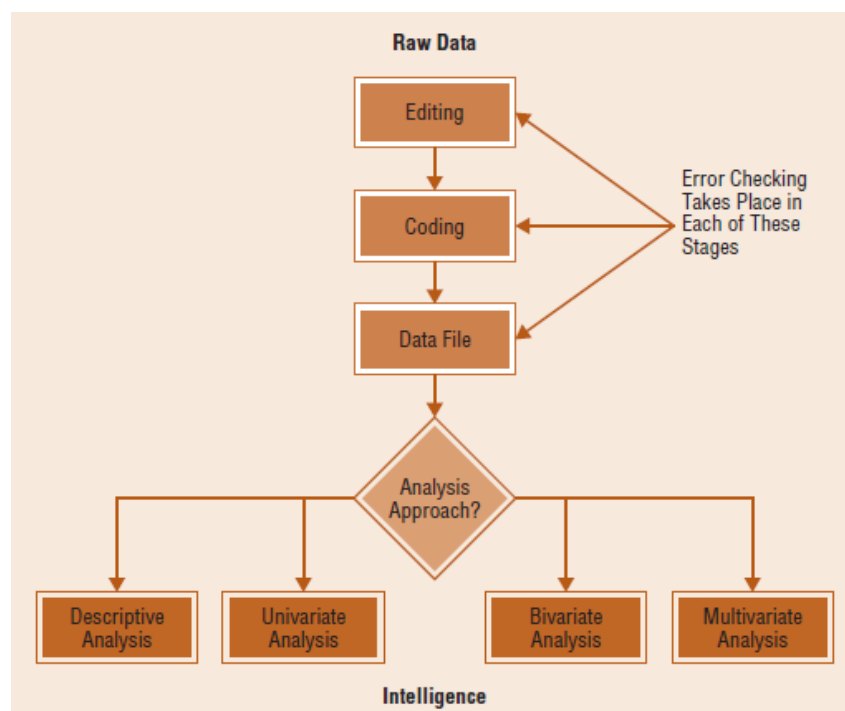


## EDITING AND CODING: TRANSFORMING RAW DATA INTO INFORMATION

### Stages of Data Analysis

Practically all researchers will be very anxious to begin data analysis once the field work is complete. Now, the raw data can be transformed into intelligence. However, **raw data** may not be in a form that lends itself well to analysis. Raw data are recorded just as the respondent indicated. For an oral response, the raw data are in the words of the respondent, whereas for a questionnaire response, the actual number checked is the number stored. Raw data will often also contain errors both in the form of respondent errors and nonrespondent errors. Whereas a respondent error is a mistake made by the respondent, a **nonrespondent error** is a mistake made by an interviewer or by a person responsible for creating an electronic data file representing the responses.

Below figure provides an overview of data analysis. The first two stages result in an electronic file suitable for data analysis. This file can then be used in the application of various statistical routines including those associated with descriptive, univariate, bivariate, or multivariate analysis. Each of these data analysis approaches will be discussed in the subsequent chapters. An important part of the editing, coding, and filing stages is checking for errors. As long as error remains in the data, the process of transformation from raw data to intelligence will be made more risky and more difficult. Editing and coding are the first two stages in the data analysis process.



Overview of the Stages of Data Analysis

**Data integrity** refers to the notion that the data file actually contains the information that the researcher promised the decision maker he or she would obtain. Additionally, data integrity extends to the fact that the data have been edited and properly coded so that they are useful to the decision maker. Any errors in this process, just as with errors or shortcuts in the interview process itself, harm the integrity of the data.

## Editing

Fieldwork often produces data containing mistakes. For example, consider the following simple questionnaire item and response:

*How long have you lived at your current address? 48*

The researcher had intended the response to be in years. Perhaps the respondent has indicated the number of months rather than years he or she has lived at this address? Alternatively, if this was an interviewer's form, he or she may have marked the response in months without indicating this on the form. How should this be treated? Sometimes, responses may be contradictory. What if the same respondent above gives this response?

*What is your age? 32 years*

This answer contradicts the earlier response. If the respondent is 32 years of age, then how could he or she have lived at the same address for 48 years? Therefore, an adjustment should be made to accommodate this information. The most likely case is that this respondent has lived at the current address for four years.

This example illustrates data **editing**. Editing is the process of checking and adjusting data for omissions, consistency, and legibility. In this way, the data become ready for analysis by a computer. So, the editor's task is to check for errors and omissions on questionnaires or other data collection forms. When the editor discovers a problem, he or she adjusts the data to make them more complete, consistent, or readable.

At times, the editor may need to reconstruct data. In the example above, the researcher can guess with some certainty that the respondent entered the original questions in months instead of years. Therefore, the probable true answer can be reconstructed. While the editor should try to make adjustments in an effort to represent as much information from a respondent as possible, reconstructing responses in this fashion should be done only when the probable true response is very obvious. Had the respondent's age been 55 years, filling in the response with years would not have been advisable barring other information. Perhaps the respondent has lived in the house since childhood? That possibility would seem real enough to prevent changing the response.

## Field Editing

Field supervisors often are responsible for conducting preliminary **field editing** on the same day as the interview. Field editing is used to

1. Identify technical omissions such as a blank page on an interview form
2. Check legibility of handwriting for open-ended responses
3. Clarify responses that are logically or conceptually inconsistent.

Field editing is particularly useful when personal interviews have been used to gather data. In these cases, a daily field edit allows supervisors to deal with some questions by asking interviewers, who may still be able to remember the interviews, about facts that may allow errors to be identified and perhaps corrected. In addition, the number of unanswered questions or incomplete responses can be reduced with rapid follow up. A daily field edit allows fieldworkers to identify respondents who should be recontacted to fill in omissions in a timely fashion.

The supervisor may also use field edits to spot the need for further interviewer training or to correct faulty procedures. For example, if an interviewer did not correctly follow skip patterns, training may be indicated. The supervisor may also notice that an interviewer is not properly probing some open-ended responses.

## In-House Editing

Although simultaneous field editing is highly desirable, in many situations (particularly with mail questionnaires) early reviewing of the data is not always possible. **In-house editing** rigorously investigates the results of data collection. The research supplier or research department normally has a centralized office staff perform the editing and coding function.

For example, Arbitron measures radio audiences by having respondents record their listening behavior—time, station, and place (home or car)—in diaries. After the diaries are returned by mail, in-house editors perform usability edits in which they check that the postmark is after the last day of the survey week, verify the legibility of station call letters (station WKXY could look like KWCY), look for completeness of entries on each day of the week, and perform other editing activities. If the respondent's age or sex is not indicated, the respondent is called to ensure that this information is included.

### ■ ILLUSTRATING INCONSISTENCY-FACT OR FICTION?

Consider a situation in which a telephone interviewer has been instructed to interview only registered voters in a state that requires voters to be at least 18 years old. If the editor's review of a questionnaire indicates that the respondent was only 17 years old, the editor's task is to correct this mistake by deleting this response because this respondent should never have been considered as a sampling unit. The sampling units (respondents) should all be consistent with the defined population. The editor also should check for consistency within the data collection framework. For example, a survey on out-shopping behavior (shopping in towns other than the one in which the person resides) might have a question such as the following:

In which of the following cities have you shopped for clothing during the last year?

- San Francisco
- Sacramento
- San Jose
- Los Angeles
- Other \_\_\_\_\_

Please list the clothing stores where you have shopped during the last two months.

Suppose a respondent checks Sacramento and San Francisco to the first question. If the same respondent lists a store that has a location only in Los Angeles in the second question, an error is indicated. Either the respondent failed to list Los Angeles in the first question or listed an erroneous store in the second question. These answers are obviously inconsistent.

### ■ TAKING ACTION WHEN RESPONSE IS OBVIOUSLY AN ERROR

What should the editor do? If solid evidence exists that points to the fact that the respondent simply failed to check Los Angeles, then the response to the first question can be changed to indicate that the person shopped in that city as well. Since Los Angeles is not listed next to Sacramento or San Francisco, it is unlikely that the respondent checked the wrong city inadvertently. Perhaps the question about the stores triggered a memory that did not come to the respondent when checking off the cities. This seems quite possible, and if another question can also point strongly to the fact that the respondent actually shopped at the store in Los Angeles, then the change should be made.

However, perhaps the respondent placed a mail order with the store in Los Angeles and simply did not physically shop in the store. If other evidence suggests this possibility, then the researcher should not make an adjustment to the first question. For example, a later question may have the respondent list any clothing orders placed via mail order (or by telephone or Internet order).

Many surveys use filter or “skip” questions that direct a respondent to a specific set of questions depending on how the filter question is answered according to the respondent’s answers. Common filter questions involve age, sex, home ownership, or product usage. A survey might involve different questions for a home owner than for someone who does not own a home. A data record may sometimes contain data on variables that the respondent should never have been asked. For example, if someone indicated that he or she did not own a home, yet responses for the home questions are provided, a problem is indicated. The editor may check other responses to make sure that the screening question was answered accurately. For instance, if the respondent left the question about home value unanswered, then the editor will be confident that the person truly does not own a home. In cases like this, the editor should adjust these answers by considering all answers to the irrelevant questions as “no response” or “not applicable.”

### ■ EDITING TECHNOLOGY

Today, computer routines can check for inconsistencies automatically. Thus, for electronic questionnaires, rules can be entered which prevent inconsistent responses from ever being stored in the file used for data analysis. These rules should represent the conservative judgment of a trained data analyst. Some online survey services can assist in providing this service. In fact, the rules can even be preprogrammed to prevent many inconsistent responses. Thus, if a person who is 25 indicates that he or she has lived in the same house for 48 years, a pop-up window can appear requiring the respondent to go back and fix an earlier incorrect response. Electronic questionnaires can also prevent a respondent from being directed to the wrong set of questions based on a screening question response.

### Editing for Completeness

In some cases the respondent may have answered only the second portion of a two-part question. The following question creates a situation in which an in-house editor may have to adjust answers for completeness:

Does your organization have more than one computer network server?

☐ Yes      ☐ No

If yes, how many? \_\_\_\_\_

If the respondent checked neither yes nor no but indicated three computer installations, the editor should change the first response to a “Yes” as long as other information doesn’t indicate otherwise. Here again, a computerized questionnaire may either not allow a response to the “how many” question if someone checked yes or require the respondent to go back to the previous question once he or she tries to enter a number for the “how many” question.

**Item nonresponse** is the technical term for an unanswered question on an otherwise complete questionnaire. Missing data results from item nonresponse. Specific decision rules for handling this problem should be meticulously outlined in the editor’s instructions. In many situations the decision rule is to do nothing with the missing data and simply leave the item blank. However, when the relationship between two questions is important, such as that between a question about job satisfaction and one’s pay, the editor may be tempted to insert a **plug value**. The decision rule may be to plug in an average or neutral value in each instance of missing data. Several choices are available:

1. Leave the response blank. Because the question is so important, the risk of creating error by plugging a value is too great.

2. Plug in alternate choices for missing data (“yes” the first time, “no” the second time, “yes” the third time, and so forth).
3. Randomly select an answer. The editor may flip a coin with heads for “yes” and tails for “no.”

The editor can **impute** a missing value based on the respondent’s choices to other questions. Many different techniques exist for imputing data. Some involve complex statistical estimation approaches that use the available information to forecast a best guess for the missing response.

The editor must decide whether an entire questionnaire is usable. When a questionnaire has too many missing answers, it may not be suitable for the planned data analysis. While no exact answer exists for this question, a questionnaire with a quarter of the responses or more missing is suspect. In such a situation the editor can record that a particular incomplete questionnaire has been dropped from the sample.

## Editing Questions Answered Out of Order

Another task an editor may face is rearranging the answers given to open-ended questions such as may occur in a focus group interview. The respondent may have provided the answer to a subsequent question in his or her comments to an earlier open-ended question. Because the respondent already had clearly identified the answer, the interviewer may not have asked the subsequent question, wishing to avoid hearing “I already answered that earlier” and to maintain interview rapport. If the editor is asked to list answers to all questions in a specific order, the editor may move certain answers to the section related to the skipped question.

## Facilitating the Coding Process

While all of the previously described editing activities will help coders, several editing procedures are designed specifically to simplify the coding process. For example, the editor should check written responses for any stray marks. Respondents are often asked to circle responses. Sometimes, a respondent may accidentally draw a circle that overlaps two numbers. For example, the circle may include both 3 and 4. The editor may be able to decide which number is the most accurate response and indicate that on the form. Occasionally, a respondent may do this to indicate indecision between the 3 and the 4. Again, if the editor sees that the circle is carefully drawn to include both responses, he or she may indicate a 3.5 on the form. Such ambiguity is impossible with an electronic questionnaire.

### ■ EDITING AND TABULATING "DON'T KNOW" ANSWERS

In many situations, respondents answer “don’t know.” On the surface, this response seems to indicate unfamiliarity with the subject matter at question. A *legitimate* “don’t know” response is the same as “no opinion.” However, there may be reasons for this response other than the legitimate “don’t know.” A *reluctant* “don’t know” is given when the respondent simply does not want to answer a question. For example, asking an individual who is not the head of the household about family income may elicit a “don’t know” answer meaning, “This is personal, and I really do not want to answer the question.” If the individual does not understand the question, he or she may give a *confused* “I don’t know” answer.

In some situations the editor can separate the legitimate “don’t knows” (“no opinion”) from the other “don’t knows.” The editor may try to identify the meaning of the “don’t know” answer from other data provided on the questionnaire.

## Pitfalls of Editing

Subjectivity can enter into the editing process. Data editors should be intelligent, experienced, and

*objective.* A *systematic procedure* for assessing the questionnaires should be developed by the research analyst so that the editor has clearly defined decision rules to follow. Any inferences such as imputing missing values should be done in a manner that limits the chance for the data editor's subjectivity to influence the response.

## Pretesting Edit

Editing questionnaires during the pretest stage can prove very valuable. For example, if respondents' answers to open-ended questions were longer than anticipated, the fieldworkers, respondents, and analysts would benefit from a change to larger spaces for the answers. Answers will be more legible because the writers have enough space, answers will be more complete, and answers will be verbatim rather than summarized. Examining answers to pretests may identify poor instructions or inappropriate question wording on the questionnaire.

## Coding

Editing may be differentiated from **coding**, which is the assignment of numerical scores or classifying symbols to previously edited data. Careful editing makes the coding job easier. Codes are meant to represent the meaning in the data.

Assigning numerical symbols permits the transfer of data from questionnaires or interview forms to a computer. **Codes** often, but not always, are numerical symbols. However, they are more broadly defined as rules for interpreting, classifying, and recording data. In qualitative research, numbers are seldom used for codes.

## Coding Qualitative Responses

### ■ UNSTRUCTURED QUALITATIVE RESPONSES (LONG INTERVIEWS)

In qualitative research, the codes are usually words or phrases that represent themes. In a hermeneutic unit in which a qualitative researcher is applying a code to a text describing in detail a respondent's reactions to several different glasses of wine. The researcher is trying to understand in detail what defines the wine drinking experience. In this case, coding is facilitated by the use of qualitative software.

After reading through the text several times, and applying a word-counting routine, the researcher realizes that appearance, the nose (aroma), and guessing (trying to guess what the wine will be like or what type of wine is in the glass) are important themes. A code is assigned to these categories. Similarly, other codes are assigned as shown in the *code manager window*. The density column shows how often a code is applied. After considerable thought and questioning of the experience, the researcher builds a network, or grounded theory, that suggests how a wine may come to be associated with feelings of romance. This theory is shown in the network view.

### ■ STRUCTURED QUALITATIVE RESPONSES

Qualitative responses to structured questions such as "yes" or "no" can be stored in a data file with letters such as "Y" or "N." Alternatively, they can be represented with numbers, one each to represent the respective category. So, the number 1 can be used to represent "yes" and 2 can be used to represent "no." Since this represents a nominal numbering system, the actual numbers used are arbitrary. Even though the codes are numeric, the variable is classificatory, simply separating the positive from the negative responses. For reasons that should become increasingly apparent in later chapters, the research may consider adopting **dummy coding** for dichotomous responses like yes or no. Dummy coding assigns a 0 to one category and a 1 to the other. So, for yes/no responses, a 0 could be "no" and a 1 would be "yes." Similarly, a "1" could represent a female respondent and a "0" would be a male respondent. Dummy

coding provides the researcher with more flexibility in how structured, qualitative responses are analyzed statistically. Dummy coding can be used when more than two categories exist, but because a dummy variable can only represent two categories, multiple dummy variables are needed to represent a single qualitative response that can take on more than two categories.

#### ■ DATA FILE TERMINOLOGY

Once structured, qualitative responses are coded, they are stored in an electronic data file. Here, both the qualitative responses and quantitative responses are likely stored for every respondent involved in a survey or interview. A terminology exists that helps describe this process and the file that results.

Some of the terminology seems strange these days. For instance, what does a “card” have to do with a simple computer file? Most of the terminology describing files goes back to the early days of computers. In those days, data and the computer programs that produced results were stored on actual computer cards. Hopefully, readers will no longer have to use physical cards to store data.

Researchers organize coded data into cards, fields, records, and files. Cards are the collection of records that make up a file. A **field** is a collection of characters (a *character* is a single number, letter, or special symbol such as a question mark) that represents a single piece of data, usually a variable. Some variables may require a large field, particularly for text data; other variables may require a field of only one character. Text variables are represented by **string characters**, which is computer terminology for a series of alphabetic characters (non-numeric characters) that may form a word. String characters often contain long fields of eight or more characters. In contrast, a dummy variable is a numeric variable that needs only one character to form a field.

A **record** is a collection of related fields. A record was the way a single, complete computer card was represented. Researchers may use the term *record* to refer to one respondent’s data. A **data file** is a collection of related records that make up a data set.

### The Data File

Data are generally stored in a matrix that resembles a common spreadsheet file. A data file stores the data from a research project and is typically represented in a rectangular arrangement (matrix) of data in rows and columns. Typically, each row represents a respondent’s scores on each variable and each row represents a variable for which there is a value for every respondent.

### Code Construction

There are two basic rules for code construction. First, the coding categories should be **exhaustive**, meaning that a coding category should exist for all possible responses. With a categorical variable such as sex, making categories exhaustive is not a problem. However, trouble may arise when the response represents a small number of subjects or when responses might be categorized into a class not typically found. For example, when questioned about automobile ownership, an antique car collector might mention that he drives a Packard Clipper. This may present a problem if separate categories have been developed for all possible makes of cars. Solving this problem frequently requires inclusion of an “other” code category to ensure that the categories are all-inclusive. For example, household size might be coded 1, 2, 3, 4, and 5 or more. The “5 or more” category assures all subjects of a place in a category.

Missing data should also be represented with a code. In the “good old days” of computer cards, a numeric value such as 9 or 99 was used to represent missing data. Today, most software will understand that either a period or a blank response represents missing data.

Second, the coding categories should be **mutually exclusive** and **independent**. This means that there should be no overlap among the categories to ensure that a subject or response can be placed in only one category.

## Precoding Fixed-Alternative Questions

When a questionnaire is highly structured, the categories may be precoded before the data are collected. Exhibit 19.5 presents a questionnaire for which the precoded response categories were determined before the start of data collection. The codes in the data file will correspond to the small numbers beside each choice option. In most instances, the codes will not actually appear on the questionnaire before the start of data collection. The codes in the data file will correspond to the small numbers beside each choice option. In most instances, the codes will not actually appear on the questionnaire.

The questionnaire in Exhibit 19.5 shows several demographic questions classifying individuals' scores. Question 29 has three possible answers, and they are precoded 1, 2, 3. Question 30 asks a person to respond "yes" (1) or "no" (2) to the question "Are you the male or female head of the household?" Once again, technology is making things easier and much of this type of coding is automated. For users of Web-based survey services, all that one need do is submit a questionnaire and in return he or she will receive a coded data file in the software of his or her choice.

Telephone interviews are still widely used. The partial questionnaire in Exhibit 19.6 on the next page shows a precoded format for a telephone interview. In this situation the interviewer circles the coded numerical score as the answer to the question.

Precoding can be used if the researcher knows what answer categories exist before data collection occurs. Once the questionnaire has been designed and the structured (or closed-form) answers identified, coding then becomes routine. In some cases, predetermined responses are based on standardized classification schemes

**EXHIBIT 19.5 Precoding Fixed-Alternative Responses**

**29. Do you—or does anyone else in your immediate household—belong to a labor union?**

<sup>1</sup>☐ Yes, I personally belong to a labor union.

<sup>2</sup>☐ Yes, another member of my household belongs to a labor union.

<sup>3</sup>☐ No, no one in my household belongs to a labor union.

**30. Are you the male or female head of the household—that is, the person whose income is the chief source of support of the household?**

<sup>1</sup>☐ Yes <sup>2</sup>☐ No

**31. Would you please check the appropriate combined yearly income (before income taxes and any other payroll deductions) from all sources of all those in your immediate household? (Please include income from salaries, investments, dividends, rents, royalties, bonuses, commissions, etc.) Please remember that your individual answers will not be divulged.**

<sup>1</sup> <input type="checkbox"/> Less than \$4,000	<sup>7</sup> <input type="checkbox"/> \$8,000–\$8,999	<sup>13</sup> <input type="checkbox"/> \$25,000–\$29,999
<sup>2</sup> <input type="checkbox"/> \$4,000–\$4,999	<sup>8</sup> <input type="checkbox"/> \$9,000–\$9,999	<sup>14</sup> <input type="checkbox"/> \$30,000–\$39,999
<sup>3</sup> <input type="checkbox"/> \$5,000–\$5,999	<sup>9</sup> <input type="checkbox"/> \$10,000–\$12,499	<sup>15</sup> <input type="checkbox"/> \$40,000–\$49,999
<sup>4</sup> <input type="checkbox"/> \$6,000–\$6,999	<sup>10</sup> <input type="checkbox"/> \$12,500–\$14,999	<sup>16</sup> <input type="checkbox"/> \$50,000–\$74,999
<sup>5</sup> <input type="checkbox"/> \$7,000–\$7,499	<sup>11</sup> <input type="checkbox"/> \$15,000–\$19,999	<sup>17</sup> <input type="checkbox"/> \$75,000–\$99,999
<sup>6</sup> <input type="checkbox"/> \$7,500–\$7,999	<sup>12</sup> <input type="checkbox"/> \$20,000–\$24,999	<sup>18</sup> <input type="checkbox"/> \$100,000 or more

**32. a. Do you personally own corporate stocks?** <sup>1</sup>☐ Yes <sup>2</sup>☐ No

**b. Do you own stocks in the corporation for which you work?**

**Do you own them in a corporation for which you do not work?**

**(Please check as many as apply.)**

Own STOCK in:

<sup>1</sup>☐ Company for which I work <sup>2</sup>☐ Other company

THANK YOU VERY MUCH FOR YOUR COOPERATION

If you would like to make any comments on any of the subjects covered in this study, please use the space below:

---



---



## Code Book

A **code book** gives each variable in the study and its location in the data matrix. Researchers commonly identify individual respondents by giving each an identification number or questionnaire number. When each interview is identified with a number entered into each computer record, errors discovered in the tabulation process can be checked on the questionnaire to verify the answer.

## Editing and Coding Combined

Frequently the person coding the questionnaire performs certain editing functions, such as translating an occupational title provided by the respondent into a code for socioeconomic status. A question that asks for a description of the job or business often is used to ensure that there will be no problem in classifying the responses. For example, respondents who indicate “salesperson” as their occupation might write their job description as “selling shoes in a shoe store” or “selling IBM supercomputers to the defense department.” Generally, coders are instructed to perform this type of editing function, seeking the help of a tabulation supervisor if questions arise.

## Computerized Survey Data Processing

In most studies with large sample sizes, a computer is used for data processing. The process of transferring data from a research project, such as answers to a survey questionnaire, to computers is referred to as **data entry**. Several alternative means exist for entering data into a computer. In studies involving highly structured paper and pencil questionnaires, an **optical scanning system** may be used to read material directly into the computer’s memory from *mark-sensed questionnaires*. The form may look similar to the type a student uses to take a multiple-choice test. As seen in the Research Snapshot above, even mobile phone technology is now being used to aid data processing.

# BASIC DATA ANALYSIS: DESCRIPTIVE STATISTICS

## Introduction

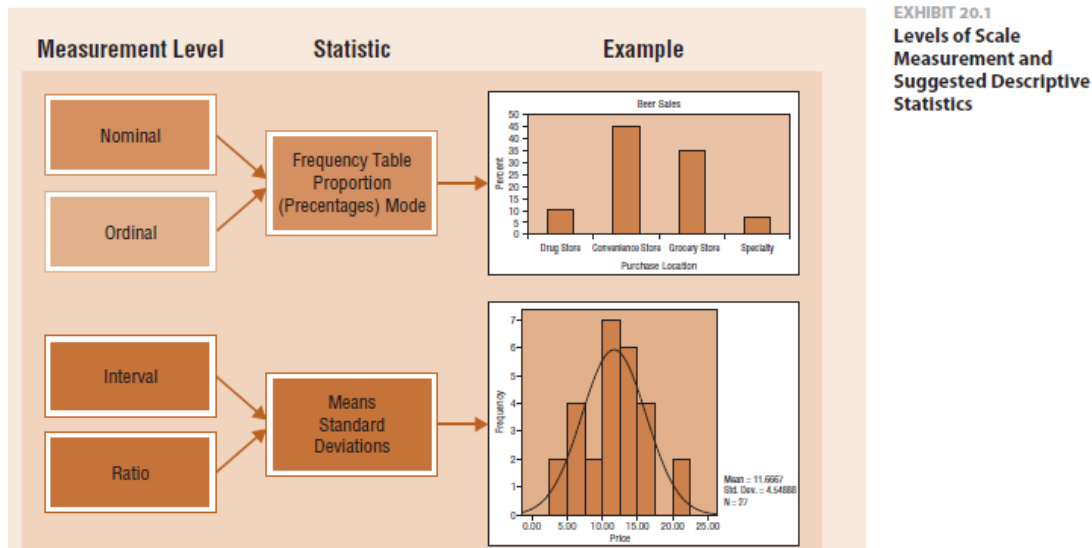
Perhaps the most basic statistical analysis is descriptive analysis. Descriptive statistics can summarize responses from large numbers of respondents in a few simple statistics. When a sample is obtained, the sample descriptive statistics are used to make inferences about characteristics of the entire population of interest. This chapter introduces basic descriptive statistics, which are simple but powerful. This chapter also provides the foundation for Chapter 21, which will extend basic statistics into the area of univariate statistical analysis.

## The Nature of Descriptive Analysis

**Descriptive analysis** is the elementary transformation of data in a way that describes the basic characteristics such as central tendency, distribution, and variability. For example, consider the business researcher who takes responses from 1,000 American consumers and tabulates their favorite soft drink brand and the price they expect to pay for a six-pack of that product. The mean, median, and mode for favorite soft drink and the average price across all 1,000 consumers would be descriptive statistics that describe central tendency in three different ways. Means, medians, modes, variance, range, and standard deviation typify widely applied descriptive statistics.

Consider the following data. Sample consumers were asked where they most often purchased beer. The result is a nominal variable which can be described with a frequency distribution (see the bar chart in Exhibit 20.1). Ten percent indicated they most often purchased beer in a drug store, 45 percent indicated a

convenience store, 35 percent indicated a grocery store, and 7 percent indicated a specialty store. Three percent listed some “other” outlet (not shown in the bar chart).



The bottom part of Exhibit 20.1 displays example descriptive statistics for interval and ratio variables. In this case, the chart displays results of a question asking respondents how much they typically spend on a bottle of wine purchased in a store.

## Tabulation

**Tabulation** refers to the orderly arrangement of data in a table or other summary format. When this tabulation process is done by hand, the term *tallying* is used. Counting the different ways respondents answered a question and arranging them in a simple tabular form yields a **frequency table**. The actual number of responses to each category is a variable’s frequency distribution. A simple tabulation of this type is sometimes called a *marginal tabulation*.

Simple tabulation tells the researcher how frequently each response occurs. This starting point for analysis requires the researcher to count responses or observations for each category or code assigned to a variable. A frequency table showing where consumers generally purchase beer can be computed easily. The tabular results that correspond to the chart would appear as follows:

Response	Frequency	Percent	Cumulative Percentage
Drug store	50	10	10
Convenience store	225	45	55
Grocery store	175	35	90
Specialty	35	7	97
Other	15	3	100

The frequency column shows the tally result or the number of respondents listing each store, respectively. The percent column shows the total percentage in each category. From this chart, we can see the most common outlet—the mode—is convenience store since more people indicated this as their top response

than any other. The cumulative percentage keeps a running total, showing the percentage of respondents indicating this particular category and all preceding categories as their preferred place to purchase beer. The cumulative percentage column is not so important for nominal or interval data, but is quite useful for interval and ratio data, particularly when there are a large number of response categories.

## Cross-Tabulation

A frequency distribution or tabulation can address many research questions. As long as a question deals with only one categorical variable, tabulation is probably the best approach. Although frequency counts, percentage distributions, and averages summarize considerable information, simple tabulation may not yield the full value of the research. **Cross-tabulation** is the appropriate technique for addressing research questions involving relationships among multiple less-than interval variables. We can think of a cross-tabulation is a combined frequency table. *Cross-tabs* allow the inspection and comparison of differences among groups based on nominal or ordinal categories. One key to interpreting a cross-tabulation table is comparing the observed table values with hypothetical values that would result from pure chance.

## Contingency Tables

A **contingency table** is a data matrix that displays the frequency of some combination of possible responses to multiple variables. Two-way contingency tables, meaning they involve two less-than interval variables, are used most often. A three-way contingency table involves three less-than interval variables. Beyond three variables, contingency tables become difficult to analyze and explain easily.

## Percentage Cross-Tabulations

When data from a survey are cross-tabulated, percentages help the researcher understand the nature of the relationship by making relative comparisons simpler. The total number of respondents or observations may be used as a **statistical base** for computing the percentage in each cell. When the objective of the research is to identify a relationship between answers to two questions (or two variables), one of the questions is commonly chosen to be the source of the base for determining percentages.

## Elaboration and Refinement

The *Oxford Universal Dictionary* defines *analysis* as “the resolution of anything complex into its simplest elements.” Once a researcher has examined the basic relationship between two variables, he or she may wish to investigate this relationship under a variety of different conditions. Typically, a third variable is introduced into the analysis to elaborate and refine the researcher’s understanding by specifying the conditions under which the relationship between the first two variables is strongest and weakest. In other words, a more elaborate analysis asks, “Will interpretation of the relationship be modified if other variables are simultaneously considered?”

**Elaboration analysis** involves the basic cross-tabulation within various subgroups of the sample. The researcher breaks down the analysis for each level of another variable. Exhibit 20.4 breaks down the responses to the question “Do you shop at Target?” by sex and marital status. The data show women display the same preference whether married or single. However, married men are much more likely to shop at Target than are single men. The analysis suggests that the original conclusion about the relationship between sex and shopping behavior for women be retained. However, a relationship that was not discernible in the two-variable case is evident. Married men more frequently shop at Target than do single men.

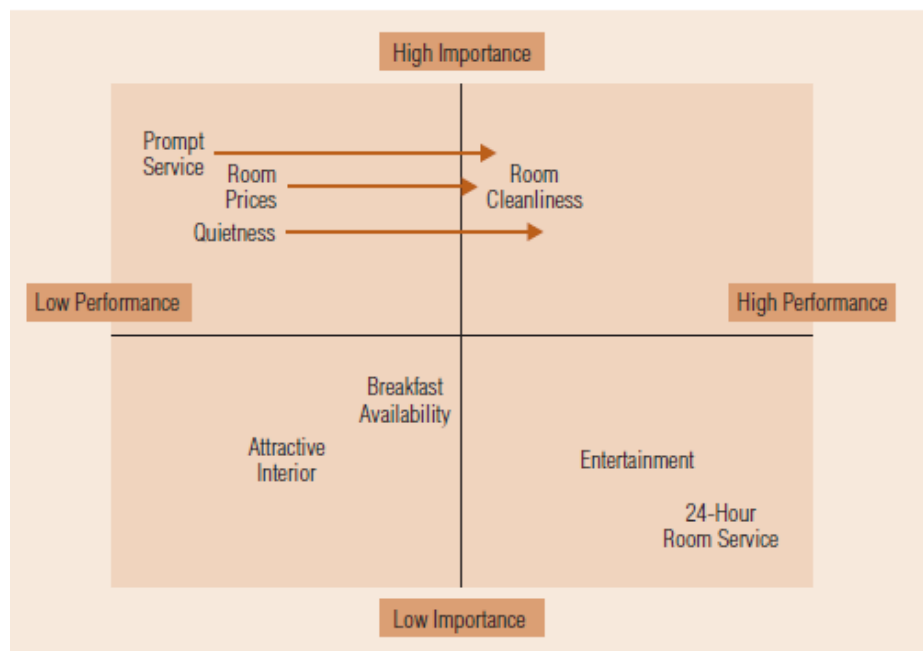
	Single		Married	
	Men	Women	Men	Women
"Do you shop at Target?"				
Yes	55%	80%	86%	80%
No	45%	20%	14%	20%

The finding is consistent with an interaction effect. The combination of the two variables, sex and marital status, is associated with differences in the dependent variable. Interactions between variables examine moderating variables. A **moderator variable** is a third variable that changes the nature of a relationship between the original independent and dependent variables. Marital status is a moderator variable in this case. The interaction effect suggests that marriage changes the relationship between sex and shopping preference.

### Quadrant Analysis

**Quadrant analysis** is a variation of cross-tabulation in which responses to two rating scale questions are plotted in four quadrants of a two-dimensional table. A common quadrant analysis in business research portrays or plots relationships between average responses about a product attribute's importance and average ratings of a company's (or brand's) performance on that product feature. The term **importance-performance analysis** is sometimes used because consumers rate perceived importance of several attributes and rate how well the company's brand performs on that attribute. Generally speaking, the business would like to end up in the quadrant indicating high performance on an important attribute.

Exhibit 20.5 illustrates a quadrant analysis for an international, mid-priced hotel chain. The chart shows the importance and the performance ratings provided by business travelers. After plotting the scores for each of eight attributes, the analysis suggests areas for improvement. The arrows indicate attributes that the hotel firm should concentrate on to move from quadrant three, which means the performance on those attributes is low but business consumers rate those attributes as important, to quadrant four, where attributes are both important and rated highly for performance.



## Data Transformation

### Simple Transformations

**Data transformation** (also called *data conversion*) is the process of changing the data from their original form to a format suitable for performing a data analysis that will achieve research objectives. Researchers often modify the values of scalar data or create new variables. For example, many researchers believe that less response bias will result if interviewers ask respondents for their year of birth rather than their age. This presents no problem for the research analyst, because a simple data transformation is possible. The raw data coded as birth year can easily be transformed to age by subtracting the birth year from the current year.

In earlier chapters, we discussed recoding and creating summated scales. These also are common data transformations.

Collapsing or combining adjacent categories of a variable is a common form of data transformation used to reduce the number of categories. A Likert scale may sometimes be collapsed into a smaller number of categories. For instance, consider the following Likert item administered to a sample of state university seniors:

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I am satisfied with my college experience at this university	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

The following frequency table describes results for this survey item:

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
110	30	15	35	210

### Problems with Data Transformations

Researchers often perform a median split to collapse a scale with multiple response points into two categories. The **median split** means respondents below the observed median go into one category and respondents above the median go into another. Although this is common, the approach is best applied only when the data do indeed exhibit bimodal characteristics. When the data are unimodal, such as would be the case with normally distributed data, a median split will throw away valuable information and lead to error.

When a sufficient number of responses exist and a variable is ratio, the researcher may choose to delete one-fourth to one-third of the responses around the median to effectively ensure a bimodal distribution. However, median splits should always be performed only with great care, as the inappropriate collapsing of continuous variables into categorical variables ignores the information contained within the untransformed values. Rather than splitting a continuous variable into two categories to conduct a frequency distribution or cross-tabulation.

### Index Numbers

The consumer price index and wholesale price index are secondary data sources that are frequently used by business researchers. Price indexes, like other **index numbers**, represent simple data transformations that allow researchers to track a variable's value over time and compare a variable(s) with other variables. Recalibration allows scores or observations to be related to a certain base period or base number.

Consider the information in Exhibit 20.8. Weekly television viewing statistics are shown grouped by household size. Index numbers can be computed for these observations in the following manner:

1. A base number is selected. The U.S. household average of 52 hours and 36 minutes represents the central tendency and will be used.
2. Index numbers are computed by dividing the score for each category by the base number and multiplying by 100. The index reflects percentage changes from the base:

$$\begin{aligned}
 \text{1 person hh:} & \quad \frac{41:01}{52:36} = 0.7832 \times 100 = 78.32 \\
 \text{2 person hh:} & \quad \frac{47:58}{52:36} = 0.9087 \times 100 = 90.87 \\
 \text{3+ person hh:} & \quad \frac{60:49}{52:36} = 1.1553 \times 100 = 115.53 \\
 \text{Total U.S. average:} & \quad \frac{52:36}{52:36} = 1.0000 \times 100 = 100.00
 \end{aligned}$$

Household Size	Hours:Minutes
1	41:01
2	47:58
3+	60:49
Total U.S. average	52:36

If the data are time-related, a base year is chosen. The index numbers are then computed by dividing each year's activity by the base-year activity and multiplying by 100. Index numbers require ratio measurement scales.

## Calculating Rank Order

Survey respondents are often asked to rank order their preference for some item, issue, or characteristic. For instance, consumers may be asked to rank their three favorite brands or employee respondents may provide rankings of several different employee benefit plans. Ranking data can be summarized by performing a data transformation. The transformation involves multiplying the frequency by the ranking score for each choice to result in a new scale.

Executive	Hawaii	Paris	Greece	Hong Kong
1	1	2	4	3
2	1	3	4	2
3	2	1	3	4
4	2	4	3	1
5	2	1	3	4
6	3	4	1	2
7	2	3	1	4
8	1	4	2	3
9	4	3	2	1
10	2	1	3	4

Destination	Preference Rankings			
	1st	2nd	3rd	4th
Hawaii	3	5	1	1
Paris	3	1	3	3
Greece	2	2	4	2
Hong Kong	2	2	2	4

For example, suppose a CEO had 10 executives rank their preferences for locations in which to hold the company's annual conference. Exhibit 20.9 shows how executives ranked each of four locations: Hawaii, Paris, Greece, and Hong Kong. Exhibit 20.10 tabulates frequencies for these rankings. A ranking summary can be computed by assigning the destination with the highest preference the lowest number (1) and the least preferred destination the highest consecutive number (4). The summarized rank orderings were obtained with the following calculations:

$$\begin{aligned}
 \text{Hawaii:} & \quad (3 \times 1) + (5 \times 2) + (1 \times 3) + (1 \times 4) = 20 \\
 \text{Paris:} & \quad (3 \times 1) + (1 \times 2) + (3 \times 3) + (3 \times 4) = 26 \\
 \text{Greece:} & \quad (2 \times 1) + (2 \times 2) + (4 \times 3) + (2 \times 4) = 26 \\
 \text{Hong Kong:} & \quad (2 \times 1) + (2 \times 2) + (2 \times 3) + (4 \times 4) = 28
 \end{aligned}$$

Three executives chose Hawaii as the best destination (ranked "1"), five executives selected Hawaii as the second best destination, and so forth. The lowest total score indicates the first (highest) preference ranking. The results show the following rank ordering: (1) Hawaii, (2) Paris, (3) Greece, and (4) Hong Kong. Company employees may be glad to hear their conference will be in Hawaii!

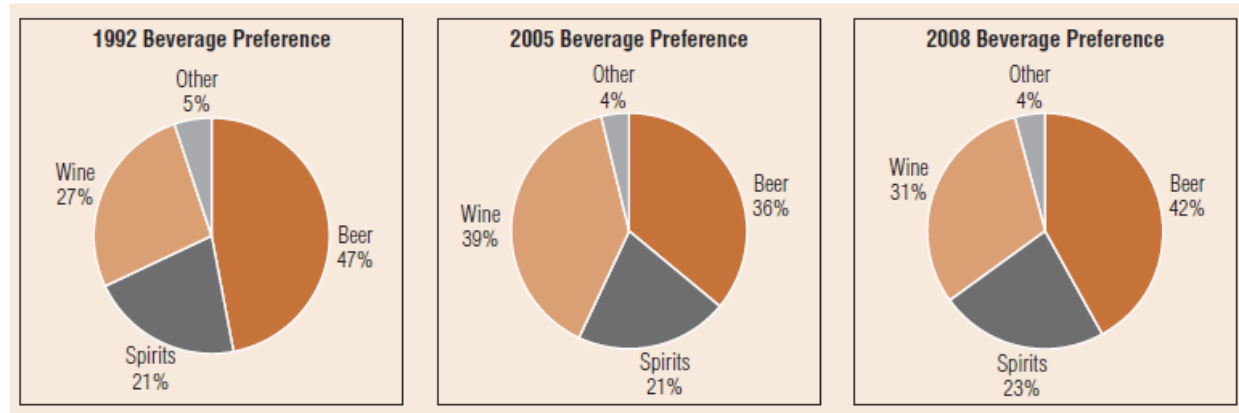
## Tabular and Graphic Methods of Displaying Data

Tables, graphs, and charts may simplify and clarify data. Graphical representations of data may take a number of forms, ranging from a computer printout to an elaborate pictograph. Tables, graphs, and charts, however, all facilitate summarization and communication. For example, see how the simple frequency table and histogram shown in Exhibit 20.7 provide a summary that quickly and easily communicates meaning that would be more difficult to see if all 350 responses were viewed separately.

Today's researcher has many convenient tools to quickly produce charts, graphs, or tables. Even common programs such as Excel and Word include chart functions that can construct the chart within the text document. Bar charts (histograms), pie charts, curve/line diagrams, and scatter plots are among the most widely used tools. Some choices match well with certain types of data and analyses.

Bar charts and pie charts are very effective in communicating frequency tabulations and simple cross-tabulations. Exhibit 20.11 displays frequency data from the chapter vignette with pie charts. Each pie summarizes preference in the respective year. The size of each pie slice corresponds to a frequency value associated with that choice. When the three pie charts are compared, the result communicates a cross-tabulation. Here, the comparison clearly communicate that wine preference increased at the expense of beer preference from 1992 to 2005, but has yielded some ground in 2008. In other words, the relative slice of pie for wine became larger, then slightly smaller.

EXHIBIT 20.11 Pie Charts Work Well with Tabulations and Cross-Tabulations



## Computer Programs for Analysis

### Statistical Packages

Just 50 years ago, the thought of a typical U.S. company performing even basic statistical analyses, like cross-tabulations, on a thousand or more observations was unrealistic. The personal computer brought this capability not just to average companies, but to small companies and individuals with limited resources. Today, computing power is very rarely a barrier to completing a research project.

In the 1980s and early 1990s, when the PC was still a relatively novel innovation, specialized statistical software formerly used on mainframe computers made their way into the personal computing market. Today, most spreadsheet packages can perform a wide variety of basic statistical options. Excel's basic data analysis tool will allow descriptive statistics including frequencies and measures of central tendency to be easily computed.<sup>10</sup> Most of the basic statistical features are now menu driven, reducing the need to memorize function labels. Spreadsheet packages like Excel continue to evolve and become more viable for performing many basic statistical analyses.

Despite the advances in spreadsheet applications, commercialized statistical software packages remain extremely popular among researchers. They continue to become easier to use and more compatible with other data interface tools including spreadsheets and word processors. Like any specialized tool, statistical packages are more tailored to the types of analyses performed by statistical analysts, including business researchers. Thus, any serious business or social science researcher should still become familiar with at least one general computer software package.

Two of the most popular general statistical packages are SAS (<http://www.sas.com>) and SPSS (<http://www.spss.com>). SAS revenues exceed \$2.15 billion in 2008 and its software can be found on computers worldwide. SAS was founded in 1976, and its statistical software historically has been widely used in engineering and other technical fields. SPSS stands for *Statistical Package for the Social Sciences*. SPSS was founded in 1968 and sales now exceed \$300 million annually. SPSS is commonly used by university business and social science students. Business researchers have traditionally used SPSS more than any other statistical software tool. SPSS has been viewed as more "user-friendly" in the past. However, today's versions of both SPSS and SAS are very user friendly and give the user the option of using drop-down menus to conduct analysis rather than writing computer code.

Excel, SAS, and SPSS account for most of the statistical analysis conducted in business research. University students may also be exposed to MINITAB, which is sometimes preferred by economists. However, MINITAB has traditionally been viewed as being less user-friendly than other choices.

In the past, data entry was an issue as specific software required different types of data input. Today,



however, all the major software packages, including SAS and SPSS, can work from data entered into a spreadsheet. The spreadsheets can be imported into the data windows or simply read by the program. Most conventional online survey tools will return data to the user in the form of either an SPSS data file, an Excel spreadsheet, or a plain text document.

### Computer Graphics and Computer Mapping

Graphic aids prepared by computers have replaced graphic presentation aids drawn by artists. Computer graphics are extremely useful for descriptive analysis. We know, decision support systems can generate two- or three-dimensional computer maps to portray data about sales, demographics, lifestyles, retail stores, and other features. Exhibit 20.14 shows a computer graphic depicting how fast-food consumption varies from state to state. The chart shows the relative frequencies of eating fast-food burgers, chicken, tacos, or other types of fast food across several states. Computer graphics like these have become more common as common applications have introduced easy ways of generating 3-D graphics and maps. Many computer maps are used by business executives to show locations of high-quality customer segments. Competitors' locations are often overlaid for additional quick and easy visual reference. Scales that show miles, population densities, and other characteristics can be highlighted in color, with shading, and with symbols.

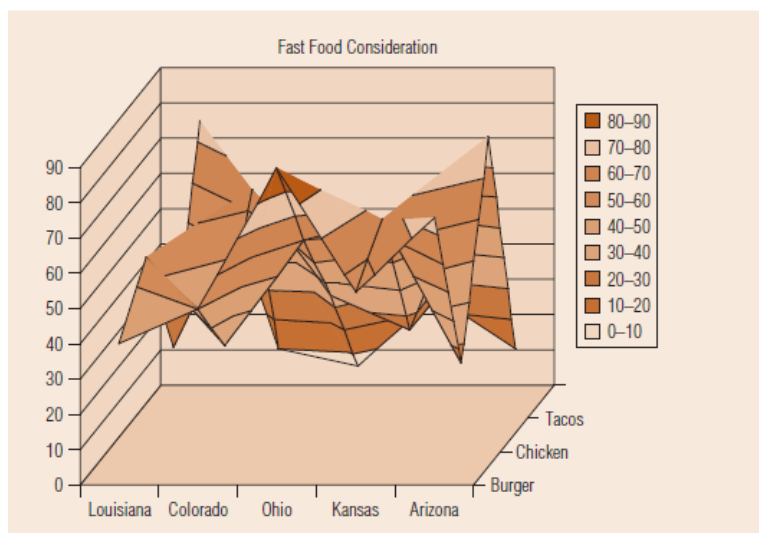


EXHIBIT 20.14  
A 3-D Graph Showing Fast-Food Consumption Patterns around the United States

Many computer programs can draw **box and whisker plots**, which provide graphic representations of central tendencies, percentiles, variabilities, and the shapes of frequency distributions

### Interpretation

An interpreter at the United Nations translates a foreign language into another language to explain the meaning of a foreign diplomat's speech. In business research, the interpretation process explains the meaning of the data. After the statistical analysis of the data, inferences and conclusions about their meaning are developed.

A distinction can be made between *analysis* and *interpretation*. **Interpretation** is drawing inferences from the analysis results. Inferences drawn from interpretations lead to managerial implications. In other words, each statistical analysis produces results that are interpreted with respect to insight into a particular decision. The logical interpretation of the data and statistical analysis are closely intertwined. When a researcher calculates a cross-tabulation of employee number of dependents with choice of health plan, an interpretation is drawn suggesting that employees with a different number of dependents may be more or less likely to choose a given health place.

# UNIVARIATE STATISTICAL ANALYSIS

## Hypothesis Testing

Descriptive research and causal research designs often climax with hypothesis tests. Hypotheses are defined as formal statements of explanations stated in a testable form. Generally, hypotheses should be stated in concrete fashion so that the method of empirical testing seems almost obvious. Types of hypotheses tested commonly in business research include the following:

1. Relational hypotheses—examine how changes in one variable vary with changes in another. This is usually tested by assessing covariance in some way, very often with regression analysis.
2. Hypotheses about differences between groups—examine how some variable varies from one group to another. These types of hypotheses are very common in causal designs.
3. Hypotheses about differences from some standard—examine how some variable differs from some preconceived standard. The preconceived standard sometimes represents the true value of the variable in a population. These tests can involve either a test of a mean for better-than ordinal variables or a test of frequencies if the variable is ordinal or nominal. These tests typify univariate statistical tests.

## The Hypothesis-Testing Procedure

### ■ PROCESS

Hypotheses are tested by comparing the researcher's educated guess with empirical reality. The process can be described as follows:

1. First, the hypothesis is derived from the research objectives. The hypothesis should be stated as specifically as possible.
2. Next, a sample is obtained and the relevant variable is measured.
3. The measured value obtained in the sample is compared to the value either stated explicitly or implied in the hypothesis. If the value is consistent with the hypothesis, the hypothesis is supported. If the value is not consistent with the hypothesis, the hypothesis is not supported.

A univariate hypothesis consistent with the chapter vignette would be

*H<sub>1</sub>. The average satisfaction at the Madison plant is greater than 3.5.*

If the average job satisfaction is 3.4, the hypothesis is not supported. If the average job satisfaction is 3.9, the hypothesis is supported.

Univariate hypotheses are typified by tests comparing some observed sample mean against a benchmark value. The test addresses the question, Is the sample mean truly different from the benchmark? But how different is really different? If the observed sample mean is 3.45 and the benchmark is 3.50, would the hypothesis still be supported? Probably not! When the observed mean is so close to the benchmark, we do not have sufficient confidence that a second set of data using a new sample taken from the same population might not produce a finding conflicting with the benchmark. In contrast, when the mean turns out well above 3.5, perhaps 3.9, then we could more easily trust that another sample would not produce a mean equal to or less than 3.5.

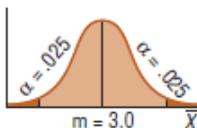
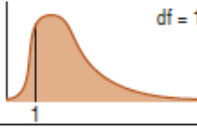
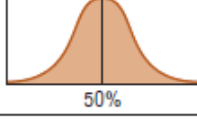
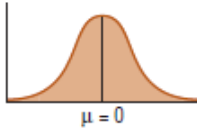
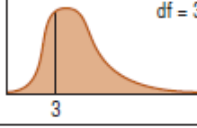
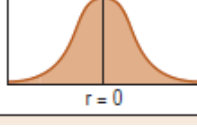
In statistics classes, students are exposed to hypothesis testing as a contrast between a *null* and an *alternative* hypothesis. A “null” hypothesis can be thought of as the expectation of findings as if no

hypothesis existed (i.e., “no” or “null” hypothesis). In other words, the state implied by the null hypothesis is the opposite of the state represented by the actual hypothesis.

### ■ SIGNIFICANCE LEVELS AND p-VALUES

A **significance level** is a critical probability associated with a statistical hypothesis test that indicates how likely it is that an inference supporting a difference between an observed value and some statistical expectation is true. The term **p-value** stands for probability-value and is essentially another name for an *observed* or *computed* significance level. Exhibit 21.1 discusses interpretations of p-values in different kinds of statistical tests. The probability in a p-value is that the statistical expectation (null) for a given test is true. So, low p-values mean there is little likelihood that the statistical expectation is true. This means the researcher’s hypothesis positing (suggesting) a difference between an observed mean and a population mean, or between an observed frequency and a population frequency, or for a relationship between two variables, is likely supported.

**EXHIBIT 21.1**  
**p-Values and Statistical Tests**

Test Description	Test Statistic	
Compare an Observed Mean with Some Predetermined Value	Z or t-test—Low p-values Indicate the Observed Mean Is Different Than Some Predetermined Value (Often 0)	
Compare an Observed Frequency with a Predetermined Value	$\chi^2$ —Low p-values Indicate That Observed Frequency Is Different Than Predetermined Value	
Compare an Observed Proportion with Some Predetermined Value	Z or t-test for Proportions—Low p-values Indicate That the Observed Proportion Is Different Than the Predetermined Value	
<b>Bivariate Tests:</b>		
Compare Whether Two Observed Means Are Different from One Another.	Z or t-test—Low p-values Indicate the Means Are Different	
Compare Whether Two Less-Than Interval Variables Are Related Using Cross-tabs	$\chi^2$ —Low p-values Indicate the Variables Are Related to One Another	
Compare Whether Two Interval or Ratio Variables Are Correlated to One Another	t-test for Correlation—Low p-values Indicate the Variables Are Related to One Another	

Traditionally, researchers have specified an acceptable significance level for a test prior to the analysis. Later, we will discuss this as an acceptable amount of Type I error. For most applications, the acceptable amount of error, and therefore the acceptable significance level, is 0.1, 0.05, or 0.01. If the p-value resulting from a statistical test is less than the pre-specified significance level, then a hypothesis about differences is supported.

## Type I and Type II Errors

Hypothesis testing using sample observations is based on probability theory. We make an observation of a sample and use it to infer the probability that some observation is true within the population the sample represents. Because we cannot make any statement about a sample with complete certainty, there is always the chance that an error will be made. When a researcher makes the observation using a census, meaning that every unit (person or object) in a population is measured, then conclusions are certain. Researchers very rarely use a census.

The researcher using sampling runs the risk of committing two types of errors. Exhibit 21.4 summarizes the state of affairs in the population and the nature of Type I and Type II errors. The four possible situations in the exhibit result because the null hypothesis (using the example above,  $\mu = 3.0$ ) is actually either true or false and the observed statistics ( $\bar{X} = 3.78$ ) will result in acceptance or rejection of this null hypothesis.

Actual State in the Population	Decision	
	Accept $H_0$	Reject $H_0$
$H_0$ is true	Correct—no error	Type I error
$H_0$ is false	Type II error	Correct—no error

EXHIBIT 21.4  
Type I and Type II  
Errors in Hypothesis Testing

**TO THE POINT**

### ■ TYPE I ERROR

Suppose the observed sample mean described above leads to the conclusion that the mean is greater than 3.0 when in fact the true population mean is equal to 3.0. A **Type I error** has occurred. A Type I error occurs when a condition that is true in the population is rejected based on statistical observations. When a researcher sets an acceptable significance level a priori a he or she is determining tolerance for a Type I error. Simply put, a Type I error occurs when the researcher concludes that there is a statistical difference when in reality one does not exist. When testing for relationships, a Type I error occurs when the researcher concludes a relationship exists when in fact one does not exist.

### ■ TYPE II ERROR

If the alternative condition is in fact true (in this case the mean is not equal to 3.0) but we conclude that we should not reject the null hypothesis (accept that the mean is equal to 3.0), we make what is called a **Type II error**. A Type II error is the probability of failing to reject a false null hypothesis. This incorrect decision is called beta ( $\beta$ ). In practical terms, a Type II error means that we fail to reach the conclusion that some difference between an observed mean and a benchmark exists when in fact the difference is very real. In terms of a bivariate correlation, a Type II error would mean the idea that a relationship exists between two variables is rejected when in fact the relationship does indeed exist. The Research Snapshot on the next page provides further clarification of the Type I and Type II conditions.

## Choosing the Appropriate Statistical Technique

Numerous statistical techniques are available to assist the researcher in interpreting data. Choosing the right tool for the job is just as important to the researcher as to the mechanic. Making the correct choice can be determined by considering

1. The type of question to be answered
2. The number of variables involved
3. The level of scale measurement

Today, the researcher rarely has to perform a paper and pencil calculation. Hypotheses are tested by using a correct click-through sequence in a statistical software package. The mathematics of these packages is highly reliable. Therefore, if the researcher can choose the right statistic, know the right click-through sequence, and read the output that results, the right statistical conclusion should be easy to reach.

### **Type of Question to Be Answered**

The type of question the researcher is attempting to answer is a consideration in the choice of statistical technique. For example, a researcher may be concerned simply with the central tendency of a variable or with the distribution of a variable. Comparison of different business divisions' sales results with some target level will require a one-sample t-test. Comparison of two salespeople's average monthly sales will require a t-test of two means, but a comparison of quarterly sales distributions will require a chi-square test.

The researcher should consider the method of statistical analysis before choosing the research design and before determining the type of data to collect. Once the data have been collected, the initial orientation toward analysis of the problem will be reflected in the research design.

### **Number of Variables**

The number of variables that will be simultaneously investigated is a primary consideration in the choice of statistical technique. A researcher who is interested only in the average number of times a prospective home buyer visits financial institutions to shop for interest rates can concentrate on investigating only one variable at a time. However, a researcher trying to measure multiple complex organizational variables cannot do the same. Simply put, univariate, bivariate, and multivariate statistical procedures are distinguished based on the number of variables involved in an analysis.

### **Level of Scale of Measurement**

The scale measurement level helps choose the most appropriate statistical techniques and appropriate empirical operations. Testing a hypothesis about a mean, as we have just illustrated, is appropriate for interval scaled or ratio scaled data. Suppose a researcher is working with a nominal scale that identifies users versus nonusers of bank credit cards. Because of the type of scale, the researcher may use only the mode as a measure of central tendency. In other situations, where data are measured on an ordinal scale, the median may be used as the average or a percentile may be used as a measure of dispersion. For example, ranking brand preferences generally employs an ordinal scale. Nominal and ordinal data are often analyzed using frequencies or cross-tabulation.

### **Parametric versus Nonparametric Hypothesis Tests**

The terms **parametric statistics** and **nonparametric statistics** refer to the two major groupings of statistical procedures. The major distinction between them lies in the underlying assumptions about the data to be analyzed. Parametric statistics involve numbers with known, continuous distributions. When the data are interval or ratio scaled and the sample size is large, parametric statistical procedures are appropriate. Nonparametric statistics are appropriate when the numbers do not conform to a known distribution.

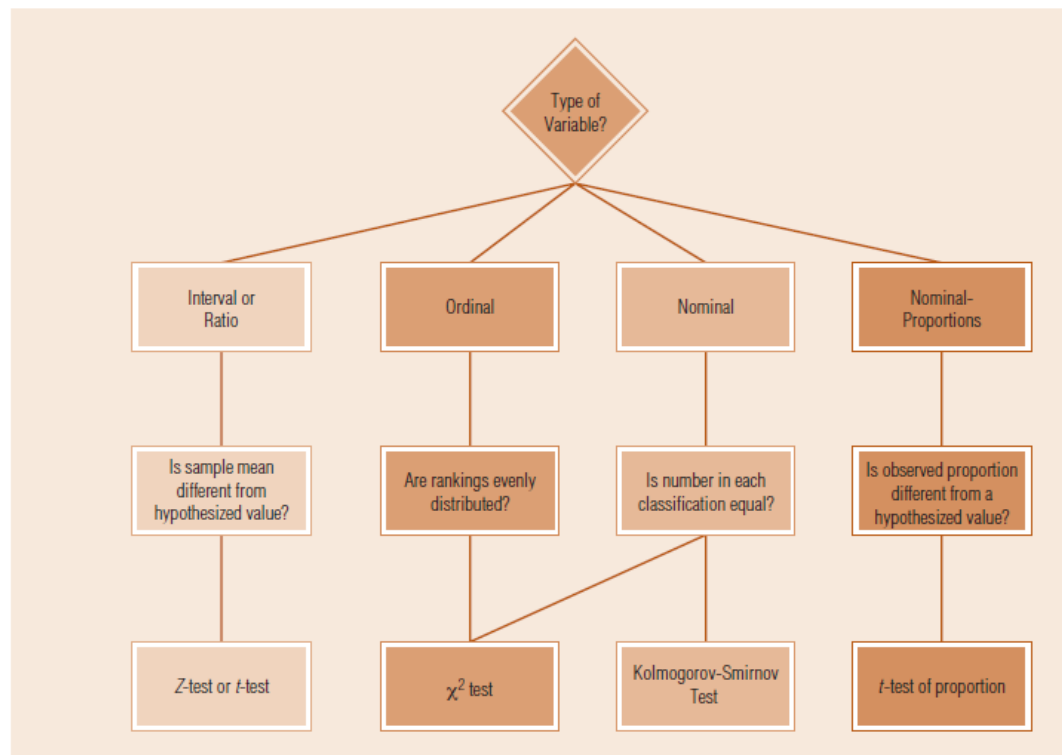
Parametric statistics are based on the assumption that the data in the study are drawn from a population with a normal (bell-shaped) distribution and/or normal sampling distribution. For example, if an investigator has two interval-scaled measures, such as gross national product (GNP) and industry sales

volume, parametric tests are appropriate. Possible statistical tests might include product-moment correlation analysis, analysis of variance, regression, or a f-test for a hypothesis about a mean.

Nonparametric methods are used when the researcher does not know how the data are distributed. Making the assumption that the population distribution or sampling distribution is normal generally is inappropriate when data are either ordinal or nominal. Thus, nonparametric statistics are referred to as distribution free. Data analysis of both nominal and ordinal scales typically uses nonparametric statistical tests.

Exhibit 21.5 illustrates how an appropriate univariate statistical method can be selected. The exhibit illustrates how statistical techniques vary according to scale properties and the type of question being asked.

EXHIBIT 21.5 Univariate Statistical Choice Made Easy

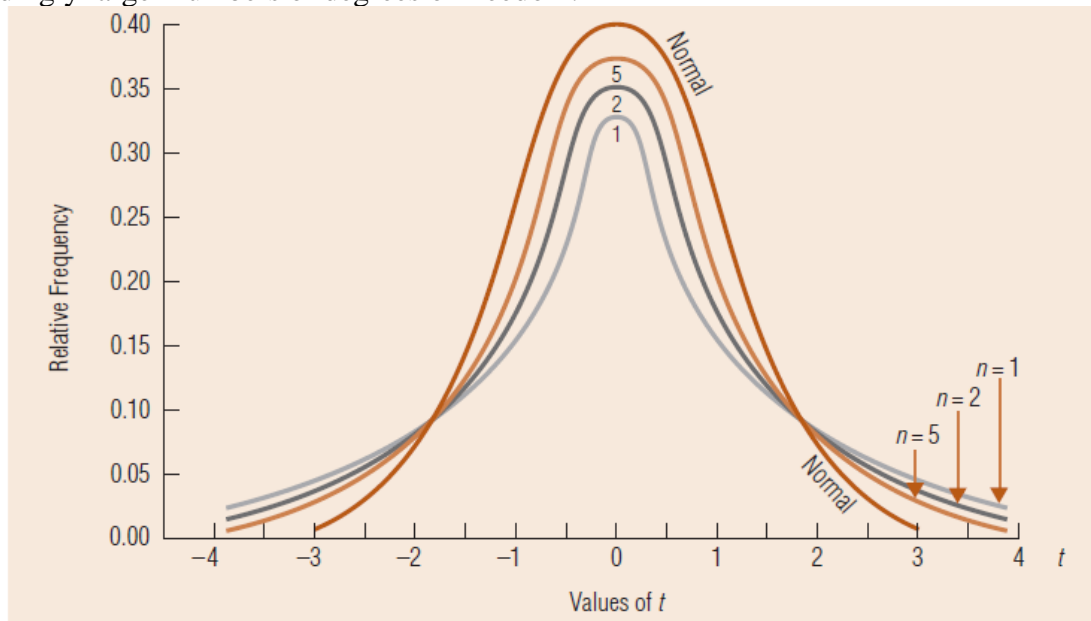


## The t-Distribution

A univariate **t-test** is appropriate for testing hypotheses involving some observed mean against some specified value. The **t-distribution**, like the standardized normal curve, is a symmetrical, bell-shaped distribution with a mean of 0 and a standard deviation of 1.0. When sample size ( $n$ ) is larger than 30, the t-distribution and Z-distribution are almost identical. Therefore, while the t-test is strictly appropriate for tests involving small sample sizes with unknown standard deviations, researchers commonly apply the f-test for comparisons involving the mean of an interval or ratio measure. The precise height and shape of the f-distribution vary with sample size. More specifically, the shape of the t-distribution is influenced by its **degrees of freedom (df)**. The degrees of freedom are determined by the number of distinct calculations that are possible given a set of information. In the case of a univariate t-test, the degrees of freedom are equal to the sample size ( $n$ ) minus one.

Exhibit 21.6 illustrates t-distributions for 1, 2, 5, and an infinite number of degrees of freedom. Notice that the t-distribution approaches a normal distribution rapidly with increasing sample size. This is

why, in practice, marketing researchers usually apply a t-test even with large samples. The practical effect is that the conclusion will be the same since the distributions are so similar with large samples and the correspondingly larger numbers of degrees of freedom.



### Calculating a Confidence Interval Estimate Using the t-Distribution

Suppose a business organization is interested in finding out how long newly hired MBA graduates remain on their first jobs. On the basis of a small sample of employees with MBAs, the researcher wishes to estimate the population mean with 95 percent confidence. The data from the sample are presented below.

Number of years on first job: 3 5 7 1 12 1 2 2 5  
4 2 3 1 3 4 2 6 7

To find the confidence interval estimate of the population mean for this small sample, we use the formula

$$\mu = \bar{X} \pm t_{c.l.} S_{\bar{X}}$$

or

$$\text{Upper limit} = \bar{X} + t_{c.l.} \left( \frac{S}{\sqrt{n}} \right)$$

$$\text{Lower limit} = \bar{X} - t_{c.l.} \left( \frac{S}{\sqrt{n}} \right)$$

where

- $\mu$  = population mean
- $(\bar{X})$  = sample mean
- $t_{c.l.}$  = critical value of  $t$  at a specified confidence level
- $S_{\bar{X}}$  = standard error of the mean
- $S$  = sample standard deviation
- $n$  = sample size

More specifically, the step-by-step procedure for calculating the confidence interval is as follows:



1. We calculate  $(\bar{X})$  from the sample. Summing our data values yields  $\Sigma X = 70$ , and  $(\bar{X}) = \Sigma X/n = 70/18 = 3.89$ .
2. Since  $\sigma$  is unknown, we estimate the population standard deviation by finding  $S$ , the sample standard deviation. For our example,  $S = 2.81$ .
3. We estimate the standard error of the mean using the formula  $S_{\bar{X}} = S/\sqrt{n}$ . Thus,  $S_{\bar{X}} = 2.81/\sqrt{18}$  or  $S_{\bar{X}} = 0.66$ .
4. We determine the  $t$ -values associated with the desired confidence level. To do this, we go to Table A.3 in the appendix. Although the  $t$ -table provides information similar to that in the  $Z$ -table, it is somewhat different. The  $t$ -table format emphasizes the chance of error, or significance level ( $\alpha$ ), rather than the 95 percent chance of including the population mean in the estimate. Our example is a two-tailed test. Since a 95 percent confidence level has been selected, the significance level equals  $0.05(1.00 - 0.95 = 0.05)$ . Once this has been determined, all we have to do to find the  $t$ -value is look under the 0.05 column for *two-tailed tests* at the row

in which degrees of freedom ( $df$ ) equal the appropriate value ( $n - 1$ ). Below 17 degrees of freedom ( $n - 1 = 18 - 1 = 17$ ), the  $t$ -value at the 95 percent confidence level (0.05 level of significance) is  $t = 2.12$ .

5. We calculate the confidence interval:

$$\text{Lower limit} = 3.89 - 2.12 \left( \frac{2.81}{\sqrt{18}} \right) = 2.49$$

$$\text{Upper limit} = 3.89 + 2.12 \left( \frac{2.81}{\sqrt{18}} \right) = 5.28$$

In our hypothetical example it may be concluded with 95 percent confidence that the population mean for the number of years spent on the first job by MBAs is between 2.49 and 5.28.

### ■ ONE- AND TWO-TAILED $t$ -TESTS

Univariate 2-tests and  $t$ -tests can be one- or two-tailed. A two-tailed test is one that tests for differences from the population mean that are either greater or less. Thus, the extreme values of the normal curve (or tails) on both the right and the left are considered. In practical terms, when a research question does not specify whether a difference should be greater than or less than, a two-tailed test is most appropriate. For instance, the following research question could be examined using a two-tailed test:

The number of take-out pizza restaurants within a postal code in Germany is not equal to 5.

A one-tailed univariate test is appropriate when a research hypothesis implies that an observed mean can only be greater than or less than a hypothesized value. Thus, only one of the “tails” of the bell-shaped normal curve is relevant. For instance, the following hypothesis could be appropriately examined with a one-tailed test:

$H_1$ : The number of pizza restaurants within a postal code in Florida is greater than five.

In this case, if the observed value is significantly less than five, the hypothesis is still not supported. Practically, a one-tailed test can be determined from a two-tailed test result by taking half of the observed  $p$ -value. When the researcher has any doubt about whether a one- or two-tailed test is appropriate, he or she should opt for the less conservative two-tailed test. Most computer software will assume a two-tailed test unless otherwise specified.



## Univariate Hypothesis Test Using the $t$ -Distribution

The step-by-step procedure for a  $t$ -test is conceptually similar to that for hypothesis testing with the  $Z$ -distribution. Suppose a Pizza-In store manager believes that the average number of customers who return a pizza or ask for a refund is 20 per day. The store records the number of returns and exchanges for each of the 25 days it was open during a given month. Are the return/complaint observations different than 20 per day? The substantive hypothesis is

$$H_1: \mu \neq 20$$

1. The researcher calculates a sample mean and standard deviation. In this case,  $\bar{X} = 22$  and  $S$  (sample standard deviation) = 5.
2. The standard error is computed ( $S_{\bar{X}}$ ):

$$\begin{aligned} S_{\bar{X}} &= \frac{S}{\sqrt{n}} \\ &= \frac{5}{\sqrt{25}} \\ &= 1 \end{aligned}$$

3. The researcher then finds the  $t$ -value associated with the desired level of confidence level or statistical significance. If a 95 percent confidence level is desired, the significance level is 0.05.
4. The critical values for the  $t$ -test are found by locating the upper and lower limits of the confidence interval. The result defines the regions of rejection. This requires determining the value of  $t$ . For 24 degrees of freedom ( $n = 25$ ,  $df = n - 1$ ), the  $t$ -value is 2.064. The critical values are

$$\begin{aligned} \text{Lower limit} &= \mu - t_{c.l.} S_{\bar{X}} = 20 - 2.064 \left( \frac{5}{\sqrt{25}} \right) \\ &= 20 - 2.064(1) \\ &= 17.936 \\ \text{Upper limit} &= \mu + t_{c.l.} S_{\bar{X}} = 20 + 2.064 \left( \frac{5}{\sqrt{25}} \right) \\ &= 20 + 2.064(1) \\ &= 22.064 \end{aligned}$$

Finally, the researcher makes the statistical decision by determining whether the sample mean falls between the critical limits. For the pizza store sample,  $\bar{X} = 22$ . The sample mean is *not* included in the region of rejection. Even though the sample result is only slightly less than the critical value at the upper limit, the null hypothesis cannot be rejected. In other words, the pizza store manager's assumption appears to be correct.

As with the  $Z$ -test, there is an alternative way to test a hypothesis with the  $t$ -statistic. This is by using the formula

$$\begin{aligned} t_{\text{obs}} &= \frac{\bar{X} - \mu}{S_{\bar{X}}} \\ t_{\text{obs}} &= \frac{22 - 20}{1} = \frac{2}{1} = 2 \end{aligned}$$

We can see that the observed  $t$ -value is less than the critical  $t$ -value of 2.064 at the 0.05 level when there are  $25 - 1 = 24$  degrees of freedom. As a result, the  $p$ -value is greater than 0.05 and the hypothesis is not supported. We cannot conclude with 95 percent confidence that the mean is not 20.

## The Chi-Square Test for Goodness of Fit

A **chi-square ( $\chi^2$ ) test** is one of the most basic tests for statistical significance and is particularly appropriate for testing hypotheses about frequencies arranged in a frequency or contingency table. Univariate tests involving nominal or ordinal variables are examined with a  $\chi^2$ . More generally, the  $\chi^2$  test is associated with **goodness-of-fit (GOF)**. GOF can be thought of as how well some matrix (table) of numbers matches or *fits* another matrix of the same size. Most often, the test is between a table of observed frequency counts and another table of expected values (central tendency) for those counts.

In statistical terms, a  $\chi^2$  test determines whether the difference between an observed frequency distribution and the corresponding expected frequency distribution is due to sampling variation. Computing a  $\chi^2$  test is fairly straightforward and easy. Students who master this calculation should have little trouble understanding future significance tests since the basic logic of the  $\chi^2$  test underlies these tests as well.

The steps in computing a  $\chi^2$  test are as follows:

1. Gather data and tally the observed frequencies for the categorical variable.
2. Compute the expected values for each value of the categorical variable.
3. Calculate the  $\chi^2$  value, using the observed frequencies from the sample and the expected frequencies.
4. Find the degrees of freedom for the test.
5. Make the statistical decision by comparing p-value associated with the calculated  $\chi^2$  against the predetermined significance level (acceptable Type I error rate).

Consider the following hypothesis that relates back to the chapter vignette:

*H<sub>p</sub>. Papa John's Pizza stores are more likely to be located in a stand-alone location than in a shopping center.*

A competitor may be interested in this hypothesis as part of the competitor analysis in a marketing plan. A researcher for the competitor gathers a random sample of 100 Papa John's locations in California (where the competitor is located). The sample is selected from phone directories and the locations are checked by having an assistant drive to each location. The following observations are recorded in a frequency table.

Location	One-Way Frequency Table
Stand-Alone	60 stores
Shopping Center	40 stores
Total	100 stores

The next step asks, "What are the expected frequencies for the location variable? We would expect that half (50) of the locations would be stand-alone and half (50) would be in a shopping center. This is another way of saying that the expected probability of being one type of location is 50 percent. The expected values also can be placed in a frequency table:

Location	Expected Frequencies
Stand-Alone	$100/2 = 50$ stores
Shopping Center	$100/2 = 50$ stores
Total	100 stores

- The actual  $\chi^2$  value is computed using the following formula:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where

$\chi^2$  = chi-square statistic

$O_i$  = observed frequency in the  $i$ th cell

$E_i$  = expected frequency in the  $i$ th cell

Sum the squared differences:

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

Thus, we determine that the chi-square value equals 4:

$$\begin{aligned}\chi^2 &= \frac{(60 - 50)^2}{50} + \frac{(40 - 50)^2}{50} \\ &= 4\end{aligned}$$

## The t-Test for Comparing Two Means

### Independent Samples $t$ -Test

Most typically, the researcher will apply the **independent samples  $t$ -test**, which tests the differences between means taken from two independent samples or groups.

#### ■ INDEPENDENT SAMPLES $t$ -TEST CALCULATION

The  $t$ -test actually tests whether or not the differences between two means is zero. Not surprisingly, this idea can be expressed as the difference between two population means:

$$\mu_1 = \mu_2, \text{ which is equivalent to, } \mu_1 - \mu_2 = 0$$

However, since this is inferential statistics, we test the idea by comparing two sample means ( $\bar{X}_1 - \bar{X}_2$ ).

A verbal expression of the formula for  $t$  is

$$t = \frac{\text{Sample Mean 1} - \text{Sample Mean 2}}{\text{Variability of random means}}$$

In almost all situations, we will see from the calculation of the two sample means that they are not exactly equal. The question is actually whether the observed differences have occurred by chance, or likely exist in the population. The  $t$ -value is a ratio with information about the difference between means (provided by the sample) in the numerator and the standard error in the denominator. To calculate  $t$ , we use the following formula:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}}$$

where

$\bar{X}_1$  = mean for group 1

$\bar{X}_2$  = mean for group 2

$S_{\bar{X}_1 - \bar{X}_2}$  = pooled, or combined, standard error of difference between means

## Paired-Samples t-Test

What happens when means need to be compared that are not from independent samples? Such might be the case when the same respondent is measured twice—for instance, when the respondent is asked to rate both how much he or she likes shopping on the Internet and how much he or she likes shopping in traditional stores. Since the liking scores are both provided by the same person, the assumption that they are independent is not realistic. Additionally, if one compares the prices the same retailers charge in their stores with the prices they charge on their Web sites, the samples cannot be considered independent because each pair of observations is from the same sampling unit.

A **paired-samples t-test** is appropriate in this situation. The idea behind the paired-samples *t-test* can be seen in the following computation:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

where  $\bar{d}$  is the difference between means,  $s_d$  is the standard deviation of the observed differences,  $d$  and  $n$  is the number of observations. Researchers also can compute the paired-samples *t-test* using statistical software. For example, using SPSS, the click-through sequence would be:

Analyze → Compare Means → Paired-Samples *t-test*

A dialog box then appears in which the “paired variables” should be entered. When a paired-samples *t-test* is appropriate, the two numbers being compared are usually scored as separate variables.

## The Z-Test for Comparing Two Proportions

What type of statistical comparison can be made when the observed statistics are proportions? Suppose a researcher wishes to test the hypothesis that wholesalers in the northern and southern United States differ in the proportion of sales they make to discount retailers. Testing whether the population proportion for group 1 ( $p_1$ ) equals the population proportion for group 2 ( $p_2$ ) is conceptually the same as the *t-test* of two means. This section illustrates a **Z-test for differences of proportions**, which requires a sample size greater than 30.

The test is appropriate for a hypothesis of this form:

$$H_0: \pi_1 = \pi_2$$

which may be restated as

$$H_0: \pi_1 - \pi_2 = 0$$

Comparison of the observed sample proportions  $p_1$  and  $p_2$  allows the researcher to ask whether the difference between two *large* (greater than 30) random samples occurred due to chance alone. The Z-test statistic can be computed using the following formula:

$$Z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{S_{p_1-p_2}}$$

where

$p_1$  = sample proportion of successes in group 1

$p_2$  = sample proportion of successes in group 2

$\pi_1 - \pi_2$  = hypothesized population proportion 1 minus hypothesized population proportion 2

$S_{p_1-p_2}$  = pooled estimate of the standard error of differences in proportions

The statistic normally works on the assumption that the value of  $\pi_1 - \pi_2$  is zero, so this formula is actually much simpler than it looks at first inspection. Readers also may notice the similarity between this and the paired-samples  $t$ -test.

To calculate the standard error of the differences in proportions, use the formula

$$S_{p_1-p_2} = \sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where

$\bar{p}$  = pooled estimate of proportion of successes in a sample

$\bar{q} = 1 - \bar{p}$ , or pooled estimate of proportion of failures in a sample

$n_1$  = sample size for group 1

$n_2$  = sample size for group 2

To calculate the pooled estimator,  $\bar{p}$ , use the formula

$$\bar{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

Suppose the survey data are as follows:

Northern Wholesalers	Southern Wholesalers
$p_1 = 0.35$	$p_2 = 0.40$
$n_1 = 100$	$n_2 = 100$

First, the standard error of the difference in proportions is

$$\begin{aligned} S_{p_1-p_2} &= \sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \\ &= \sqrt{(0.375)(0.625)\left(\frac{1}{100} + \frac{1}{100}\right)} = 0.068 \end{aligned}$$

where

$$\bar{p} = \frac{(100)(0.35) + (100)(0.40)}{100 + 100} = 0.375$$

If we wish to test the two-tailed question of no difference, we must calculate an observed  $Z$ -value. Thus,

$$\begin{aligned}
 Z &= \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{S_{p_1 - p_2}} \\
 &= \frac{(0.35 - 0.40) - (0)}{0.068} \\
 &= -0.73
 \end{aligned}$$

In this example the idea that the proportion of sales differs by region is not supported. The calculated Z-value is less than the critical Z-value of 1.96. Therefore, the p-value associated with the test is greater than 0.05.

## What Is ANOVA?

So far, we have discussed tests for differences between two groups. However, what happens when we have more than two groups? For example, what if we want to test and see if employee turnover differs across our five production plants? When the means of more than two groups or populations are to be compared, one-way **analysis of variance (ANOVA)** is the appropriate statistical tool. ANOVA involving only one grouping variable is often referred to as *one-way ANOVA* because only one independent variable is involved. Another way to define ANOVA is as the appropriate statistical technique to examine the effect of a less-than interval independent variable on an at-least interval dependent variable. Thus, a categorical independent variable and a continuous dependent variable are involved. An independent samples *t*-test can be thought of as a special case of ANOVA in which the independent variable has only two levels. When more levels exist, the *t*-test alone cannot handle the problem.

The statistical null hypothesis for ANOVA is stated as follows:

$$\mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

The symbol k is the number of groups or categories for an independent variable. In other words, all group means are equal. The substantive hypothesis tested in ANOVA is

*At least one group mean is not equal to another group mean.*

As the term *analysis of variance* suggests, the problem requires comparing variances to make inferences about the means.

The Papa Johns example considered locations that were stand-alone and shopping center, compared to the categorical variable of profitable or not profitable. However, if we knew the exact amount of profit or loss for each store, this becomes a good example of a *t*-test. Specifically, the independent variable could be thought of as “location,” meaning either stand-alone or shopping center. The dependent variable is the amount of profit/loss. Since only two groups exist for the independent variable, either an independent samples *t*-test or one-way ANOVA could be used. The results would be identical.

## Simple Illustration of ANOVA

ANOVA’s logic is fairly simple. Look at the data table below that describes how much coffee respondents report drinking each day based on which shift they work, day shift, second shift, or nights.

Day 1  
Day 3  
Day 4  
Day 0  
Day 2



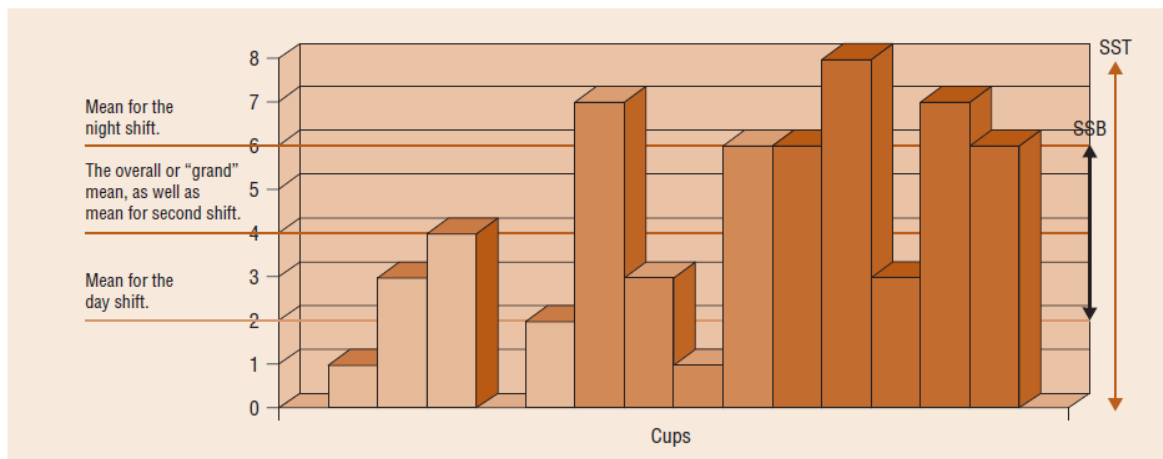
Second 7  
 Second 2  
 Second 1  
 Second 6  
 Night 6  
 Night 8  
 Night 3  
 Night 7  
 Night 6

The following table displays the means for each group and the overall mean:

Shift	Mean	Std. Deviation	N
Day	2.00	1.58	5
Second	4.00	2.94	4
Night	6.00	1.87	5
Total	4.00	2.63	14

Exhibit 22.5 plots each observation with a bar. The long vertical line illustrates the total range of observations. The lowest is 0 cups and the highest is 8 cups of coffee for a range of 8. The overall mean is 4 cups. Each group mean is shown with a different colored line that matches the bars corresponding to the group. The day shift averages 2 cups of coffee a day, the second shift 4 cups, and the night shift 6 cups of coffee per day.

EXHIBIT 22.5 Illustration of ANOVA Logic



Here is the basic idea of ANOVA. Look at the dark double-headed arrow in Exhibit 22.5. This line represents the range of the differences between group means. In this case, the lowest mean is 2 cups and the highest mean is 6 cups. Thus, the middle vertical line corresponds to the total variation (range) in the data and the thick double-headed black line corresponds to the variance accounted for by the group differences. As the thick black line accounts for more of the total variance, then the ANOVA model suggests that the group means are not all the same, and in particular, not all the same as the overall mean. This also means that the independent variable, in this case work shift, explains

the dependent variable. Here, the results suggest that knowing when someone works explains how much coffee they drink. Night-shift workers drink the most coffee.

## Partitioning Variance in ANOVA

### • TOTAL VARIABILITY

An implicit question with the use of ANOVA is, “How can the dependent variable best be predicted?” Absent any additional information, the error in predicting an observation is minimized by choosing the central tendency, or mean for an interval variable. For the coffee example, if no information was available about the work shift of each respondent, the best guess for coffee drinking consumption would be four cups. The Sum of Squares Total (SST) or variability that would result from using the **grand mean**, meaning the mean over all observations, can be thought of as

$$SST = \text{Total of } (\text{Observed Value} - \text{Grand Mean})^2$$

Although the term error is used, this really represents how much total variation exists among the measures.

Using the first observation, the error of observation would be

$$(1 \text{ cup} - 4 \text{ cups})^2 = 9$$

The same squared error could be computed for each observation and these squared errors totaled to give SST.

### ■ BETWEEN-GROUPS VARIANCE

ANOVA tests whether “grouping” observations explains variance in the dependent variable. In Exhibit 22.5, the three colors reflect three levels of the independent variable, work shift. Given this additional information about which shift a respondent works, the prediction changes. Now, instead of guessing the grand mean, the group mean would be used. So, once we know that someone works the day shift, the prediction would be that he or she consumes 2 cups of coffee per day. Similarly, the second and night-shift predictions would be 4 and 6 cups, respectively. Thus, the **between-groups variance** or Sum of Squares Between-groups (SSB) can be found by taking the total sum of the weighted difference between group means and the overall mean as shown:

$$SSB = \text{Total of } n (\text{Group Mean} - \text{Grand Mean})^2$$

The weighting factor ( $n_{\text{group}}$ ) is the specific group sample size. Let’s consider the first observation once again. Since this observation is in the day shift, we predict 2 cups of coffee will be consumed. Looking at the day shift group observations in Exhibit 22.5, the new error in prediction would be

$$(2 \text{ cups} - 4 \text{ cups})^2 = (2)^2 = 4$$

The error in prediction has been reduced from 3 using the grand mean to 2 using the group mean. This squared difference would be weighted by the group sample size of 5, to yield a contribution to SSB of 20. Next, the same process could be followed for the other groups yielding two more contributions to SSB. Because the second shift group mean is the same as the grand mean, that group’s contribution to SSB is 0. Notice that the night-shift group mean is also 2 different than the grand mean, like the day shift, so this group’s contribution to SSB is likewise 20. The total SSB then represents the variation explained by the experimental or independent variable. In this case, total SSB is 40.



### • WITHIN-GROUP ERROR

Finally, error within each group would remain. Whereas the group means explain the variation between the total mean and the group mean, the distance from the group mean and each individual observation remains unexplained. This distance is called **within-group error or variance** or the Sum of Squares Error (SSE). The values for each observation can be found by

$$\text{SSE} = \text{Total of (Observed Mean — Group Mean)}^2$$

Again, looking at the first observation, the SSE component would be

$$\text{SSE} = (1 \text{ cup} — 2 \text{ cups})^2 = 1 \text{ cup}$$

This process could be computed for all observations and then totaled. The result would be the total error variance—a name used to refer to SSE since it is variability not accounted for by the group means. These three components are used in determining how well an ANOVA model explains a dependent variable.

### The F-Test

The **F-test** is the key statistical test for an ANOVA model. The F-test determines whether there is more variability in the scores of one sample than in the scores of another sample. The key question is whether the two sample variances are different from each other or whether they are from the same population. Thus, the test breaks down the variance in a total sample and illustrates why ANOVA is *analysis of variance*.

The F-statistic (or F-ratio) can be obtained by taking the larger sample variance and dividing by the smaller sample variance.

### ■ USING VARIANCE COMPONENTS TO COMPUTE F-RATIOS

In ANOVA, the basic consideration for the F-test is identifying the relative size of variance components. The three forms of variation described briefly above are:

1. SSE—variation of scores due to random error or within-group variance due to individual differences from the group mean. This is the error of prediction.
2. SSB—systematic variation of scores between groups due to manipulation of an experimental variable or group classifications of a measured independent variable or between-groups variance.
3. SST—the total observed variation across all groups and individual observations.

Thus, we can partition **total variability** into *within-group variance* (SSE) and *between-groups variance* (SSB).

The *F*-distribution is a function of the ratio of these two sources of variances:

$$F = f\left(\frac{SSB}{SSE}\right)$$

A larger ratio of variance between groups to variance within groups implies a greater value of *F*. If the *F*-value is large, the results are likely to be statistically significant.

# MULTIVARIATE STATISTICAL ANALYSIS

## What Is Multivariate Data Analysis?

Research that involves three or more variables, or that is concerned with underlying dimensions among multiple variables, will involve multivariate statistical analysis. Multivariate statistical methods analyze multiple variables or even multiple sets of variables simultaneously. How do we know when someone has experienced nostalgia and whether or not the experience has altered behavior? Nostalgia itself is a latent construct that involves multiple indicators that together represent nostalgia. As such, the measurement and outcomes of nostalgia lend themselves well to multivariate analysis. Likewise, many other business problems involve multivariate data analysis including most employee motivation research, customer psychographic profiles, and research that seeks to identify viable market segments.

## The "Variate" in Multivariate

Another distinguishing characteristic of multivariate analysis is the variate. The variate is a mathematical way in which a set of variables can be represented with one equation. A variate is formed as a linear combination of variables, each contributing to the overall meaning of the variate based upon an empirically derived weight. Mathematically, the variate is a function of the measured variables involved in an analysis:

$$V_k = f(X_1, X_2, \dots, X_m)$$

$V_k$  is the  $k$ th variate. Every analysis could involve multiple sets of variables, each represented by a variate.  $X_1$  to  $X_m$  represent the measured variables.

Here is a simple illustration. Recall that constructs are distinguished from variables by the fact that multiple variables are needed to measure a construct. Let's assume we measured nostalgia with five questions on our survey. With these five variables, a variate of the following form could be created:

$$V_k = L_1X_1 + L_2X_2 + L_3X_3 + L_4X_4 + L_5X_5$$

$V_k$  represents the score for nostalgia,  $X_1$  to  $X_5$  represent the observed scores on the five scale items (survey questions) that are expected to indicate nostalgia, and  $L_1$  to  $L_5$  are parameter estimates much like regression weights that suggest how highly related each variable is to the overall nostalgia score.

## Classifying Multivariate Techniques

Two basic groups of multivariate techniques are *dependence methods* and *interdependence methods*

### Dependence Techniques

When hypotheses involve distinction between independent and dependent variables, dependence techniques are needed. For instance, when we hypothesize that nostalgia is related positively to purchase intentions, nostalgia takes on the character of an independent variable and purchase intentions take on the character of a dependent variable. Predicting the dependent variable "sales" on the basis of numerous independent variables is a problem frequently investigated with dependence techniques. *Multiple regression analysis*, *multiple discriminant analysis*, *multivariate analysis of variance*, and *structural equations modeling* are all dependence methods.

## Interdependence Techniques

When researchers examine questions that do not distinguish between independent and dependent variables, interdependence techniques are used. No one variable or variable subset is to be predicted from or explained by the others. The most common interdependence methods are *factor analysis*, *cluster analysis*, and *multidimensional scaling*. A manager might utilize these techniques to determine which employee motivation items tend to group together (factor analysis), to identify profitable customer market segments (cluster analysis), or to provide a perceptual map of cities being considered for a new plant (multidimensional scaling).

## Analysis of Dependence

Multivariate dependence techniques are variants of the general linear model (GLM). Simply, the GLM is a way of modeling some process based on how different variables cause fluctuations from the average dependent variable. Fluctuations can come in the form of group means that differ from the overall mean as in ANOVA or in the form of a significant slope coefficient as in regression. The basic idea can be thought of as follows:

$$\hat{Y}_i = \mu + \Delta X + \Delta F + \Delta XF$$

Here,  $\mu$  represents a constant, which can be thought of as the overall mean of the dependent variable,  $\Delta X$  and  $\Delta F$  represent changes due to main effect independent variables (such as experimental variables) and blocking independent variables (such as covariates or grouping variables), respectively, and  $\Delta XF$  represents the change due to the combination (interaction effect) of those variables. Realize that  $Y_i$  in this case could represent multiple dependent variables, just as  $X$  and  $F$  could represent multiple independent variables. Multiple regression analysis, n-way ANOVA, and MANOVA represent common forms that the GLM can take.

## Multiple Regression Analysis

Multiple regression analysis is an extension of simple regression analysis allowing a metric dependent variable to be predicted by multiple independent variables. Thus, one dependent variable (sales volume) is explained by one independent variable (number of building permits). Yet reality is more complicated and several additional factors probably affect construction equipment sales. Other plausible independent variables include price, seasonality, interest rates, advertising intensity, consumer income, and other economic factors in the area. The simple regression equation can be expanded to represent multiple regression analysis:

$$Y_i = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_nX_n + e_i$$

Thus, as a form of the GLM, dependent variable predictions ( $\hat{Y}$ ) are made by adjusting the constant ( $b_0$ , which would be equal to the mean if all slope coefficients are 0, based on the slope coefficients associated with each independent variable ( $b_1, b_2, \dots, b_n$ )).<sup>10</sup>

Less-than interval (nonmetric) independent variables can be used in multiple regression. This can be done by implementing dummy variable coding. A dummy variable is a variable that uses a 0 and a 1 to code the different levels of dichotomous variable (for instance, residential or commercial building permit). Multiple dummy variables can be included in a regression model.

## ■ REGRESSION COEFFICIENTS IN MULTIPLE REGRESSION

Recall that in simple regression, the coefficient  $b_1$  represents the slope of  $X$  on  $Y$ . Multiple regression involves multiple slope estimates, or regression weights. One challenge in regression models is to understand how one independent variable affects the dependent variable, considering the effect of other independent variables. When the independent variables are related to each other, the regression weight associated with one independent variable is affected by the regression weight of another. Regression coefficients are unaffected by each other only when independent variables are totally independent.

Conventional regression programs can provide standardized parameter estimates,  $\beta_1$ ,  $\beta_2$  and so on, that can be thought of as *partial regression coefficients*. The correlation between  $Y$  and  $X_1$ , controlling for the correlation that  $X_2$  has with the  $Y$ , is called partial correlation. Consider a standardized regression model with only two independent variables:

$$Y = \beta_1 X_1 + \beta_2 X_2 + e_i$$

The coefficients  $\beta_1$ ,  $\beta_2$  are partial regression coefficients, which express the relationship between the independent variable and dependent variable taking into consideration that the other variable also is related to the dependent variable. As long as the correlation between independent variables is modest, partial regression coefficients adequately represent the relationships. When the correlation between two independent variables becomes high, the regression coefficients may not be reliable.

When researchers want to know which independent variable is most predictive of the dependent variable, the standardized regression coefficient ( $\beta$ ) is used. One huge advantage of  $\beta$  is that it provides a constant scale. In other words, the  $\beta$ s are directly comparable. Therefore, the greater the absolute value of the standardized regression coefficient, the more that particular independent variable is responsible for explaining the dependent variable.

## ■ $R^2$ IN MULTIPLE REGRESSION

The coefficient of multiple determination in multiple regression indicates the percentage of variation in  $Y$  explained by the combination of *all* independent variables. For example, a value of  $R^2 = 0.845$  means that 84.5 percent of the variance in the dependent variable is explained by the independent variables. If two independent variables are truly independent (uncorrelated with each other), the  $R^2$  for a multiple regression model is equal to the separate  $R^2$  values that would result from two separate simple regression models. More typically, the independent variables are at least moderately related to one another, meaning that the model  $R^2$  from a multiple regression model will be less than the separate  $R^2$  values resulting from individual simple regression models. This reduction in  $R^2$  is proportionate to the extent to which the independent variables exhibit multicollinearity.

## ■ STATISTICAL SIGNIFICANCE IN MULTIPLE REGRESSION

Following from simple regression, an  $F$ -test is used to test statistical significance by comparing the variation explained by the regression equation to the residual error variation. The  $F$ -test allows for testing of the relative magnitudes of the sum of squares due to the regression ( $SSR$ ) and the error sum of squares ( $SSE$ ).

$$F = \frac{(SSR)/k}{(SSE)/(n - k - 1)} = \frac{MSR}{MSE}$$

where

$k$  = number of independent variables

$n$  = number of observations

$MSR$  = Mean Squares Regression

$MSE$  = Mean Squares Error

Degrees of freedom for the  $F$ -test ( $df$ ) are:

$df$  for the numerator =  $k$

$df$  for the denominator =  $n - k - 1$

For our toy sales example,

$df$  (numerator) = 3

$df$  (denominator) =  $24 - 3 - 1 = 20$

### ■ STEPS IN INTERPRETING A MULTIPLE REGRESSION MODEL

Multiple regression models often are used to test some proposed theoretical model. For instance, a researcher may be asked to develop and test a model explaining business unit performance. Why do some business units outperform others? Multiple regression models can be interpreted using these steps:

1. Examine the model  $F$ -test. If the test result is not significant, the model should be dismissed and there is no need to proceed to further steps.
2. Examine the individual statistical tests for each parameter estimate. An independent variable with significant results can be considered a significant explanatory variable. If an independent variable is not significant, the model should be run again with nonsignificant predictors deleted. Often, it is best to eliminate predictor variables one at a time, then rerun the reduced model.
3. Examine the model  $R^2$ . No cutoff values exist that can distinguish an acceptable amount of explained variation across all regression models. However, the absolute value of  $R^2$  is more important when the researcher is more interested in prediction than in explanation. In other words, the regression is run for pure forecasting purposes. When the model is more oriented toward explanation of which variables are most important in explaining the dependent variable, cutoff values for the model  $R^2$  are not really appropriate.

Examine collinearity diagnostics. Multicollinearity in regression analysis refers to how strongly interrelated the independent variables in a model are. When multicollinearity is too high, the individual parameter estimates become difficult to interpret. Most regression programs can compute variance inflation factors (VIF) for each variable. As a rule of thumb, VIF above 5.0 suggests problems with multicollinearity.

### ■ N-WAY (UNIVARIATE) ANOVA

The interpretation of an  $n$ -way ANOVA model follows closely from the regression results described above. The steps involved are essentially the same with the addition of interpreting differences between means:

1. Examine the overall model  $F$ -test result. If significant, proceed.
2. Examine individual  $F$ -tests for each individual independent variable.
3. For each significant categorical independent variable, interpret the effect by examining the group means.

4. For each significant continuous variable (covariate), interpret the parameter estimate (b).
5. For each significant interaction, interpret the means for each combination.

## ■ INTERPRETING MANOVA

Compared to ANOVA, a MANOVA model produces an additional layer of testing. The first layer of testing involves the multivariate F-test, which is based on a statistic called Wilke's Lambda (A). This test examines whether or not an independent variable explains significant variation among the dependent variables within the model. If this test is significant, then the F-test results from individual univariate regression models nested within the MANOVA model are interpreted. The rest of the interpretation results follow from the one-way ANOVA or multiple regression model results above. The Research Snapshot on the next page provides an example of how to run and interpret MANOVA.

## Discriminant Analysis

Researchers often need to produce a classification of sampling units. This process may involve using a set of independent variables to decide if a sampling unit belongs in one group or another. A physician might record a person's blood pressure, weight, and blood cholesterol level and then categorize that person as having a high or low probability of a heart attack. A researcher interested in retailing failures might be able to group firms as to whether they eventually failed or did not fail on the basis of independent variables such as location, financial ratios, or management changes. A bank might want to discriminate between potentially successful and unsuccessful sites for electronic fund transfer system machines. A human resource manager might want to distinguish between applicants to hire and those not to hire. The challenge is to find the discriminating variables to use in a predictive equation that will produce *better than chance* assignment of the individuals to the correct group.

Discriminant analysis is a multivariate technique that predicts a categorical dependent variable (rather than a continuous, interval-scaled variable, as in multiple regression) based on a linear combination of independent variables. In each problem above, the researcher determines which variables explain why an observation falls into one of two or more groups. A linear combination of independent variables that explains group memberships is known as a discriminant function. Discriminant analysis is a statistical tool for determining such linear combinations. The researcher's task is to derive the coefficients of the discriminant function (a straight line).

We will consider an example of the two-group discriminant analysis problem where the dependent variable, Y, is measured on a nominal scale. (Although n-way discriminant analysis is possible, it is beyond the scope of this discussion.) Suppose a personnel manager for an electrical wholesaler has been keeping records on successful versus unsuccessful sales employees. The personnel manager believes it is possible to predict whether an applicant will succeed on the basis of age, sales aptitude test scores, and mechanical ability scores. As stated at the outset, the problem is to find a linear function of the independent variables that shows large differences in group means. The first task is to estimate the coefficients of the applicant's discriminant function. To calculate the individuals' discriminant scores, the following linear function is used.



$$Z_i = b_1X_{1i} + b_2X_{2i} + \cdots + b_nX_{ni}$$

where

$Z_i$  =  $i$ th applicant's discriminant score

$b_n$  = discriminant coefficient for the  $n$ th variable

$X_{ni}$  =  $i$ th applicant's value on the  $n$ th independent variable

Using scores for all the individuals in the sample, a discriminant function is determined based on the criterion that the groups be maximally differentiated on the set of independent variables.

## Analysis of Interdependence

Rather than attempting to predict a variable or set of variables from a set of independent variables, we use techniques like *factor analysis*, *cluster analysis*, and *multidimensional scaling* to better understand the relationships and structure among a set of variables or objects.

### 1-Factor Analysis

Factor analysis is a prototypical multivariate, interdependence technique. Factor analysis is a technique of statistically identifying a reduced number of factors from a larger number of measured variables. The factors themselves are not measured, but instead, they are identified by forming a variate using the measured variables. Factors are usually latent constructs like attitude or satisfaction, or an index like social class. A researcher need not distinguish between independent and dependent variables to conduct factor analysis. Factor analysis can be divided into two types:

1. Exploratory factor analysis (EFA)—performed when the researcher is uncertain about how many factors may exist among a set of variables. The discussion here concentrates primarily on EFA.
2. Confirmatory factor analysis (CFA)—performed when the researcher has strong theoretical expectations about the factor structure (number of factors and which variables relate to each factor) before performing the analysis. CFA is a good tool for assessing construct validity because it provides a test of how well the researcher's "theory" about the factor structure fits the actual observations. Many books exist on CFA alone and the reader is advised to refer to any of those sources for more on CFA.

Exhibit 24.6 illustrates factor analysis graphically. Suppose a researcher is asked to examine how feelings of nostalgia in a restaurant influence customer loyalty. Three hundred fifty customers at themed restaurants around the country are interviewed and asked to respond to the following Likert scales (1 = Strongly Disagree to 7 = Strongly Agree):

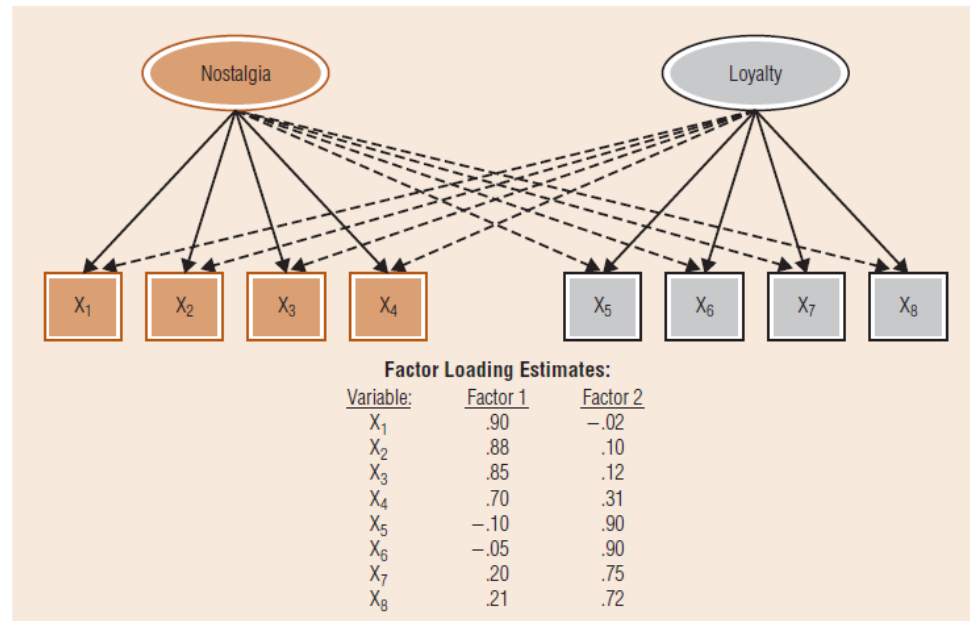
- $X_1$ —I feel a strong connection to the past when I am in this place.
- $X_2$ —This place evokes memories of the past.
- $X_3$ —I feel a yearning to relive past experiences when I dine here.
- $X_4$ —This place looks like a page out of the past.
- $X_5$ —I am willing to pay more to dine in this restaurant.
- $X_6$ —I feel very loyal to this establishment.
- $X_7$ —I would recommend this place to others.
- $X_8$ —I will go out of my way to dine here.

Factor analysis can summarize the information in the eight variables in a smaller number of variables. How many dimensions, or groups of variables, are likely present in this case? More than one technique

exists for estimating the variates that form the factors. In this example, the factor analysis indicates there are two dimensions, or factors, as shown in Exhibit 24.6. Thus, EFA provides two important pieces of information:

1. How many factors exist among a set of variables?
2. What variables are related to or “load on” which factors?

**EXHIBIT 24.6**  
**A Simple Illustration of**  
**Factor Analysis**



## ■ HOW MANY FACTORS

One of the first questions the researcher asks is, “How many factors will exist among a large number of variables?” While a detailed discussion is beyond the scope of this text, the question is usually addressed based on the eigenvalues for a factor solution. Eigenvalues are a measure of how much variance is explained by each factor. The most common rule—and the default for most statistical programs—is to base the number of factors on the number of eigenvalues greater than 1.0. The basic thought is that a factor with an eigenvalue of 1.0 has the same total variance as one variable. It usually does not make sense to have factors, which are a combination of variables, that have less information than a single variable. So, unless some other rule is specified, the number of factors shown in a factor solution is based on this rule.

## ■ FACTOR LOADINGS

Each arrow connecting a factor (represented by an oval in Exhibit 24.6) to a variable (represented by a box in Exhibit 24.6) is associated with a factor loading. A factor loading indicates how strongly correlated a measured variable is with that factor. In other words, to what extent does a variable “load” on a factor? EFA depends on the loadings for proper interpretation. A latent construct can be interpreted based on the pattern of loadings and the content of the variables. In this way, the latent construct is measured indirectly by the variables.

Loading estimates are provided by factor analysis programs. In Exhibit 24.6, the factor loading estimates are shown beneath the factor diagram. The thick arrows indicate high loading estimates and the thin dotted lines correspond to weak loading estimates. Factors are interpreted by examining any patterns that emerge from the factor results. Here, a clear pattern emerges. The first four variables produce high



loadings on factor 1 and the last four variables produce high loadings on factor 2.

When a clear pattern of factor loadings emerges, interpretation is easy. Because the first four variables all have content consistent with nostalgia and the second four variables all have content consistent with customer loyalty, the two factors can easily be labeled. Factor one represents the latent construct nostalgia and factor 2 represents the latent construct customer loyalty.

### ■ FACTOR ROTATION

Factor rotation is a mathematical way of simplifying factor results. The most common type of factor rotation is a process called varimax. A discussion of the technical aspects of the concept of factor rotation is far beyond the scope of this book. However, factor rotation involves creating new reference axes for a given set of variables. An initial factor solution is often difficult to interpret. Rotation “clears things up” by producing more obvious patterns of loadings. Users can observe this by looking at the unrotated and rotated solutions in the factor analysis output. An example of how to run factor analysis is provided in the Research Snapshot above.

### ■ DATA REDUCTION TECHNIQUE

Factor analysis is considered a data reduction technique. Data reduction techniques allow a researcher to summarize information from many variables into a reduced set of variates or composite variables. Data reduction is advantageous for many reasons. In general, the rule of parsimony suggests that an explanation involving fewer components is better than one involving more. Factor analysis accomplishes data reduction by capturing variance from many variables with a single variate. Data reduction is also a way of identifying which variables among a large set might be important in some analysis. Thus, data reduction simplifies decision making.

In our example, the researcher can now form two composite factors representing the latent constructs nostalgia and customer loyalty. These can be formed using factor equations of this form:

$$F_k = L_1X_1 + L_2X_2 + L_3X_3 + L_4X_4 + L_5X_5 + L_6X_6 + L_7X_7 + L_8X_8$$

where

$F_k$  is the factor score for the  $k$ th factor (in this case there are two factors)

$L$  represents factor loadings (ith) 1 through 8 for the corresponding factor

$X$  represents the value of the corresponding measured variable

### ■ CREATING COMPOSITE SCALES WITH FACTOR RESULTS

When a clear pattern of loadings exists as in this case, the researcher may take a simpler approach.  $F_1$  could be created simply by summing the four variables with high loadings and creating a summated scale representing nostalgia.  $F_2$  could be created by summing the second four variables (those loading highly on  $F_2$ ) and creating a second summated variable. While not necessary, it is often wise to divide these summated scales by the number of items so the scale of the factor is the same as the original items. For example,  $F_1$  would be

$$((X_1 + X_2 + X_3 + X_4)/4)$$

The result provides a composite score on the 1—7 scale. The composite score approach would introduce very little error given the pattern of loadings. In other words, very low loadings suggest a variable does not contribute much to the factor. The reliability of each summated scale can be tested by computing a

coefficient alpha estimate. Then, the researcher could conduct a bivariate regression analysis that would test how much nostalgia contributed to loyalty.

### ■ COMMUNALITY

While factor loadings show the relationship between a variable and each factor, a researcher may also wish to know how much a single variable has in common with all factors. Communality is a measure of the percentage of a variable's variation that is explained by the factors. A relatively high communality indicates that a variable has much in common with the other variables taken as a group. A low communality means that the variable does not have a strong relationship with the other variables. The item might not be part of one of the common factors or might represent a separate dimension. Communality for any variable is equal to the sum of the squared loadings for that variable. The communality for  $X_1$  is

$$0.90^2 + 0.02^2 = 0.8104$$

Communality values are shown on factor analysis printouts.

### ■ TOTAL VARIANCE EXPLAINED

Along with the factor loadings, the percentage of total variance of original variables explained by the factors can be useful. Recall that common variance is correlation squared. Thus, if each loading is squared and totaled, that total divided by the number of factors provides an estimate of the variance in a set of variables explained by a factor. This explanation of variance is much the same as  $R^2$  in multiple regression. Again, these values are computed by the statistics program so there is seldom a need to compute them manually. In this case, though, the variance accounted for among the eight variables by the nostalgia factor is 0.36 and the variance among the eight variables explained by the loyalty factor is 0.35. Thus, the two factors explain 71 percent of the variance in the eight variables:

$$0.36 + 0.35 = 0.71$$

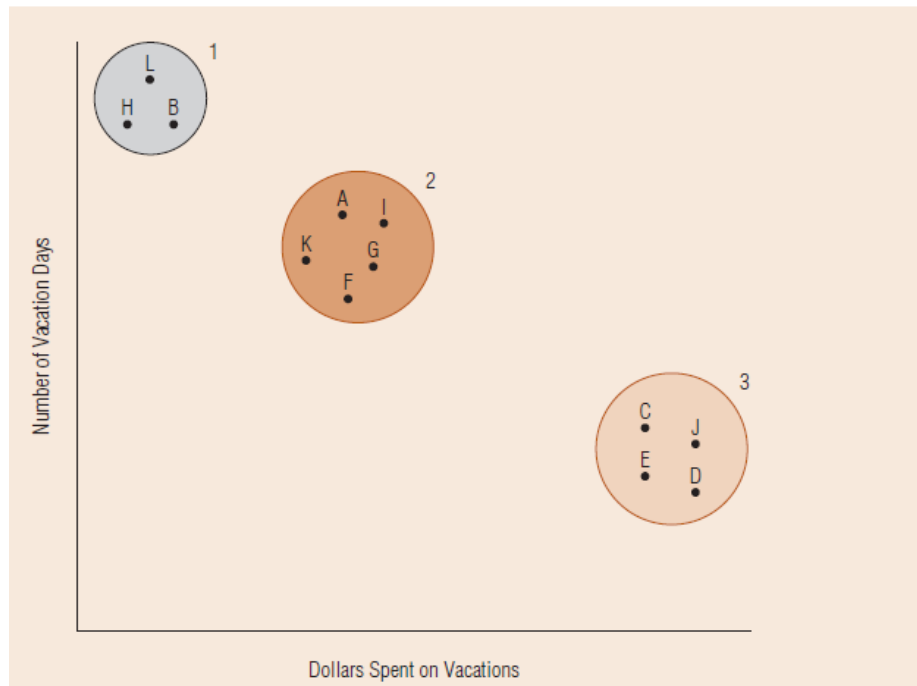
In other words, the researcher has 71% of the information in two factors that are in the original eight items, another example of the rule of parsimony.

## 2-Cluster Analysis

Cluster analysis is a multivariate approach for identifying objects or individuals that are similar to one another in some respect. Cluster analysis classifies individuals or objects into a small number of mutually exclusive and exhaustive groups. Objects or individuals are assigned to groups so that there is great similarity within groups and much less similarity between groups. The cluster should have high internal (within-cluster) homogeneity and high external (between-cluster) heterogeneity.

Cluster analysis is an important tool for the business researcher. For example, an organization may want to group its employees based on their insurance or retirement needs, or on job performance dimensions. Similarly, a business may wish to identify market segments by identifying subjects or individuals who have similar needs, lifestyles, or responses to marketing promotions. Clusters, or subgroups, of recreational vehicle owners may be identified on the basis of their similarity with respect to recreational vehicle usage and the benefits they want from recreational vehicles. Alternatively, the researcher might use demographic or lifestyle variables to group individuals into clusters identified as market segments.

We will illustrate cluster analysis with a hypothetical example relating to the types of vacations taken by 12 individuals. Vacation behavior is represented on two dimensions: number of vacation days and dollar expenditures on vacations during a given year. Exhibit 24.7 is a scatter diagram that represents the geometric distance between each individual in two-dimensional space. The diagram portrays three clear-cut clusters. The first subgroup—consisting of individuals L, H, and B—suggests a group of individuals who have many vacation days but do not spend much money on their vacations. The second cluster—represented by individuals A, I, K, G, and F—represents intermediate values on both variables: average amounts of vacation days and average dollar expenditures on vacations.



**EXHIBIT 24.7**  
**Clusters of Individuals on**  
**Two Dimensions**

The third group—individuals C, J, E, and D—consists of individuals who have relatively few vacation days but spend large amounts on vacations.

In this example, individuals are grouped on the basis of their similarity or proximity to one another. The logic of cluster analysis is to group individuals or objects by their similarity to or distance from each other. The mathematical procedures for deriving clusters will not be dealt with here, as our purpose is only to introduce the technique.

### Multidimensional Scaling

Multidimensional scaling provides a means for placing objects in multidimensional space on the basis of respondents' judgments of the similarity of objects. The perceptual difference among objects is reflected in the relative distance among objects in the multidimensional space.

In the most common form of multidimensional scaling, subjects are asked to evaluate an object's similarity to other objects. For example, a sports car study may ask respondents to rate the similarity of an Acura TSX to a Chevrolet Corvette, then an Acura NSX to the Corvette, followed by a Lotus Elise to the Corvette, a Mustang to the Corvette, and so forth. Then, the comparisons are rotated (i.e., Acura NSX to the TSX, Lotus Elise to the TSX, and so on until all pairs are exhausted). Multidimensional scaling would then generate a plot of the cars, and the analyst then attempts to explain the difference in objects on the basis of the plot. The interpretation of the plot is left to the researcher.

In one study MBA students were asked to provide their perceptions of relative similarities among six graduate schools.

Next, the overall similarity scores for all possible pairs of objects were aggregated for all individual respondents and arranged in a matrix. With the aid of a computer program, the judgments about similarity were statistically transformed into distances by placing the graduate schools into a specified multidimensional space. The distance between similar objects on the perceptual map was small for similar objects; dissimilar objects were farther apart.

Exhibit 24.9 on the next page shows a perceptual map in two dimensional space. Inspection of the map illustrates that Harvard and Stanford were perceived as quite similar to each other. MIT and Carnegie also were perceived as very similar. MIT and Harvard, on the other hand, were perceived as dissimilar. The researchers identified the two axes as “quantitative versus qualitative curriculum” and “less versus more prestige.” The labeling of the dimension axes is a task of interpretation for the researcher and is not statistically determined. As with other multivariate techniques in the analysis of interdependence, there are several alternative mathematical techniques for multidimensional scaling.

**EXHIBIT 24.9**  
**Perceptual Map of Six**  
**Graduate Business Schools:**  
**Simple Space**

