the following (Feick, 1984, p. 376):

1. It provides a means of data display and analysis that is clearly interpretable, even to the less statistically inclined researcher or manager.

202 • ANALYSIS AND MODELING

2. A series of bivariate tabulations provides clear insights into complex marketing phenomena that might be lost in an analysis with many variables.
3. The clarity of interpretation affords a more readily constructed link between market research and market action.
4. Bivariate cross-tabulations may lessen the problems of sparse cell values that often plague the interpretation of discrete multivariate analyses (bivariate cross-tabulations require that the expected number of respondents in any table cell be 5).

The entities being cross-classified are often called units of association—usually people, objects, or events. The cross-tabulation, at its simplest, consists of a simple count of the number of entities that fall into each of the possible categories of the cross-classification. Excellent discussions of ways to analyze cross-tabulations can be found in Hellevik (1984) and Zeisel (1957).

However, we usually want to do more than show the raw frequency data. At the very least, row or column percentages (or both) are usually computed.

Percentages

The simple mechanics of calculating percentages are known to all of us. We are also aware that the general purpose of percentages is to serve as a relative measure; that is, they are used to indicate more clearly the relative size of two or more numbers.

The ease and simplicity of calculation, the general understanding of its purpose, and the near universal applicability have made the percent statistic or its counterpart, the proportion, the most widely used statistical tool in marketing research. Yet its simplicity of calculation is sometimes deceptive, and the understanding of its purpose is frequently insufficient to ensure sound application and interpretation. The result is that the percent statistic is often the source of misrepresentations, either inadvertent or intentional.

The sources of problems in using percentages are largely the following:

- Identifying the direction in which percentages should be computed
- Knowing how to interpret percentage of change

Both these problems can be illustrated by a small numerical example. Let us assume that KEN's Original, a small regional manufacturer of salad dressings, is interested in testing the effectiveness of spot TV ads in increasing consumer awareness of a new brand—called Life. Two geographic areas are chosen for the test: (a) test area A and (b) control area B. The test area receives a media weight of five 15-second television spots per week over an 8-week period, whereas the control area receives no spot TV ads at all. (Other forms of advertising were equal in each area.)

Assume that telephone interviews were conducted before and after the test in each of the areas. Respondents were asked to state all the brands of salad dressing they could think of on an aided basis. If Life was mentioned, it was assumed that this constituted consumer awareness of the brand. However, as it turned out, sample sizes differed across all four sets of interviews. This common fact of survey life (i.e., variation in sample sizes) increases the value of computing percentages.

Table 11.3 shows a set of frequency tables that were compiled before and after a TV ad for Life salad dressing was aired. (All four samples were independent samples.) Interpretation of Table 11.3 would be hampered if the data were expressed as raw frequencies and different percentage bases were reported. Accordingly, Table 11.3 shows the data, with percentages based on column and row totals. Which of these percentages is more useful for analytical purposes?

Direction in Which to Compute Percentages

In examining the relationship between two variables, it is often clear from the context that one variable is more or less the independent or control variable and the other is the dependent or criterion variable. In cases where this distinction is clear, the rule is to compare percentages within levels of the dependent variable.

In Table 11.3, the control variable is the experimental area (test vs. control), and the

Table 11.3 Aware of Life Salad Dressing—Before and After Spot TV

	<i>Before Spot TV</i>			<i>Area</i>	<i>After Spot TV</i>		
	<i>Aware</i>	<i>Not Aware</i>	<i>Total Area</i>		<i>Aware</i>	<i>Not Aware</i>	<i>Total Area</i>
Test area							
Frequency	250	350	600		330	170	550
Column %	61	59	60		67	44	57
Row %	42	58			66	34	
Control area							
Frequency	160	240	400		160	220	380
Column %	39	41	40		33	56	43
Row %	40	60			42	58	
Total	410	590	1,000	Total	490	390	880
Before TV spot (%)	41	59	100	After (%)	56	44	100

dependent variable is awareness. When comparing awareness in the test and control areas, row percentages are preferred. We note that before the spot TV campaign, the percentage of respondents who are aware of Life is almost the same between test and control areas: 42% and 40%, respectively.

However, after the campaign the test-area awareness level moves up to 66%, whereas the control-area awareness (42%) stays almost the same. The small increase of 2 percentage points reflects either sampling variability or the effect of other factors that might be serving to increase awareness of Life in the control area.

On the other hand, computing percentages across the independent variable (column percent) makes little sense. We note that 61% of the aware group (before the spot TV campaign) originates from the test area; however, this is mainly a reflection of the differences in total sample sizes between test and control areas.

After the campaign, we note that the percentage of aware respondents in the control area is only 33% versus 39% before the campaign. This may be erroneously interpreted as indicating that spot TV in the test area depressed awareness in the control area. But we know this to be false from our earlier examination of raw percentages.

It is not always the case that one variable is clearly the independent or control variable and

the other is the dependent or criterion variable. This should pose no particular problem as long as we agree, for analysis purposes, which variable is to be considered the control variable. Indeed, cases often arise in which each of the variables in turn serves as the independent and dependent variable.

Interpretation of the Percentage Change

A second problem that arises in the use of percentages in cross-tabulations is the choice of which method to use in measuring differences in percentages. There are three principal ways to portray percentage change:

1. The absolute difference in percentages
2. The relative difference in percentages
3. The percentage of possible change in percentages

The same example can be used to illustrate the three methods.

Absolute Percentage Increase

Table 11.4 shows the percentage of respondents who are aware of Life before and after the spot TV campaign in the test and control areas. First, we note that the test-area respondents displayed a greater absolute increase in

204 • ANALYSIS AND MODELING

Table 11.4 Aware of Life—Percentages Before and After the Spot TV Campaign

	<i>Before the Campaign</i>	<i>After the Campaign</i>
Test area	42	66
Control area	40	42

awareness. The increase for the test-area respondents was 24 percentage points, whereas the control-area awareness increased by only 2 percentage points.

Relative Percentage Increase

The relative increase in percentage is $[(66 - 42)/42] \times 100 = 57\%$ and $[(42 - 40)/40] \times 100 = 5\%$, respectively, for test- and control-area respondents.

Percentage Possible Increase

The percentage of possible increase for the test area is computed by first noting that the maximum percentage point increase that could have occurred is $100 - 42 = 58$ points. The increase actually registered is 24 percentage points, or $100(24/58) = 41\%$ of the maximum possible. That of the control area is $100(2/60) = 3\%$ of the maximum possible.

In terms of the illustrative problem, all three methods give consistent results in the sense that the awareness level in the test area undergoes greater change than that in the control area. However, in other situations, conflicts among the measures may arise.

The absolute-difference method is simple to use and requires only that the distinction between percentage and percentage points be understood. The relative-difference method can be misleading, particularly if the base for computing the percentage change is small. The percentage-of-possible-difference method takes cognizance of the greater difficulty associated with obtaining increases in awareness as the difference between potential-level and realized-level decreases. In some studies, all three measures are used, inasmuch as they emphasize different aspects of the relationship.

Introducing a Third Variable Into the Analysis

Cross-tabulation analysis to investigate relationships need not stop with two variables. Often, much can be learned about the original two-variable association through the introduction of a third variable that may refine or explain the original relationship. In some cases, it may show that the two variables are related even though no apparent relationship exists before the third variable is introduced. These ideas are most easily explained by example.

Consider the situation facing MCX, a company that specializes in telecommunications equipment for the residential market. The company has recently test-marketed a new device for the automatic recording of home telephone messages without an answering machine. Several months after the introduction, a telephone survey was taken in which respondents in the test area were asked whether they had adopted the innovation. The total number of respondents interviewed was 600.

One of the variables of major interest in this study was the age of the respondent. Based on earlier studies of the residential market, it appeared that adopters of the firm's new products tended to be younger than 35 years old. Accordingly, the market analyst decides to cross-tabulate adoption and respondent age. Respondents are classified into the categories "under 35 years" (< 35) and "equal to or greater than 35 years" (≥ 35) and then cross-classified by adoption or not. Table 11.5 shows the full three-variable cross-tabulation. It seems that the total sample of 600 is split evenly between those who are younger than 35 years of age and those who are 35 years of age or older. Younger respondents display a higher percentage of adoption ($37\% = (100 + 11)/300$) than older respondents ($23\% = (60 + 9)/300$).

Analysis and Interpretation

The researcher is primarily interested in whether this finding differs when gender of the respondent is introduced into the analysis. As it turned out, 400 respondents in the total sample were men, whereas 200 were women.

Table 11.5 Adoption—Percentage by Gender and Age

Frequency	Men			Women		
	< 35 Years	≥ 35 Years	Total %	< 35 Years	≥ 35 Years	Total %
Adopters						
Number of cases	100	60	160	11	9	20
Column %	50	30	40	11	9	10
Row %	62.5	37.5		55	45	
Nonadopters						
Number of cases	100	140	240	89	91	180
Column %	50	70	60	89	91	90
Row %	41.7	58.3		49.4	50.6	
Total	200	200	400	100	100	200
Percentage	50	50		50	50	

Table 11.5 shows the results of introducing gender as a third classificatory variable. In the case of men, 50% of the younger men adopt compared with only 30% of the older men. In the case of women, the percentages of adoption are much closer. Even here, however, younger women show a slightly higher percentage of adoption (11%) than older women (9%).

The effect of gender on the original association between adoption and age is to refine that association without changing its basic character; younger respondents show a higher incidence of adoption than older respondents. However, what can now be said is the following: If the respondent is a man, the differential effect of age on adoption is much more pronounced than if the respondent is a woman.

Figure 11.2 shows this information graphically. The height of the bars within each rectangle represents the percentage of respondents who are adopters. The relative width of the bars denotes the relative size of the categories—men versus women—representing the third variable, gender. The shaded portions of the bars denote the percentage adopting by gender, and the dashed line represents the weighted average percentage adopting by and across genders.

It is easy to see from Figure 11.2 that adoption differs by age group (37% vs. 23%). Furthermore, the size of the difference depends on the gender of the respondent: Men display a

relatively higher rate of adoption, compared with women, in the younger age category.

Other Possible Relationships in Cross-Tabulation Data

The relationships shown in Table 11.5 were consistent for men and women and for age categories. Often, the relationships portray other forms. For example, suppose the introduction of gender as a third variable shows that there is a strong association between adoption and age but that this association runs in opposite directions for men versus women. The overall effect is to suggest that adoption and age are not associated (when the effect of gender is not held constant).

This is often called a suppressor effect. That is, failure to control on gender differences suppresses the relationship between adoption and age to the point where there appears to be no association at all. However, once we tabulate adoption by age within the level of gender, the association becomes evident, as shown in Figure 11.3.

As an additional illustration of what might happen when a third variable (gender) is introduced, consider an example where the association between adoption and age is not affected at all by the introduction of gender. In this case, gender is independent of the association between adoption and age. Although a figure is not shown for this simple case, it should be clear

206 • ANALYSIS AND MODELING

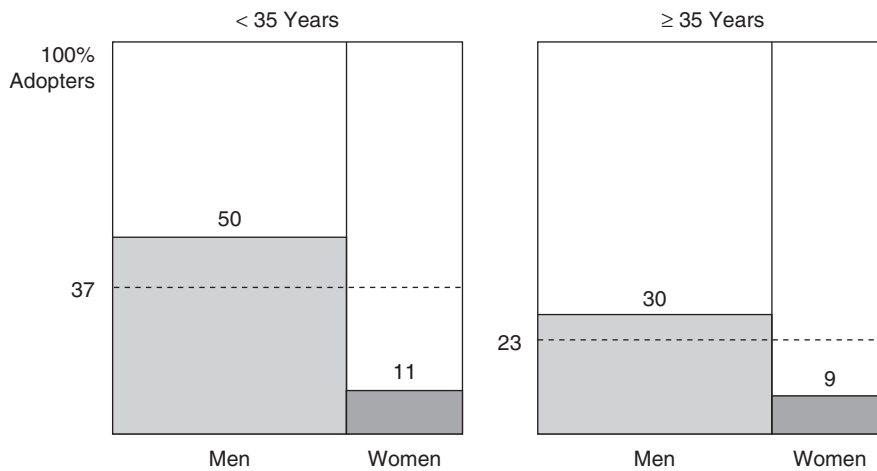


Figure 11.2 Adoption—Percentage by Age and Gender

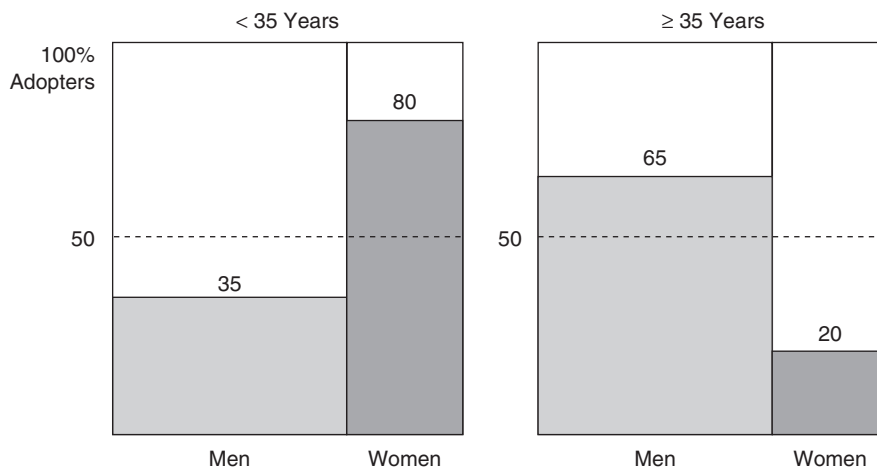


Figure 11.3 Adoption—Percentage by Age and Gender

that the bars for the separate men and women classes will look the same.

Our discussion so far about the various ways the original relationship can be modified has assumed that gender was not related to the initial independent variable, age. However, a fourth possibility exists in which the original relationship disappears upon the introduction of a third variable. Behavioral scientists often use the term *explanation* for this case. In order for the original association to vanish, it is necessary

that the third variable, gender, be associated with the original independent variable, age.

To illustrate the idea of third-variable explanation, consider the new association between adoption and age, where a higher percentage of adopters is drawn from the younger age group. However, within each separate category of gender, there is an equal percentage of adopters. The apparent relationship between adoption and gender is due solely to the difference in the relative size of

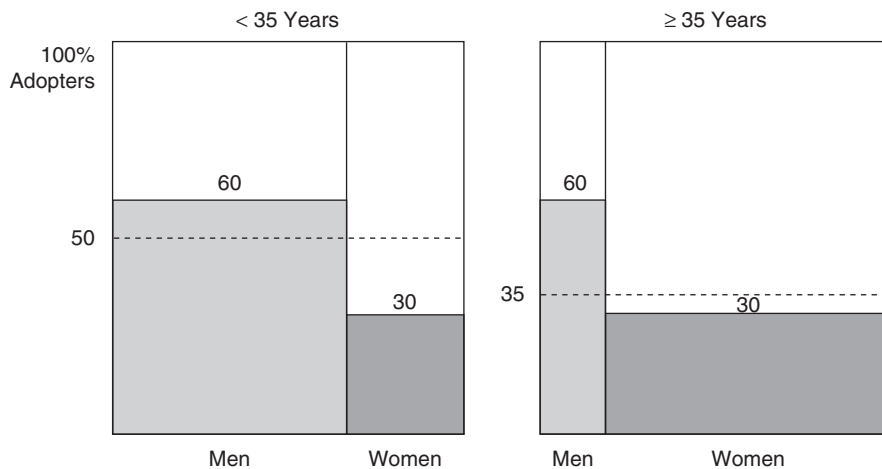


Figure 11.4 Adoption—Percentage by Age and Gender

the subsamples of men versus women within the two age categories.

Figure 11.4 shows an example of this effect graphically. In the case of the under-35 age group, there are twice as many men as women. However, in the 35-and-over age group, there are five times as many women as men. These differences in subsample size affect the weighted-average percentages that are shown as dashed lines in the rectangles.

In the present case, the gender variable is said to explain the (apparent) relationship between adoption and age. As observed from Figure 11.4, the percentage of adopters is not associated with age, once the data are examined separately for men and women.

Recapitulation

Representatives of three-variable association can involve many possibilities that could be illustrated by the preceding adoption-age-gender example:

1. In the example presented, adoption and age exhibit initial association; this association is still maintained in the aggregate but is refined by the introduction of the third variable, gender.
2. Adoption and age do not appear to be associated. However, adding and controlling on the third variable, gender, reveals suppressed association between the first two variables within

the separate categories of men and women. In the two-variable cases, men and women exhibit opposite patterns, canceling each other out.

3. Adoption and age are not associated to begin with; furthermore, introducing a third independent variable, gender, does not change the situation.
4. Adoption and age exhibit initial association, which then disappears upon the introduction of the explanatory variable, gender.

Although the preceding examples were contrived to illustrate the concepts, the results are not unusual in practice. It goes almost without saying that the introduction of a third variable can often be useful in the interpretation of two-variable cross-tabulations.

However, the reader should be aware of the fact that we have deliberately used the phrase *associated with* rather than *caused by*. Association of two or more variables does not imply causation, and this statement is true regardless of our preceding efforts to refine some observed two-variable association through the introduction of a third variable.

In principle, of course, we could cross-tabulate four or even more variables with the possibility of obtaining further insight into lower-order (e.g., two-variable) associations. However, somewhere along the line, a problem arises in maintaining an adequate cell size for all

208 • ANALYSIS AND MODELING

categories. Unless sample sizes are extremely large in the aggregate and the number of categories per variable is relatively small, cross-tabulations rarely can deal with more than three variables at a time. A further problem, independent of sample size, concerns the high degree of complexity of interpretation that is introduced by cross-tabulations involving four or more variables. In practice, most routine applications of cross-tabulation involve only two variables at a time.

As noted in Table 11.5, there are definite advantages associated with having a two-category criterion variable, such as adoption versus nonadoption. In many applications, however, the criterion variable will have more than two categories. Cross-tabulations can still be prepared in the usual manner, although they become somewhat more tedious to examine.

BIVARIATE ANALYSIS: DIFFERENCE BETWEEN SAMPLE GROUPS

Marketing activities largely focus on the identification and description of market segments. These segments may be defined demographically, attitudinally, by the quantity of the product used, by activities participated in or interests, by opinions, or by a multitude of other measures. The key component of each of these variables is the ability to group respondents into market segments. Often this segmentation analysis involves identifying differences and asking questions about the marketing implications of those differences: Do differences in satisfaction exist for the two or more groups that are defined by age categories?

Bivariate statistical analysis refers to the analysis of relationships between two variables. These analyses are often differences between respondent groups. In the following discussion, we explore bivariate statistical analysis and focus on the two-variable case as a bridge between the comparatively simple analyses already discussed and the more sophisticated techniques that will command our attention in later chapters. We begin with what is perhaps the most used test of market researchers: cross-tabulation. Next, we consider analysis of differences in group means. First, we discuss the *t* test

of differences in means of two independent samples, and then we look at one-way analysis of variance (ANOVA) for *k* groups. Finally, we provide a discussion of some of the more widely used nonparametric techniques. These are but a few of the possible parametric and nonparametric analyses that could be discussed (see Table 11.6).

Bivariate Cross-Tabulation

The chi-square statistic in a contingency table analysis is used to answer the following question: Is the observed association between the variables in the cross-tabulation statistically significant?

Often called chi-square analysis, this technique is used when the data consist of counts or frequencies within categories of a tabulation or cross-tabulation table. In conjunction with the cross-tabulation we will introduce the chi-square statistic, χ^2 , to determine the significance of observed association in cross-tabulations involving two or more variables. This is typically called a test of independence.

Cross-tabulation represents the simplest form of associative data analysis. At the minimum, we can start out with a bivariate cross-tabulation of two variables, such as occupation and education, each of which identifies a set of exclusive and exhaustive categories. We know that such data are called qualitative or categorical because each variable is assumed to be only nominal scaled. Bivariate cross-tabulation is widely used in marketing applications to analyze variables at all levels of measurement. In fact, it is the single most widely used bivariate technique in applied settings.

In marketing research, observations may be cross-classified, such as when we are interested in testing whether occupational status is associated with brand loyalty. Suppose, for illustrative purposes, that a marketing researcher has assembled data on brand loyalty and occupational status—white collar, blue collar, and unemployed or retired—that describe consumers of a particular product class. The data for our hypothetical problem appear in Table 11.7.

A total of four columns, known as banner points, are shown. Four rows or stubs are also shown. Professional cross-tabulation software

Table 11.6 Selected Nonparametric Statistical Tests for Two-Sample Cases

Sample Cases	Level of Measurement		
	Nominal	Ordinal	Interval/Ratio
Two-sample related samples	McNemar test for the significance of changes	Sign test Wilcoxon matched-pairs signed-ranks test	
Independent samples	Fischer exact probability test Chi-square test for two independent samples	Median test Mann-Whitney <i>U</i> test Kolmogorov-Smirnov two-sample test Wald-Wolfowitz runs test	<i>t</i> test One-way ANOVA

will output tables that join multiple variables on the column banner points such that loyalty and another variable, such as usage occasion, could be analyzed simultaneously.

In a study of 230 customers, we are interested in determining if occupational status may be associated with the characteristic loyalty status. The data suggest that a relationship exists, but is the observed association a reflection of sampling variation, or is the variation great enough that we can conclude that a true relationship exists? Expressed in probability terms, we may ask, "Are the conditional probabilities of being highly loyal, moderately loyal, and brand switcher, given the type of occupational status, equal to their respective marginal probabilities?"

In analyzing the problem by means of chi-square, we make use of the marginal totals (column and row totals) in computing theoretical frequencies given that we hypothesize independence (no relationship) between the attributes loyalty status and occupational status. For example, we note from Table 11.7 that 33.9% (78/320) of the respondents are highly loyal. If possession of this characteristic is independent of occupational status, we would expect that 33.9% (78/320) of the 90 respondents classified as white-collar workers (i.e., 30.5) would be highly loyal. Similarly, 37.8% (87/320) of the 90 (34.1) would be moderately loyal, and 28.3% (65/320) of the 90 (25.4) would be brand

switchers. In a similar fashion, we can compute theoretical frequencies for each cell on the null hypothesis that loyalty status is statistically independent of occupational status. (It should be noted that the frequencies are the same, whether we start with the percentage of the row or the percentage of the column.)

The theoretical frequencies (under the null hypothesis) are computed and appear in parentheses in Table 11.7. The chi-square statistic is then calculated (and shown in the table) for each of the data cells in the table using the observed and theoretical frequencies:

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - F_i)^2}{F_i},$$

where f_i = actual observed frequency, F_i = theoretical expected frequency, and k = number of cells ($r \times c$).

The appropriate number of degrees of freedom to use in this example is four. In general, if we have R rows and C columns, the degrees of freedom associated with the chi-square statistic are equal to the product

$$(R - 1)(C - 1).$$

If we use a significance level of .05, the probability table value of chi-square is 9.488. Hence,

210 • ANALYSIS AND MODELING

Table 11.7 Contingency Table of Observed Versus Theoretical Frequencies for Brand-Loyalty Illustration*N*, Expected *n* χ^2 Contribution

<i>Occupational Status</i>	<i>Highly Loyal</i>	<i>Moderately Loyal</i>	<i>Brand Switchers</i>	<i>Total Number (% Distribution)</i>
White collar	30 (30.5) .01	42 (34.1) 1.86	18 (25.4) 2.17	90 (39.1)
Blue collar	14 (22.1) 2.93	20 (24.5) .86	31 (18.4) 8.68	65 (28.3)
Unemployed/retired	34 (25.4) 2.88	25 (28.4) .40	16 (21.2) 1.27	75 (32.6)
Total, <i>n</i> (%)	78 (33.9)	87 (37.8)	65 (28.3)	230 $\chi^2 =$ 21.08

we reject the null hypothesis of independence between the characteristics loyalty status and occupational status because the computed χ^2 of 21.08 is greater than the table value of 9.488.

When the number of observations in a cell is less than 10 (or where a 2×2 contingency table is involved), a correction factor must be applied to the formula for chi-square. The numerator within the summation sign becomes $(|f_i - F_i| - \frac{1}{2})^2$, where the value $\frac{1}{2}$ is the Yates continuity correction. This correction factor adjusts for the use of a continuous distribution to estimate probability in a discrete distribution.

Chi-square analysis can be extended to deal with more than two variables. No new principles are involved. Three characteristics of the technique should be borne in mind, however. First, chi-square analysis deals with counts (frequencies) of data. If the data are expressed in percentage form, they should be converted to absolute frequencies. Second, the technique assumes that the observations are drawn independently. Third, the chi-square statistic cannot describe the relationship; it only gauges its statistical significance, regardless of logic or sense (Semon, 1999). To assess the nature of the relationship, the researcher must look at the table and indicate how the variables appear to be related—a type of eyeball approach. This may involve examining any of the following combinations: (a) the variable combinations that produce large χ^2 values in the cells, (b) those with a large difference between the observed

and expected frequencies, or (c) those where the cell frequency count, expressed as a percentage of the row total, is most different from the total column percentage (marginal column percent). When variables are ordered or loosely ordered, a pattern can sometimes be observed by marking cells with higher than expected observed frequencies with a (+) and those with lower than expected observed frequencies with a (–).

Bivariate Analysis: Differences in Means and Proportions

A great amount of marketing research is concerned with estimating parameters of one or more populations. In addition, many studies go beyond estimation and compare such population parameters by testing hypotheses about differences between them. Means, proportions, and variances are often the summary measures of concern. Our concern at this point is with differences in means and proportions. Direct comparisons of variances are a special case of the more general technique of analysis of variance, which is covered later in this chapter.

Standard Error of Differences

Here we extend the topic of sampling distributions and standard errors as they apply to a single statistic—to cover differences in statistics and show a traditional hypothesis for differences.

Standard Error of Difference of Means

The starting point is the standard error of the difference. For two samples, A and B , that are independent and randomly selected, the standard error of the differences in means is calculated by

$$\sigma_{\bar{x}_A - \bar{x}_B} = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$$

This estimate of the standard error is appropriate for use in the denominator of the Z test formula.

If the population standard deviations, σ_p , are not known, then the estimated standard error becomes

$$est. S_{\bar{x}_A - \bar{x}_B} = \sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}$$

For relatively small samples, the correction factor $N/n_i - 1$ is used, and the resulting formula for the estimated standard error is

$$est. S_{\bar{x}_A - \bar{x}_B} = \sqrt{\frac{S_A^2}{n_A - 1} + \frac{S_B^2}{n_B - 1}}$$

Of course, these formulas would be appropriate for use in the denominator of the t test.

Standard Error of Differences of Proportions

Turning now to proportions, the derivation of the standard error of the differences is somewhat similar. Specifically, for large samples,

$$est. \sigma_{p_A - p_B} = \sqrt{\frac{p_A(1 - p_A)}{n_A} + \frac{p_B(1 - p_B)}{n_B}}$$

For small samples, the correction factor is applied, resulting in

$$est. \sigma_{p_A - p_B} = \sqrt{\frac{p_A(1 - p_A)}{n_A - 1} + \frac{p_B(1 - p_B)}{n_B - 1}}$$

This estimate would again be appropriate for use in the denominator of the Z test of proportions.

Testing of Hypotheses

When applying the standard error formulas for hypotheses testing concerning parameters, the following conditions must be met:

1. Samples must be independent.
2. Individual items in samples must be drawn in a random manner.
3. The population being sampled must be normally distributed (or the sample must be of sufficiently large size).
4. For small samples, the population variances must be equal.
5. The data must be at least interval scaled.

When these five conditions are met or can at least be reasonably assumed to exist, the traditional approach is as follows.

1. The null hypothesis (H_0) is specified such that there is no difference between the parameters of interest in the two populations (e.g., $H_0: \mu_A - \mu_B = 0$); any observed difference is assumed to occur solely because of sampling variation.

2. The alpha risk is established ($\alpha = .05$ or other value).

3. A Z value is calculated by the appropriate adaptation of the Z formula. For testing the difference between two means, Z is calculated in the following way:

$$Z = \frac{(\bar{X}_A - \bar{X}_B) - (\mu_A - \mu_B)}{\sigma_{\bar{x}_A - \bar{x}_B}} = \frac{(\bar{X}_A - \bar{X}_B) - 0}{\sigma_{\bar{x}_A - \bar{x}_B}}$$

and, for proportions,

$$Z = \frac{(p_A - p_B) - (\pi_A - \pi_B)}{\sigma_{p_A - p_B}} = \frac{(p_A - p_B) - 0}{\sigma_{p_A - p_B}}$$

212 • ANALYSIS AND MODELING

For unknown population variance and small samples, the student t distribution must be used, and for means, t is calculated from

$$t = \frac{(\bar{X}_A - \bar{X}_B) - (\mu_A - \mu_B)}{S_{\bar{x}_A - \bar{x}_B}} = \frac{(\bar{X}_A - \bar{X}_B) - 0}{S_{\bar{x}_A - \bar{x}_B}}$$

4. The probability of the observed difference of the two sample statistics having occurred by chance is determined from a table of the normal distribution (or the t distribution, interpreted with $[n_A + n_B - 2]$ degrees of freedom).

5. If the probability of the observed difference having occurred by chance is greater than the alpha risk, the null hypothesis is accepted; it is concluded that the parameters of the two universes are not significantly different. If the probability of the observed difference having occurred by chance is less than the alpha risk, the null hypothesis is rejected; it is concluded that the parameters of the two populations differ significantly. In an applied setting, there are times when the level at which significance occurs (the alpha level) is reported, and management decides whether to accept or reject.

An example will illustrate the application of this procedure. Let us assume we have conducted a survey of detergent and paper goods purchases from supermarkets among urban (population A) and rural (population B) families (see Table 11.8).

The question facing the researcher is the following: "Do urban families spend more on these items, or is the \$1.20 difference in means caused by sampling variations?" We proceed as follows. The hypothesis of no difference in means is established. We assume the alpha risk is set at .05. Since a large sample test is called for, the Z value is calculated using the separate variances estimate of the standard error of differences in means:

$$S_{\bar{x}_A - \bar{x}_B} = \sqrt{\frac{(10.0)^2}{400} + \frac{(9.0)^2}{225}} = \$0.78.$$

The Z value is then determined to be

$$Z = \frac{(32.0 - 30.8) - 0}{0.78} = +1.54.$$

The probability of the observed difference in the sample means having been due to sampling is specified by finding the area under the normal curve that falls to the right of the point $Z = +1.54$. Consulting a table of the cumulative normal distribution, we find this area to be $1.0 - .9382 = 0.0618$. Since this probability associated with the observed difference ($p = 0.06$) is greater than the preset alpha, a strict interpretation would be that there is no difference between the two types of families concerning the average expenditure on nonfood items. In a decision setting, however, the manager would have to determine whether this probability (0.06) is low enough to conclude, on pragmatic grounds, that the families do not differ in their behavior. As stated previously, often there is no preset alpha, and decisional considerations require the manager to determine the meaning of the reported alpha.

To illustrate the small sample case, let us assume that we obtain the same mean values and get values for s_A and s_B such that $s_{x_A} - s_{x_B} = \$0.78$ from samples $n_A = 15$ and $n_B = 12$. With these data, we calculate t as follows:

$$t = \frac{(32.0 - 30.8) - 0}{0.78} = 1.54.$$

The critical value of t is obtained from a table of percentiles for the t distribution. For, say, $\alpha = .05$, we determine the critical value of t for $(n_A + n_B - 2) = 25$ degrees of freedom to be 1.708 (one-tailed test).

Table 11.8 Sample Group and Average Expenditure

Family Type	Sample	Average Amount Spent	Standard Deviation
Urban (Sample A)	$n_A = 400$	$\bar{X}_A = \$32.00$	$s_A = \$10.00$
Rural (Sample B)	$n_B = 225$	$\bar{X}_B = \$30.80$	$s_B = \$9.00$

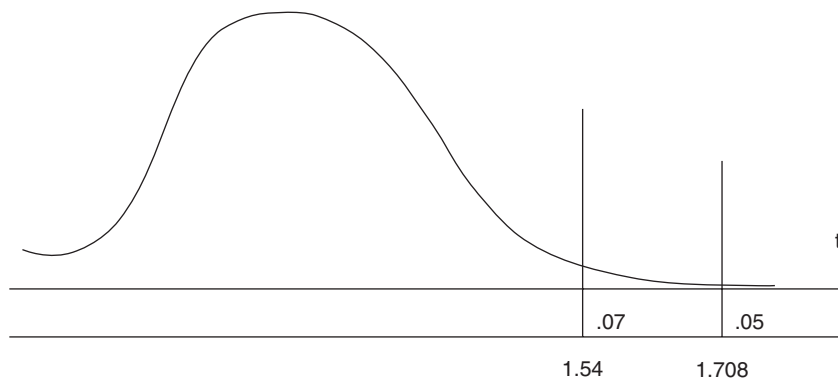


Figure 11.5 *t* Distribution

Since the calculated t of $1.54 < 1.708$, we cannot reject the hypothesis of no difference in average amount spent by the two types of families. This result is shown in Figure 11.5.

When samples are not independent, the same general procedure is followed. The formulas for calculating the test statistics differ, however.

Testing the Means of Two

Groups: The Independent Samples *t* Test

The t distribution revolutionized statistics and the ability to work with small samples. Prior statistical work was based largely on the value of Z , which was used to designate a point on the normal distribution where population parameters were known. For most market research applications, it is difficult to justify the Z test's assumed knowledge of μ and σ . The t test relaxes the rigid assumptions of the Z test by focusing on sample means and variances (X and s). The t test is a widely used market research statistic to test for differences between two groups of respondents.

In the previous section, the t statistic was described. Most computer programs recognize two versions of this statistic. In the previous section, we presented what is called the separate variance estimate formula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

This formula is appropriate where differences in large samples are tested.

The second method, called the pooled variance estimate, computes an average for the samples that is used in the denominator of the statistic:

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

The pooling of variances is a simple averaging of variances that is required when (a) testing for the same population proportion in two populations or (b) testing the difference in means between two small samples.

Table 11.9 shows a sample output of t tests from SPSS. Two respondent groups were identified for the supermarket study:

1. Respondents who were males
2. Respondents who were females

The analysis by gender is shown in Part A of the output for the attitudinal dimensions, friendliness of clerks and helpfulness of clerks. It will be noted that the sample sizes are approximately equal. To show a contrast, we have included in Part B of the output a t test from another study where the number of males and females varies widely. In all three examples, the differences between separate-variance and pooled-variance analysis are very small. Table 11.9 also shows an F statistic. This is Levene's test for equality of variances. Using this test, the researcher knows which t test output to use—equal variances assumed or equal variances not assumed.

214 • ANALYSIS AND MODELING

Table 11.9 Selected Output From SPSS *t* Test

	<i>Mean</i>	<i>n</i>	<i>F</i>	<i>Significance</i>	<i>t</i>	<i>Equal Variances Not Assumed (Separate)</i>		<i>Equal Variances Assumed (Pooled)</i>		
						<i>df</i>	<i>Significance (Two-Tailed)</i>	<i>t</i>	<i>df</i>	<i>Significance (Two-Tailed)</i>
Part A										
“Friendliness of clerks”			2,713	.102	1.301	126.213	.19	1.298	128	.197
Male	4.55	64								
Female	4.82	66								
“Helpfulness of clerks”			15,401	.000	.308	105.177	.758	.308	126	.758
Male	4.50	64								
Female	4.56	64								
Part B										
“Shopping experience”										
Male	1.977	44	.089	.765	.569	66,993	.571	.590	189	.556
Female	1.857	147								

Testing of Group Means: Analysis of Variance

ANOVA is a logical extension of the independent groups *t* test methodology. Rather than test differences between two group means, we test the overall difference in *k* group means, where the *k* groups are thought of as levels of a treatment or control variable(s) or factor(s). ANOVA is a general set of methodologies that handle many different types of research and experimental designs. Traditional use of analysis of variance has been to isolate the effects of experimental variables that are related in various experimental situations. The respondent receives a combination of the treatment levels from the different factors, and the response is measured. More specifically, ANOVA is used to test the statistical significance of differences in mean responses given the introduction of one or more treatment effects.

Much experimental work has been conducted in medicine and agriculture. In pharmaceutical drug testing, positive and negative effects of dosage and formulation are measured over time

and for different types of ailments and patient illness levels. The variables influencing the results are called experimental factors. In agriculture, crop yields are measured for plots of land, each of which receives a different treatment level that is defined by a factor or control variable. Control variables in this application might include seed type, fertilizer type, fertilizer dosage, temperature, moisture, and many other variables thought to influence production. In each of these plots, the average yield is measured and analyzed to determine the effect of the specific measured levels of the factors being evaluated. Marketing research experiments have control variables that are certainly different from agricultural experiments, but the principles are the same.

The proper planning and design of the relationships between the experimental factors results in methodologies having unfamiliar names such as completely randomized, randomized block, Latin square, and factorial designs (see Smith & Albaum, 2005, chap. 8). Here we discuss the two most basic forms of the methodology: the one-way ANOVA and the two-factor ANOVA designs (see Exhibit 11.1).

Example: It is well-known that interest ratings for TV ads are related to the advertising message. A simple one-way ANOVA to investigate this relationship might compare three messages:

Advertising message A	Advertising message B	Advertising message C
-----------------------	-----------------------	-----------------------

Two-factor ANOVA includes a second factor, possibly the type of advertisement (magazine or TV):

	Message A	Message B	Message C
Magazine Ad			
TV Ad			

Each of the cells in this matrix would contain an average interest rating for the measures taken for the particular combination of message and media.

Exhibit 11.1 ANOVA Designs

It is hoped that this brief introduction to the idea behind an ANOVA will reduce the impression that the technique is used to test for significant differences among the variances of two or more sample universes. This is not strictly the case. ANOVA is used to test the statistical significance of differences in mean responses given the introduction of one or more treatments effects.

The ANOVA Methodology

The appropriateness of the label *analysis of variance* comes from explaining the variation in responses to the various treatment combinations. The methodology for explaining this variation is explained in Exhibit 11.2, which presents an example regarding responses to messages.

The basic idea of ANOVA is to compare the between-treatment groups sum of squares (after dividing by degrees of freedom to get a mean square) with the within-treatment group sum of squares (also divided by the appropriate number of degrees of freedom). This is the *F* statistic, which indicates the strength of the grouping factor. Conceptually,

$$F = \frac{\text{Sampling variance} + \text{Variance due to effect of treatment}}{\text{Sampling variance}}$$

The larger the ratio of between to within, the more we are inclined to reject the null

hypothesis that the group mean $\mu_1 = \mu_2 = \mu_3$. Conversely, if the three sample means were very close to each other, the between-samples sum of squares would be close to zero, and we would conclude that the population means are the same, once we consider the variability of individual cases within each sample group.

However, to make this comparison, it is necessary to assume that the error term distribution has constant variance over all observations. This is the same assumption as was made for the *t* test.

In the next section, we shall (a) use more efficient computational techniques, (b) consider the adjustment for degrees of freedom to obtain mean squares, and (c) show the case of the *F* ratio in testing significance. Still, the foregoing remarks represent the basic ANOVA idea for comparing between- with within-sample variability.

One-Way (Single-Factor) Analysis of Variance

One-way ANOVA is analysis of variance in its simplest (single-factor) form. Suppose a new product manager for the hypothetical Friskie Corp. is interested in the effect of shelf height on supermarket sales of canned dog food. The product manager has been able to secure the cooperation of a store manager to run

216 • ANALYSIS AND MODELING

Exhibit 11.2 Example of ANOVA Methodology

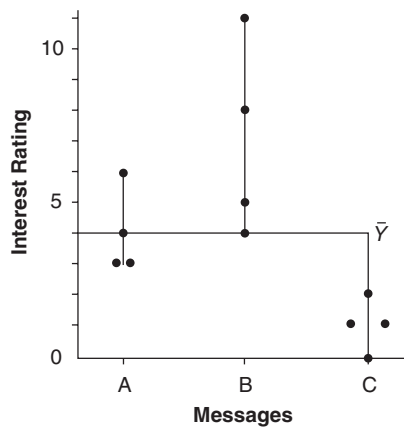
Using the single-factor ANOVA design for the problem described in Exhibit 11.1, the actual data for 12 respondents appear in tabular and graphical format as follows:

<i>Obs #</i>	<i>Msg. A</i>	<i>Msg. B</i>	<i>Msg. C</i>
1	6	8	0
2	4	11	2
3	3	4	1
4	3	5	1
Mean	4	7	1

It is apparent that the messages differ in terms of their distribution, but how do we perform ANOVA to explain these differences? There are three values that must be computed to analyze this pattern of values.

1. Total sum of squares: The grand mean of the 12 observations is computed, followed by the variance of the individual observations from this mean.

$$(X_j - \bar{X})^2: (6 - 4)^2 + (4 - 4)^2 + \dots + (1 + 4)^2 = 110$$

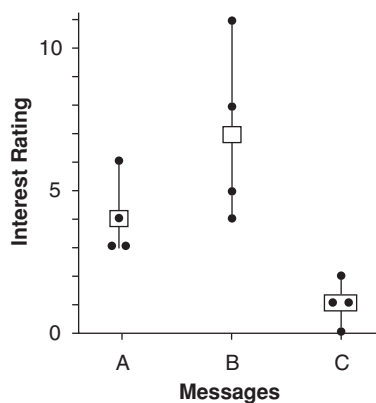


Total Sum of Squares:
Grand Mean = $4.0 = \bar{X}_G$

Computation of Total Sum of Squares:
 $(6-4)^2 + (4-4)^2 + \dots + (1-4)^2 = 110$

2. Between-treatment sum of squares: The means of the factor levels (Messages A, B, and C) are computed, followed by the deviation of the factor-level means from the overall mean, weighted by the number of

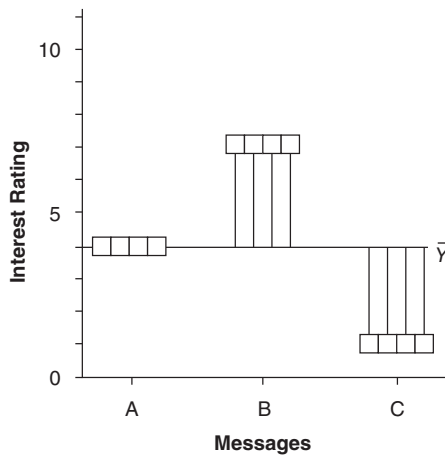
observations $[n(\bar{X}_j - \bar{X})^2]$



Between-Treatment Sum of Squares:
Grand Mean = 4.0
Means of Groups: A = 4; B = 7; C = 1

Computation of Between-Treatment Sum of Squares:
 $4(4 - 4)^2 + 4(7 - 4)^2 + 4(1 - 4)^2 = 72$

3. Within-treatment sum of squares: The means of the factor levels are computed, followed by the deviation of the observations within each factor level from that factor-level mean.



Within-Treatment

Sum of Squares:

Grand Mean = 4.0

Means of Groups: A = 4; B = 7; C = 1

Computation of

Within-Treatment

Sum of Squares:

$$A: (6 - 4)^2 + (4 - 4)^2 + (3 - 4)^2 + (3 - 4)^2 = 6$$

$$B: (8 - 7)^2 + (11 - 7)^2 + (5 - 7)^2 + (4 - 7)^2 = 30$$

$$C: (0 - 1)^2 + (2 - 1)^2 + (1 - 1)^2 + (1 - 1)^2 = \frac{2}{38}$$

Thus, an observation may be decomposed into three terms that are additive, and each explains a particular type of variance:

$$\begin{aligned} \text{Observation} &= \text{Overall mean} + \text{Deviation of the group mean from the overall mean} \\ &+ \text{Deviation of the observation from the group mean} \end{aligned}$$

The overall mean is constant and common to all observations: The deviation of a group mean from the overall mean represents the effect on each observation of belonging to that particular group; the deviation of an observation from its group mean represents the effect on that observation of all variables other than the group variable.

an experiment involving three levels of shelf height (knee level, waist level, and eye level) on sales of a single brand of dog food, which we shall call Snoopy. Assume further that our experiment must be conducted in a single supermarket and that our response variable will be sales, in cans, of Snoopy dog food for some appropriate unit of time. But what shall we use for our unit of time? Sales of dog food in a single store may exhibit week-to-week variation, day-to-day variation, and even hour-to-hour variation. In addition, sales of this particular brand may be influenced by the price or special promotions of competitive brands, the store management's knowledge that an experiment is going on, and other variables that we cannot control at all or would find too costly to control.

Assume that we have agreed to change the shelf height position of Snoopy three times per

day and run the experiment over 8 days. We shall fill the remaining sections of our gondola with a filler brand, which is not familiar to customers in the geographical area in which the test is being conducted. Furthermore, since our primary emphasis is on explaining the technique of analysis of variance in its simplest form (analysis of one factor: shelf height), we shall assign the shelf heights at random over the three time periods per day and not design an experiment to explicitly control and test for within-day and between-day differences. Our experimental results are shown in Table 11.10.

Here, we let X_{ij} denote the sales (in units) of Snoopy during the i th day under the j th treatment level. If we look at mean sales by each level of shelf height, it appears as though the waist-level treatment, the average response to which is $X_2 = 90.9$, results in the highest mean

218 • ANALYSIS AND MODELING

Table 11.10 Sales of Snoopy Dog Food (in Units) by Level of Shelf Height

Shelf Height							
Knee Level		Waist Level		Eye Level		Grand Total	
X_{11}	77	X_{12}	88	X_{13}	85		
X_{21}	82	X_{22}	94	X_{23}	85		
X_{31}	86	X_{32}	93	X_{33}	87		
X_{41}	78	X_{42}	90	X_{43}	81		
X_{51}	81	X_{52}	91	X_{53}	80		
X_{61}	86	X_{62}	94	X_{63}	79		
X_{71}	77	X_{72}	90	X_{73}	87		
X_{81}	81	X_{82}	87	X_{83}	93		
$X_{T1} = 648$		$X_{T2} = 727$		$X_{T3} = 677$		$X_{TT} = 2,052$	
$\bar{X}_1 = 81.0$		$= \bar{X}_2 = 90.9$		$\bar{X}_3 = 84.6$		$= \bar{X}_{TT} = 85.5$	

Table 11.11 Analysis of Variance: Snoopy Dog Food Experiment

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F Ratio
Among treatments	$t - 1 = 2$	399.3	199.7	14.6 ($p < .01$)
Within treatments	$n - t = 21$	288.7	13.7	
Total	$n - 1 = 23$	688.0	$MS_T = SS_T/df_T$	
			$MS_w = SS_w/df_w$	$F = \frac{MS_T}{MS_w}$
Correction factor	$C = \frac{(X_{TT})^2}{n} = \frac{(2,052)^2}{24} = 175,446.0$			
Total sum of squares	$\sum x_{ij}^2 - C = (77)^2 + (82)^2 + \dots + (87)^2 + (93)^2 + -175,466.0 = 688.0$			
Treatment sum of squares	$\frac{\sum x_{Tj}^2}{n_j} - C = \frac{(648)^2 + (727)^2 + (677)^2}{8} - 175,466.0 = 399.3$			
Within-treatment sum of squares	$\sum x_{ij}^2 \frac{\sum x_{Tj}^2}{n_j} = (77)^2 + (82)^2 + \dots + (87)^2 + (93)^2 - \frac{(648)^2 + (727)^2 + (677)^2}{8} = 288.7$			

sales over the experimental period. However, we note that the last observation (93) under the eye-level treatment exceeds the waist-level treatment mean. Is this a fluke observation? We know that these means are, after all, sample means, and our interest lies in whether the three

population means from which the samples are drawn are equal.

Now we shall show what happens when one goes through a typical one-way analysis of variance computation for this problem. These calculations are shown in Table 11.11.

Table 11.11 shows the mechanics of developing the among-treatments, within-treatments, and total sums of squares, the mean squares, and the F ratio. Had the experimenter used an alpha risk of 0.01, the null hypothesis of no differences among treatment levels would have been rejected. A table of F ratios is found in most statistics and research texts.

Note that Table 11.11 shows shortcut procedures for finding each sum of squares. For example, the total sum of squares is given by

$$\sum X_{ij}^2 - \frac{(X_{TT})^2}{n} = 688.0.$$

This is the same quantity that would be obtained by subtracting the grand mean of 85.5 from each original observation, squaring the result, and adding up the 24 squared deviations. This mean-corrected sum of squares is equivalent to the type of formula used earlier in this chapter.

The interpretation of this analysis is, like the t test, a process of comparing the F value of 9.6 ($df = 2, 21$) with the table value of 4.32 ($p = 0.05$). Because ($9.6 > 4.32$), we reject the null hypothesis that Treatments 1, 2, and 3 have equivalent appeals.

The important consideration to remember is that, aside from the statistical assumptions underlying the analysis of variance, the variance of the error distribution will markedly influence the significance of the results. That is, if the variance is large relative to differences among treatments, then the true effects may be swamped, leading to an acceptance of the null hypothesis when it is false. As we know, an increase in sample size can reduce this experimental error. Though beyond the scope of this chapter, specialized experimental designs are available, the objectives of which are to increase the efficiency of the experiment by reducing the error variance.

Follow-Up Tests of Treatment Differences

The question that now must be answered is the following: Which treatments differ? The F ratio only provides information that differences exist. The question of where differences exist is answered by follow-up analysis, usually a series

of independent sample t tests, to compare the treatment level combinations ((1, 2), (1, 3), and (2, 3)). Because of our previous discussion of the t test, we will not discuss these tests in detail. We will only allude to the fact that various forms of the t statistic may be used when conducting a series of two group tests. These test statistics (which include techniques known as the least significant difference, Bonferroni's test, Duncan's multiple-range tests, Scheffé's test, and others) control the probability that a Type I error will occur when a series of statistical tests are conducted. Recall that if in a series of statistical tests, each test has a .05 probability of a Type I error, then in a series of 20 such tests, we would expect that one ($20 \times .05 = 1$) of these tests would report a significant difference that did not exist (Type I error). These tests typically are options provided by the standard statistical packages, such as the SPSS program Oneway.

BIVARIATE ANALYSIS: MEASURES OF ASSOCIATION

Bivariate measures of association include the two-variable case in which both variables are interval or ratio scaled. Our concern is with the nature of the associations between the two variables and the use of these associations in making predictions.

Correlation Analysis

When referring to a simple two-variable correlation, we refer to the strength and direction of the relationship between the two variables. As an initial step in studying the relationship between the X and Y variables, it is often helpful to graph this relationship in a scatter diagram (also known as an X - Y plot). Each point on the graph represents the appropriate combination of scale values for the associated X and Y variables, as shown in Figure 11.6. The values of the correlation coefficient may range from +1 to -1. These extreme values indicate perfect positive and negative linear correlations. Other relationships may appear to be curvilinear or even random plots and have coefficients between zero and the extreme values.

220 • ANALYSIS AND MODELING

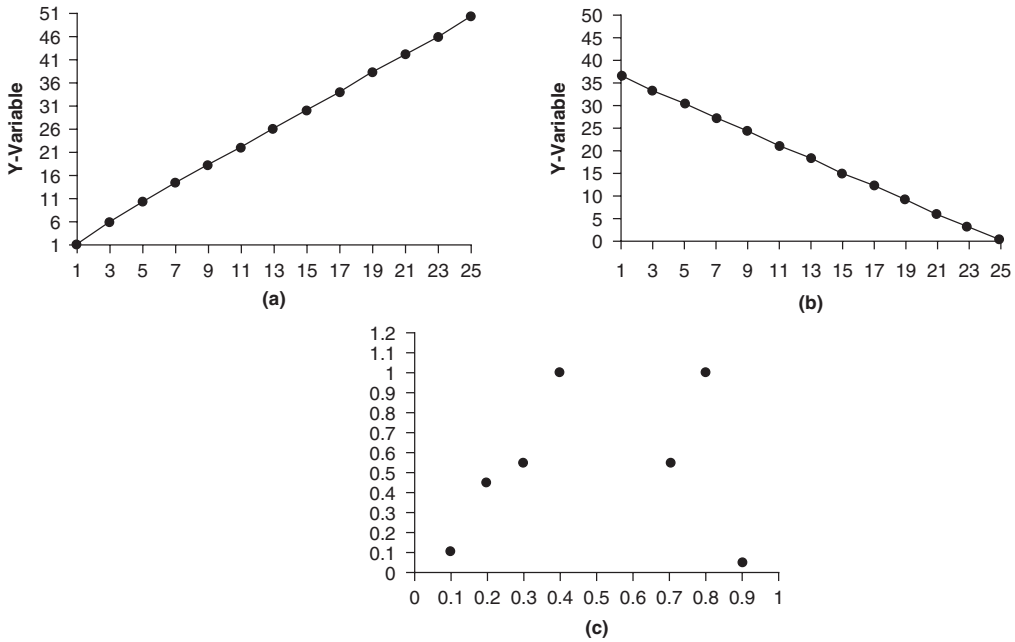


Figure 11.6 Scatter Diagrams

The objective of correlation analysis, then, is to obtain a measure of the degree of linear association (correlation) that exists between the two variables. The Pearson correlation coefficient is commonly used for this purpose and is defined by the formula

$$\rho_{XY} = \frac{1}{n} \sum_{i=1}^n \frac{Y_i - \bar{Y}}{S_Y} \frac{X_i - \bar{X}}{S_X} = \sum_{i=1}^n \frac{Z_X Z_Y}{n},$$

where n pairs of (X_p, Y_p) values provide a sample size n , and X , Y , S_X , and S_Y represent the sample means and sample standard deviations of the X and Y variables.

The alternate formulation shows the correlation coefficient to be the product of the Z scores for the X and Y variables. In this method of computing the correlation coefficient, the first step is to convert the raw data to a Z score by finding the deviation from the respective sample mean. The Z scores will be centered as a normally distributed variable (mean of zero and standard deviation of 1).

The transformation of the X and Y variables to Z scores means that the scale measuring the original variable is no longer relevant, as a

Z score variable originally measured in dollars can be correlated with another Z score variable originally measured on a satisfaction scale. The original metric scales are replaced by a new abstract scale (called correlation) that is the product of the two Z distributions.

By continuing our digression one step further, we can show how the correlation coefficient becomes positive or negative.

$$\rho_{XY} = \frac{\sum_{i=1}^n Z_X Z_Y}{n},$$

where

$$Z_Y = \frac{Y_i - \bar{Y}}{S_Y} \quad Z_X = \frac{X_i - \bar{X}}{S_X}$$

We know that the Z_X and Z_Y values will generally fall in the range -3 to $+3$. When both Z_X and Z_Y are positive or both are negative, r_{XY} is positive, as shown in Figure 11.7. When Z_X is positive and Z_Y negative (or the opposite), a negative correlation will exist. Of course, we are talking of individual pairs of the Z_X and Z_Y

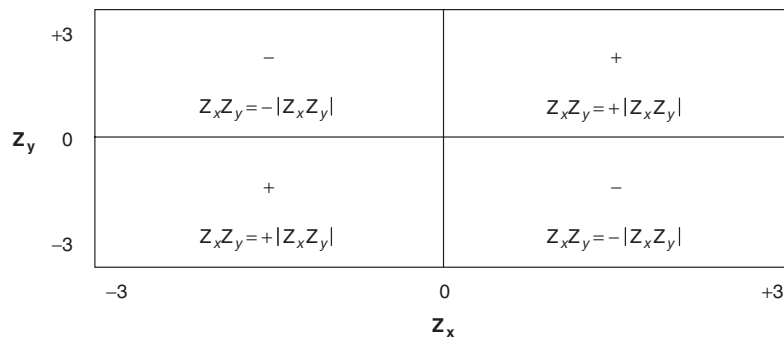


Figure 11.7 Bivariate Products of Standard Z Scores

X	39	43	21	64	57	47	28	75	34	52
Y	68	82	56	86	97	94	77	103	59	79

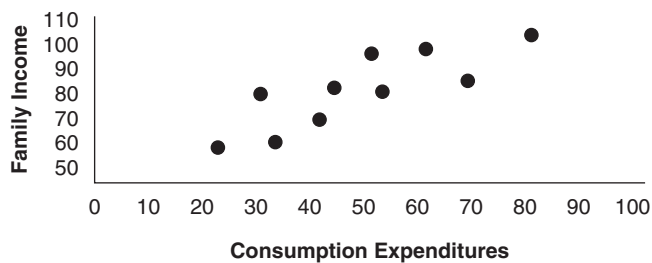


Figure 11.8 Consumption Expenditure and Income Data

variables, which, when summed, produce the overall correlation coefficient.

To summarize, the sign (+ or -) indicates the direction of the correlation, and the absolute value of the coefficient indicates the degree of correlation (from 0 to 1). Thus, correlations of +.7 and -.7 are of exactly the same strength, but the relationships are in opposite directions.

We are cautioned that near-perfect positive or negative linear correlations do not mean causality. Often, other factors underlie and are even responsible for the relationship between the variables. For example, at one time, Kraft Foods reported that sales of its macaroni and cheese product were highly correlated (negatively) with national indices of the health of the economy. We may not imply that Kraft sales directly caused fluctuations in the national economy or vice versa. Consumer expectations and possibly

personal income vary as a function of the national economy. In times of reduced family income, macaroni and cheese is a low-price dietary substitute for more expensive meals.

To demonstrate, we will consider a brief example of a correlation analysis that examines the relationships between (a) family income and (b) family consumption expenditures. The data and plot appear in Figure 11.8.

To calculate the correlation coefficient, we will reduce the previous formula to the basic computation consisting of sums for the X and Y variables. This equation looks formidable but allows for easy computation by simply entering the appropriate summation values from the bottom of Table 11.12.

Needless to say, computations such as this are rarely done today. Researchers routinely perform their analyses by using Excel spreadsheets

222 • ANALYSIS AND MODELING

Table 11.12 Family Income and Family Consumption Expenditures

<i>Respondent</i>	<i>Y</i>	<i>X</i>	<i>XY</i>	<i>Y</i> ²	<i>X</i> ²
1	68	39	2,652	4,624	1,521
2	82	43	3,526	6,724	1,849
3	56	21	1,176	3,136	441
4	86	64	5,504	7,396	4,096
5	97	57	5,529	9,409	3,249
6	94	47	4,418	8,836	2,209
7	77	28	2,156	15,929	784
8	103	75	7,725	609	5,625
9	59	34	2,006	3,481	1,156
10	79	52	4,108	6,241	2,704
Sum	801	460	38,800	66,385	23,634
Average	80.1	46	3,880	6,638.5	2,363.4

or computer packages such as SPSS. However, our brief discussion is included to provide understanding of underlying processes.

Introduction to Bivariate Regression Analysis

In the analysis of associative data, the marketing researcher is almost always interested in problems of predictions:

- Can we predict a person's weekly fast-food and restaurant food purchases from that person's gender, age, income, or education level?
- Can we predict the dollar volume of purchase of our new product by industrial purchasing agents as a function of our relative price, delivery schedules, product quality, and technical service?

The list of such problems is almost endless. Not surprisingly, the linear regression model—as applied in either the bivariate (single predictor) or multivariate form (multiple predictors)—is one of the most popular methods in the marketing researcher's tool kit. The bivariate form is also known as simple regression.

The regression model has been applied to problems ranging from estimating sales quotas to predicting demand for new shopping centers. As an illustration, one of the leading ski resorts in the United States used a regression model

to predict the weekend ticket sales, based on variables including the following:

- Highway driving conditions
- Average temperature in the 3-day period preceding the weekend
- Local weather forecast for the weekend
- Amount of newspaper space devoted to the resort's advertisements in the surrounding city newspapers
- A moving average of the 3 preceding weekends' ticket sales

The model's accuracy was within $\pm 6\%$ of actual attendance throughout the season.

One firm used a regression model to predict physicians' readership of various medical journals, based on physician ratings of several attributes of each journal:

- Writing style
- Quality of illustrations
- Informativeness of advertisements
- Relevance to physician needs
- Authoritativeness in the medical profession
- Frequency of issue

The model predicted actual readership in future time periods quite well. Moreover, it provided diagnostic information as to how various journals' editorial policies and advertising could be improved.

A bakery explored the use of a regression model to predict sales of hamburger buns as a guide to production policy. The following factors were included as independent variables:

- The weather
- The day of month
- The proximity to a holiday

The model was able to predict well enough to reduce the average returned goods by 4 percentage points (from 10.4% to 6.4%).

Regression analysis, in its simplest bivariate form, involves a single dependent (criterion) variable and a single independent (predictor) variable. In its more advanced multiple regression form, a set of predictors is used to form an additive linear combination that predicts the single dependent variable. In considering the use of either simple or multiple regression, the researcher is interested in four main questions:

1. Can we find a predictor variable (a linear composite of the predictor variables in the multiple case) that will compactly express the relationship between a criterion variable and the predictor (set of predictors)?
2. If we can, how strong is the relationship; that is, how well can we predict values of the criterion variable from values of the predictor (linear composite)?
3. Is the overall relationship statistically significant?
4. Which predictor is most important in accounting for variation in the criterion variable? (Can the original model be reduced to fewer variables but still provide adequate prediction of the criterion?)

Regression analysis is discussed in detail in Chapters 13 and 14 of this book.

Rank Correlation

Our discussion in this chapter is based on the premise that the dependent variable is at least interval scaled or can be treated as such with little error. There are marketing problems, however, where the dependent and independent

Table 11.13 Rankings on Two Methods of Salesperson Evaluation Rank

Salesperson	Performance Index (X)	New Method (Y)	d	
			d_i	d
A	8	6	2	4
B	4	7	-3	9
C	1	22	-1	1
D	6	3	3	9
E	2	1	1	1
F	10	8	2	4
G	5	5	0	0
H	3	9	-6	36
I	7	4	3	9
J	9	10	1	1

$\sum d_i^2 = 74$

variables are rank orders or are best transformed into such rankings. In this situation, rank correlation techniques can be used to estimate the association between sets of data.

For two variables, the best-known and easiest technique is that involving the use of the Spearman rank correlation coefficient, r_s . We show the use of this measure by an example. Suppose a sales manager evaluates salespersons by two different methods (performance index and a new method). Since the new method is easier to use, the manager wants to know if it will yield the same relative results as the proven existing method. The scores have been transformed into rankings so that each salesperson has two rankings. Table 11.13 shows the rankings.

To measure the extent of rank correlation, we use the statistic

$$r_s = 1 - \frac{6 \sum_{i=1}^N d^2}{N(N^2 - 1)},$$

where N is the number of pairs of ranks, and d is the difference between the two rankings for an individual (i.e., $X - Y$). Applying this formula to our example, we get

224 • ANALYSIS AND MODELING

$$r_s = 1 - \frac{6(74)}{10(100 - 1)} = .55.$$

If the subjects whose scores were used in computing r_s were randomly drawn from a population, we can test the significance of the obtained value. The null hypothesis is that the two variables are not associated, and thus the true value of ρ is zero. Under H_0 , any observed value would be due to chance. When $N \geq 10$, significance can be tested using the statistic

$$t = r_s \sqrt{\frac{N - 2}{1 - r_s^2}},$$

which is interpreted from a table of t values with $(N - 2)$ degrees of freedom. For our example, we calculate

$$t = .55 \sqrt{\frac{10 - 2}{1 - (.55)^2}} = 1.870.$$

Looking at the table of critical values of t , we find that $p \geq .10$ (two-tailed test) for $(10 - 2 = 8)$ degrees of freedom. Thus, if a strict α level is to be adhered to (.10 or less), we tentatively accept H_0 and conclude that it is unlikely that a correlation exists between the scores from the two evaluation methods.

One final point concerning the use of the Spearman rank correlation coefficient is warranted. At times, tied observations will exist. When this happens, each of them is assigned the average of the ranks that would have been assigned in the absence of ties. If the proportion of ties is small, the effect on r_s is minute. If large, however, a correction factor must be applied (see Siegel, 1956, pp. 206–210).

Another measure that gives comparable results is the Kendall rank correlation coefficient, tau. The value of tau ranges between -1 and $+1$. It measures association in a different way from the Spearman rank correlation. Tau measures the association between X and Y as the proportion of concordant pairs minus the proportion of discordant pairs in the samples. Two bivariate observations, (X_p, Y_p) and (X_j, Y_j) , are called concordant whenever the product of

the difference between each pair of observations $(X_i - X_j)(Y_i - Y_j)$ is positive, and a pair is called discordant when this product is negative (Gibbons, 1993, p. 11).

In general, the Spearman measure is used more widely than the Kendall measure. It is an easier measure to use. In addition, the Spearman rank correlation coefficient is equivalent to the Pearson product-moment correlation coefficient, with ranks substituted for the measurement observations, X and Y . Both the Spearman and Kendall measures are discussed more fully by Siegel (1956, chap. 9) and Gibbons (1993). These two measures are used when there are two sets of rankings to be analyzed. When there are three or more sets of rankings, Kendall's coefficient of concordance should be used (Gibbons, 1993; Siegel, 1956, chap. 9).

Finally, when the variables are nominally scaled, ordinal measures such as tau and r_s are not appropriate measures. Nominal variables lack the ordering property. One measure that can be used is Goodman and Kruskal's lambda (λ). Lambda is a measure of association whose calculation and interpretation are straightforward. Lambda tells us how much we can reduce our error in predicting Y once we know X and is shown as

$$\lambda = \frac{\text{reduction in prediction errors knowing } X}{\text{prediction errors in not knowing}}$$

This measure, as well as others, is discussed by Lewis-Beck (1995, chap. 4). Lambda is an option provided by the SPSS Crosstab program. We discuss lambda further in the next section.

NONPARAMETRIC ANALYSIS

Other Tests

One reason for the widespread use of chi-square in cross-tabulation analysis is that most computer computational routines show the statistic as part of the output, or at least it is an option that the analyst can choose. Sometimes, the data available are stronger than simple nominal measurement and are ordinal. In this situation, other tests are more powerful than chi-square. Three regularly used tests are

the Wilcoxon rank sum (T), the Mann-Whitney U , and the Kolmogorov-Smirnov test. Siegel (1956) and Gibbons (1993) provide more detailed discussions of these techniques.

The Wilcoxon T test is used for dependent samples in which the data are collected in matched pairs. This test takes into account both the direction of differences within pairs of observations and the relative magnitude of the differences. The Wilcoxon matched-pairs signed-ranks test gives more weight to pairs showing large differences between the two measurements than to a pair showing a small difference. To use this test, measurements must at least be ordinally scaled within pairs. In addition, ordinal measurement must hold for the differences between pairs.

This test has many practical applications in marketing research. For instance, it may be used to test whether a promotional campaign has had an effect on attitudes. An ordinal scaling device, such as a semantic differential, can be used to measure attitudes toward, say, a bank. Then, after a special promotional campaign, the same sample would be given the same scaling device. Changes in values of each scale could be analyzed by this Wilcoxon test. With ordinal measurement and two independent samples, the Mann-Whitney U test may be used to test whether the two groups are from the same population. This is a relatively powerful non-parametric test, and it is an alternative to the Student t test when the analyst cannot meet the assumptions of the t test or when measurement is at best ordinal. Both one- and two-tailed tests can be conducted. As indicated earlier, results of U and t tests often are similar, leading to the same conclusion.

The Kolmogorov-Smirnov two-sample test is a test of whether two independent samples come from the same population or from populations with the same distribution. This test is sensitive to any kind of difference in the distributions from which the two samples were drawn—differences in location (central tendency), dispersion, skewness, and so on. This characteristic of the test makes it a very versatile test. Unfortunately, the test does not by itself show what kind of difference exists. There is a Kolmogorov-Smirnov one-sample test that is concerned with the agreement

between an observed distribution of a set of sample values and some specified theoretical distribution. In this case, it is a goodness-of-fit test similar to single-classification chi-square analysis.

Indexes of Agreement

Chi-square is appropriate for making statistical tests of independence in cross-tabulations. Usually, however, we are interested in the strength of association as well as the statistical significance of association. This concern is for what is known as substantive or practical significance. An association is substantively significant when it is statistically significant and of sufficient strength. Unlike statistical significance, however, there is no simple numerical value to compare with, and considerable research judgment is necessary. Although such judgment is subjective, it need not be completely arbitrary. The nature of the problem can offer some basis for judgment, and common sense can indicate that the degree of association is too low in some cases and high enough in others (Gold, 1969, p. 44).

Statisticians have devised a plethora of indexes—often called indexes of agreement—for measuring the strength of association between two variables in a cross-tabulation. The main descriptors for classifying the various indexes are as follows:

1. Whether the table is 2×2 or larger, $R \times C$
2. Whether one, both, or neither of the variables has categories that obey some natural order (e.g., age, income level, family size)
3. Whether association is to be treated symmetrically or whether we want to predict membership in one variable's categories from (assumed known) membership in the other variable's categories

Space does not permit coverage of even an appreciable fraction of the dozens of agreement indexes that have been proposed. Rather, we shall illustrate one commonly used index for 2×2 tables and two indexes that deal with different aspects of the larger $R \times C$ (row-by-column) tables.

226 • ANALYSIS AND MODELING

Table 11.14 Does Hair Have Enough Body Versus Body Inclusion in Ideal Set

	<i>Hair Have Enough Body?</i>		<i>Total</i>
	<i>No</i>	<i>Yes</i>	
Body included in ideal set	26 (A)	8 (B)	34
Body excluded from ideal set	17 (C)	33 (D)	50
Total	43	41	84

The 2 × 2 Case

The phi correlation coefficient is a useful agreement index for the special case of 2×2 tables in which both variables are dichotomous. Moreover, an added bonus is the fact that phi equals the product-moment correlation—a cornerstone of multivariate methods—that one would obtain if he or she correlated the two variables expressed in coded 0–1 form.

To illustrate, consider the 2×2 cross-tabulation in Table 11.14, taken from a study relating to shampoos. We wish to see if inclusion of the shampoo benefit of “body” in the respondent’s ideal set is associated with the respondent’s indication that her hair lacks natural “body.” We first note from the table that high frequencies appear in the cells: (a) “body” included in the ideal set and “no” to the question of whether her hair has enough (natural) body and (b) “body” excluded from the ideal set and “yes” to the same question.

Before computing the phi coefficient, first note the labels *A*, *B*, *C*, and *D* assigned to the four cells in Table 11.14. The phi coefficient is defined as

$$\begin{aligned}\phi &= \frac{AD - BC}{\sqrt{(A + B)(C + D)(A + C)(B + D)}} \\ &= \frac{26(33) - 8(17)}{\sqrt{(26 + 8)(17 + 33)(26 + 17)(8 + 33)}} \\ &= 0.417\end{aligned}$$

The value 0.417 is also what would be found if an ordinary product-moment correlation, to be described later in the chapter, is computed across the 84 pairs of numbers where the

following code values are used to identify the responses:

Body included in ideal set \Rightarrow 1

Body excluded from ideal set \Rightarrow 0

Hair have enough body? No \Rightarrow 1

Yes \Rightarrow 0

This is a nice feature of phi in the sense that standard computer programs for calculating product-moment correlations can be used for dichotomous variables.

The phi coefficient can vary from -1 to 1 (just like the ordinary product-moment correlation). However, in any given problem, the upper limit of phi depends on the relationships among the marginals. Specifically, a phi coefficient of -1 (perfect negative association) or 1 (perfect positive association) assumes that the marginal totals of the first variable are identical to those of the second. Looking at the letters (*A*, *B*, *C*, *D*) of Table 11.14, assume that the row marginals equaled the column marginals: then, $\Phi = 1$ if $B = C = 0$; similarly, $\Phi = -1$ if $A = D = 0$. The more different the marginals, the lower the upper limit that the (absolute) value of phi can assume.

The phi coefficient assumes the value of zero if the two variables are statistically independent (as would be shown by a chi-square value that is also zero). Indeed, the absolute value of phi is related to chi-square by the expression

$$\phi = \sqrt{\frac{\chi^2}{n}} 1,$$

where n is the total frequency (sample size). This is a nice feature of phi, in the sense that it can be computed quite easily after chi-square has been computed. Note, however, that phi, unlike chi-square, is not affected by total sample size because we have the divisor n in the above formula to adjust for differences in sample size.

The R × C Case

One of the most popular agreement indexes for summarizing the degree of association between two variables in a cross-tabulation of R rows and C columns is the contingency

coefficient. This index is also related to chi-square and is defined as

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

where n is again the total sample size. From Table 11.14, we can first determine that chi-square is equal to 14.61, which, with 1 degree of freedom, is significant beyond the .01 level.

We can then find the contingency coefficient C as the following:

$$\begin{aligned} C &= \sqrt{\frac{14.61}{14.61 + 84}} \\ &= 0.385 \end{aligned}$$

As may be surmised, the contingency coefficient lies between 0 and 1, with 0 reserved for the case of statistical independence (a chi-square value of 0). However, unlike the phi coefficient, the contingency can never attain a maximum value of unity. For example, in a 2×2 table, C cannot exceed 0.707. As might be noticed by the reader, there is an algebraic relationship between phi and the contingency coefficient (if the latter is applied to the 2×2 table):

$$\phi^2 = \frac{C^2}{1 - C^2}.$$

In a 4×4 table, its upper limit is 0.87. Therefore, contingency coefficients computed from different-sized tables are not easily comparable.

However, like phi, the contingency coefficient is easy to compute from chi-square; moreover, like phi, its significance has already been tested in the course of running the chi-square test.

Both phi and the contingency coefficient are symmetric measures of association. Occasions often arise in the analysis of $R \times C$ tables (or the special case of 2×2 tables) where we desire to compute an asymmetric measure of the extent to which we can reduce errors in predicting categories of one variable from knowledge of the categories of some other variable. Goodman and Kruskal's lambda-asymmetric coefficient can be used for this purpose (Goodman & Kruskal, 1954).

To illustrate the lambda-asymmetric coefficient, let us return to the cross-tabulation of Table 11.14. Suppose that we wished to predict what category—no versus yes—a randomly selected person would fall into when asked, "Does your hair have enough body?" If we had no knowledge of the row variable (whether that person included "body" in her ideal set or not), we would have only the column marginal frequencies to rely on.

Our best bet, given no knowledge of the row variable, is always to predict no, the higher of the column marginal frequencies. As a consequence, we shall be wrong in 41 of the 84 cases, a probability error of $41/84 = 0.49$. Can we do better, in the sense of lower prediction errors, if we use information provided by the row variable?

If we know that "body" is included in the ideal set, we shall predict no and be wrong in only 8 cases. If we know that "body" is not included in the ideal set, we shall predict yes and be wrong in 17 cases. Therefore, we have reduced our number of prediction errors from 41 to $8 + 17 = 25$, a decrease of 16 errors. We can consider this error reduction relatively:

$$\begin{aligned} \lambda_{C|R} &= \frac{\text{(number of errors in first case)} - \text{numbers of errors in second case}}{\text{numbers of errors in first case}} \\ &= \frac{41 - 25}{41} = 0.39 \end{aligned}$$

In other words, 39% of the errors in predicting the column variable are eliminated by knowing the individual's row variable.

A less cumbersome (but also less transparent) formula for lambda-asymmetric is

$$\lambda_{C|R} = \frac{\sum_{k=1}^K f_{kR}^{k^*} - F_c^*}{n - F_c^*} = \frac{(26 + 33) - 43}{84 - 43} = 0.39,$$

where $f_{kR}^{k^*}$ is the maximum frequency found within each subclass of the row variable, F_c^* is the maximum frequency among the marginal totals of the column variable, and n is the total number of cases.

228 • ANALYSIS AND MODELING

Lambda-asymmetric varies between 0, indicating no ability at all to eliminate errors in predicting the column variable on the basis of the row variable, and 1, indicating an ability to eliminate all errors in the column variable predictions, given knowledge of the row variable. Not surprisingly, we could reverse the role of criterion and predictor variables and find lambda-asymmetric for the row variable, given the column variable. In the case of Table 11.14, this results in $\lambda = 0.26$. Note that in this case, we simply reverse the roles of row and column variables.

Finally, if desired, we could find a lambda-symmetric index via a weighted averaging of $\lambda_{C|R}$ and $\lambda_{R|C}$. However, in our opinion, lambda-asymmetric is of particular usefulness to the analysis of cross-tabulations because we often want to consider one variable as a predictor and the other as a criterion. Furthermore, lambda-asymmetric has a natural and useful interpretation as the percentage of total prediction errors that are eliminated in predicting one variable (e.g., the column variable) from another (e.g., the row variable).

A Concluding Comment

The indices discussed previously are optional choices when using most statistical analysis programs such as SPSS's Crosstabs, which is within the Descriptive Statistics module. That is, the researcher indicates which one(s) to include in the analysis.

SUMMARY

We began by stating that data can be viewed as recorded information useful in making decisions. In the initial sections of this chapter, we introduced the basic concepts of transforming raw data into data of quality. The introduction was followed by a discussion of elementary descriptive analyses through tabulation and cross-tabulation. The focus of this discussion was heavily oriented toward how to read the data and how to interpret the results. The competent analysis of research-obtained data requires a blending of art and science, of intuition and informal insight, and of judgment and statistical treatment, combined with a thorough

knowledge of the context of the problem being investigated.

We next focused on the necessary statistical machinery to analyze differences between groups: *t* test and one-factor and two-factor analysis of variance. These techniques are useful for both experimental and nonexperimentally obtained data. The first section of the chapter dealt with cross-tabulation and chi-square analysis. This was followed by discussing bivariate analysis of differences in means and proportions. We then looked at the process of analysis of variance. A simple numerical example was used to demonstrate the partitioning of variance into among- and within-components. The assumptions underlying various models were pointed out, and a hypothetical data experiment was analyzed to show how the ANOVA models operate. The topic of interaction plotting was introduced to aid the researcher in interpreting the results of the analysis.

We concluded by examining bivariate analyses of associations for interval- or ratio-scaled data. The concept of associations between two variables was introduced through simple two-variable correlation. We examined the strength and direction of relationships using the scatter diagram and Pearson correlation coefficient. Several alternative (but equivalent) mathematical expressions were presented, and a correlation coefficient was computed for a sample data set.

Investigations of the relationships between variables almost always involve the making of predictions. Bivariate (two-variable) regression was introduced.

We ended the chapter with a discussion of the Spearman rank correlation and Kendall tau as alternatives to the Pearson correlation coefficient when the data are of ordinal measurement and do not meet the assumptions of parametric methods. Also, the Goodman and Kruskal lambda measure for nominal measurement was briefly introduced, as were other nonparametric analyses.

A wide array of statistical techniques (parametric and nonparametric) focuses on describing and making inferences about the variables being analyzed. Some of these were shown in Table 11.6. Although somewhat dated, a useful reference for selecting an appropriate

statistical technique is the guide published by the Institute for Social Research at the University of Michigan (Andrews, Klem, Davidson, O'Malley, & Rodgers, 1981, and its corresponding software Statistical Consultant). Fink (2003, pp. 78–80) presents a summary table of which technique to use under which condition.

REFERENCES

- Andrews, F. M., Klem, L., Davidson, T. N., O'Malley, P. M., & Rodgers, W. L. (1981). *A guide for selecting statistical techniques for analyzing social science data* (2nd ed.). Ann Arbor: Institute for Social Research, University of Michigan.
- Feick, L. F. (1984). Analyzing marketing research data with associated models. *Journal of Marketing Research*, 21, 376–386.
- Fink, A. (2003). *How to manage, analyze, and interpret survey data*. Thousand Oaks, CA: Sage.
- Gibbons, J. D. (1993). *Nonparametric statistics: An introduction*. Newbury Park, CA: Sage.
- Gold, D. (1969). Statistical tests and substantive significance. *American Sociologist*, 4, 44.
- Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classification. *Journal of the American Statistical Association*, 49, 732–764.
- Hellevik, O. (1984). *Introduction to causal analysis: Exploring survey data*. Beverly Hills, CA: Sage.
- Lewis-Beck, M. S. (1995). *Data analysis: An introduction*. Thousand Oaks, CA: Sage.
- Semon, T. T. (1999). Use your brain when using a chi-square. *Marketing News*, 33, 6.
- Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.
- Smith, S., & Albaum, G. (2005). *Fundamentals of marketing research*. Thousand Oaks, CA: Sage.
- Zeisel, H. (1957). *Say it with figures* (4th ed.). New York: Harper & Row.