LEARNING OUTCOMES

# CHAPTER 20 BASIC DATA ANALYSIS: DESCRIPTIVE STATISTICS

After studying this chapter, you should be able to

1. Know what descriptive statistics are and why they are used
2. Create and interpret simple tabulation tables
3. Understand how cross-tabulations can reveal relationships
4. Perform basic data transformations
5. List different computer software products designed for descriptive statistical analysis
6. Understand a researcher's role in interpreting the data

## Chapter Vignette: Choose Your "Poison"

Most Americans enjoy an adult beverage occasionally. But not all Americans like the same drink. Many decision makers are interested in what Americans like to drink. Retailers need to have the correct product mix for their particular customers if profits are to be increased and customers made more satisfied. Restaurants need to know what their customers like to have with the types of food they serve. Policy makers need to know what types of restrictions should be placed on what types of products to prevent underage drinking and alcohol abuse. Researchers could apply sophisticated statistics to address questions related to Americans' drinking preferences, but a lot can be learned from just counting what people are buying.

A grocery store built in 1975 in Chicago allocates 15 percent of their floor space to adult beverage products. Out of this 15 percent, 60 percent is allocated to beer, 25 percent to spirits, and 15 percent to wine. Since the products are not merchandised the same way (different types of shelving, aisles, and racking are needed), adjusting the floor space to change these percentages is not an easy task. Over the three-decade history of the store, the customer base has changed. Originally, stay-at-home moms buying groceries for the family best characterized the customer base. During the 1990s, empty-nesters, including retirees with high disposable incomes, characterized the customer base. More recently, younger singles just starting careers have moved into the nearby neighborhoods. Should the store reconsider its adult beverage merchandising?

© STEPHEN OLIVER/DORLING KINDERSLEY/GETTY IMAGES

In 1992, American consumers showed a heavy preference toward beer. Among American adults who drank adult beverages,[1]

- 47 percent drank beer
- 21 percent drank spirits
- 27 percent drank wine

By 2005, Americans had changed their drinking preferences. Now,

- 36 percent drink beer
- 21 percent drink spirits
- 39 percent drink wine

A couple of other facts have become clear. A count of the preferred beverages among American adult consumers 29 and younger shows the following preferences in 2005:

- 48 percent drink beer
- 32 percent drink liquor
- 17 percent drink wine

Perhaps due to the emergence of this younger group, a 2008 study shows beer has regained the position as America's favorite adult beverage:[2]

- 42 percent drink beer
- 23 percent drink spirits
- 31 percent drink wine

Across America, grocers account for 35 percent of all beer sales, but convenience stores, where younger consumers tend to shop, account for 45 percent.[3] If the grocery store is converting more to a convenience store, maybe a continued emphasis on beer is wise. However, wine consumers are more *attractive* from several perspectives. Wine now ranks among the top 10 food categories in America, based on grocery store dollar sales volume. Forty-five percent of all wine is sold in grocery stores. What we find is that the consumer who buys wine is also more likely to buy products like prime or choice beef and imported cheeses, instead of lower quality and lower priced meat and cheese products. As a result, the average $13.44 spent on wine in a grocery store (as opposed to $11.94 on beer) is only part of the story in explaining why wine customers may be *grape* customers![4]

What should the grocer emphasize in marketing adult beverages? Perhaps the research based on counting can address this decision.

# 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**
The elementary transformation of raw data in a way that describes the basic characteristics such as central tendency, distribution, and variability.

**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.

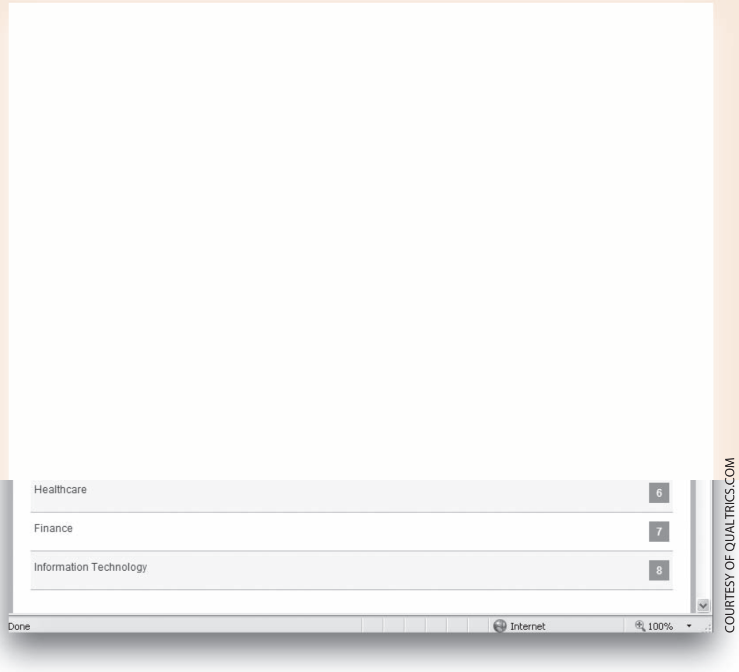
Chapter 13 indicated that the level of scale measurement helps the researcher choose the most appropriate form of statistical analysis. Exhibit 20.1 shows how the level of scale measurement influences the choice of descriptive statistics. Remember that all statistics appropriate for lower-order scales (nominal and ordinal) are suitable for higher-order scales (interval and ratio), but the reverse is not true.

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).

# SURVEY THIS!

© GEORGE DOYLE & CIARAN GRIFFIN

One item in the questionnaire asks respondents to report the career they view as most attractive. A simple way to get an understanding of this population's career aspiration is to simply count the number who rank each profession as their preferred career. Try to draw some conclusions about which job is most attractive: (1) Calculate the number of respondents who rank each profession as the most attractive (assign it a 1). Report this tabulation. (2) Do you think female and male respondents respond similarly to this item? Try to create the appropriate cross-tabulation table to show which jobs are preferred by men and women respectively.




COURTESY OF QUALTRICS.COM

The mode is convenience store since more respondents chose this than any other category. A similar distribution may have been obtained if the chart plotted the number of respondents ranking each store as their favorite type of place to purchase beer.

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. The mean and standard deviation are displayed beside the chart as 11.7 and 4.5, respectively. Additionally, a frequency distribution is shown with a histogram. A **histogram** is a graphical way of showing a frequency distribution in which the height of a bar corresponds to the frequency of a category. Histograms are useful for any

**histogram**
A graphical way of showing a frequency distribution in which the height of a bar corresponds to the observed frequency of the category.

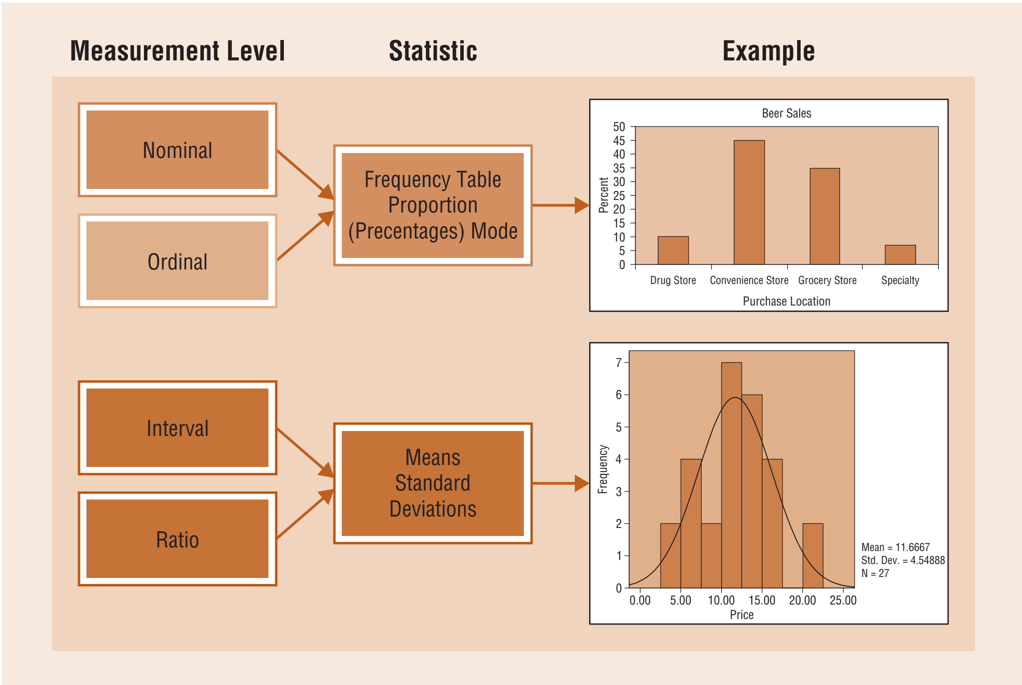



EXHIBIT 20.1
**Levels of Scale Measurement and Suggested Descriptive Statistics**

type of data, but with continuous variables (interval or ratio) the histogram is useful for providing a quick assessment of the distribution of the data. A normal distribution line is superimposed over the histogram, providing an easy comparison to see if the data are skewed or multimodal.

# Tabulation

**tabulation**
The orderly arrangement of data in a table or other summary format showing the number of responses to each response category; tallying.

**frequency table**
A table showing the different ways respondents answered a question.

**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.

Similarly, a recent tabulation of Americans' responses to the simple question of "Who is your favorite TV personality?" revealed the response varied by age. For respondents aged 18–24, Conan O'Brien was listed first. For respondents aged 30–39, Bill O'Reilly was the preferred TV personality, and among consumers 65 and older, Oprah Winfrey was the modal response.[5] The idea that age influences choice of favorite celebrity brings us to cross-tabulation.

# Cross-Tabulation

**cross-tabulation**
The appropriate technique for addressing research questions involving relationships among multiple less-than interval variables; results in a combined frequency table displaying one variable in rows and another in columns.

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. A statistical test for this comparison is discussed in Chapter 21. Here, we focus on constructing and interpreting cross-tabs.

Exhibit 20.2 summarizes several cross-tabulations from responses to a questionnaire on bonuses paid to American International Groups (AIG) executives and federal government bailouts in general.[6] Panel A presents results regarding how closely the respondents have followed the news stories regarding AIG executives receiving bonuses from the 2009 federal government bailout money. The cross-tab suggests this may vary with basic demographic variables. From the results, we can see that more men (60 percent) than women (51 percent) reported they "very closely" followed these news reports. Further, it appears that how closely one followed these news stories increases with age (from 41 percent of those 18–29 to 68 percent of those over 65). Panel B provides another example of a cross-tabulation table. The question asks if the respondents feel that most of the bailout money is going to those that created the crisis. In this case, we see very little difference between men (68 percent agree) and women (69 percent agree). However, before reaching any conclusions based on this survey, one must carefully scrutinize this finding for possible extraneous variables.

EXHIBIT 20.2 **Cross-Tabulation Tables from a Survey Regarding AIG and Government Bailouts**

**(A) Cross-Tabulation of Question "Have you followed the news stories about AIG bonuses?"**

| | | Total | Gender | | Age | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Adults | Men | Women | 18–29 | 30–39 | 40–49 | 50–64 | 65+ |
| Closely Followed News Stories about AIG Bonuses? | Very closely | 55% | 60% | 51% | 41% | 49% | 52% | 67% | 68% |
| | Somewhat closely | 33% | 30% | 35% | 37% | 38% | 39% | 26% | 22% |
| | Not very closely | 8% | 7% | 9% | 20% | 6% | 4% | 4% | 5% |
| | Not at all | 2% | 1% | 4% | 2% | 4% | 2% | 2% | 1% |
| | Not sure | 2% | 2% | 2% | 0% | 2% | 3% | 1% | 3% |

**(B) Cross-Tabulation of Question "Is the bailout money going to those that created the crisis?"**

| | | Total | Gender | | Age | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Adults | Men | Women | 18–29 | 30–39 | 40–49 | 50–64 | 65+ |
| Most Bailout Money Going to People Who Created Crisis? | Yes | 68% | 67% | 69% | 65% | 76% | 68% | 70% | 61% |
| | No | 18% | 23% | 14% | 27% | 14% | 17% | 16% | 16% |
| | Not sure | 14% | 10% | 17% | 8% | 10% | 15% | 14% | 23% |

Source: Rasmussen Reports, National Survey of 1,000 Adults (March 17–18, 2009), http://www.rasmussenreports.com/premium_content/econ_crosstabs/march_2009/crosstabs_aig_march_17_18_2009, accessed March 22, 2009.

## Contingency Tables

Exhibit 20.3 on the next page shows example cross-tabulation results using 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. For all practical purposes, a contingency table is the same as a cross-tabulation.

**contingency table**
A data matrix that displays the frequency of some combination of possible responses to multiple variables; cross-tabulation results.

Two variables are depicted in the contingency table shown in panel A:

- Row Variable: Biological Sex _____M _____F
- Column Variable: "Do you shop at Target? YES or NO"

Several conclusions can be drawn initially by examining the row and column totals:

1. 225 men and 225 women responded, as can be seen in the row totals column.
2. Out of 450 total consumers responding, 330 consumers indicated that "yes" they do shop at Target and 120 indicated "no," they do not shop at Target. This can be observed in the column totals at the bottom of the table. These row and column totals often are called **marginals** because they appear in the table's margins.

**marginals**
Row and column totals in a contingency table, which are shown in its margins.

Researchers usually are more interested in the inner cells of a contingency table. The inner cells display conditional frequencies (combinations). Using these values, we can draw some more specific conclusions:

EXHIBIT 20.3
**Possible Cross-Tabulations of One Question**

**(A) Cross-Tabulation of Question "Do you shop at Target?" by Sex of Respondent**

| | Yes | No | Total |
|---|---|---|---|
| Men | 150 | 75 | 225 |
| Women | 180 | 45 | 225 |
| Total | 330 | 120 | 450 |

**(B) Percentage Cross-Tabulation of Question "Do you shop at Target?" by Sex of Respondent, Row Percentage**

| | Yes | No | Total (Base) |
|---|---|---|---|
| Men | 66.7% | 33.3% | 100% (225) |
| Women | 80.0% | 20.0% | 100% (225) |

**(C) Percentage Cross-Tabulation of Question "Do you shop at Target?" by Sex of Respondent, Column Percentage**

| | Yes | No |
|---|---|---|
| Men | 45.5% | 62.5% |
| Women | 54.5% | 37.5% |
| Total | 100% | 100% |
| (Base) | (330) | (120) |

3. Out of 330 consumers who shop at Target, 150 are male and 180 are female.
4. Alternatively, out of the 120 respondents not shopping at Target, 75 are male and 45 are female.

This finding helps us know whether the two variables are related. If men and women equally patronized Target, we would expect that hypothetically 165 of the 330 shoppers would be female and 165 would be female. Because we have equal numbers of men and women, the 330 would be equally male and female. The hypothetical expectations (165m/165f) are not observed. What is the implication? Target shoppers are more likely to be female than male. Notice that the same meaning could be drawn by analyzing non-Target shoppers. The Research Snapshot on the next page provides an example of the information provided by cross-tabs.

A two-way contingency table like the one shown in part A is referred to as a *2 × 2 table* because it has two rows and two columns. Each variable has two levels. A two-way contingency table displaying two variables, one (the row variable) with three levels and the other with four levels, would be referred to as a *3 × 4 table*. Any cross-tabulation table may be classified according to the number of rows by the number of columns (*R* by *C*).

## Percentage Cross-Tabulations

**statistical base**
The number of respondents or observations (in a row or column) used as a basis for computing percentages.

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. For example, look at the data in parts A, B, and C of Exhibit 20.3. Compare part B with part C. In part B, we are considering gender as the base—what percentage of men and of women

© GEORGE DOYLE & CIARAN GRIFFIN

## RESEARCH SNAPSHOT

### Contingent Personalities

Who is the world's favorite celebrity? This is an important question because the answer helps to determine how much a celebrity endorsement is worth. Sports stars like Tiger Woods are effective in shaping consumers' product preferences worldwide. NBA player Tony Parker is wildly popular in France, where he can be seen endorsing International Watch Company (IWC Schaffhausen) wristwatches. In China, actress Zhang Ziyi helps pitch Maybelline, Garnier, and Asience (Japanese shampoo brand). In other parts of the world, Aishwarya Rai could do the same thing. Perhaps some celebrities are effective nearly everywhere, but others may only be effective in a given country. Their effectiveness is contingent upon region.

Television personalities also influence the public's opinion by giving their own. But all opinions may not be equal. Polling agencies like the Harris interactive poll (http://www.harrisinteractive.com) monitor the popularity of celebrities. Who is America's favorite television personality? Oprah Winfrey has achieved the top rating by Americans for several years. But is Oprah's likeability contingent upon other factors? Cross-tabulations can help answer this question. Consider the following 2-by-2 contingency table showing results of 1,000 respondents asked to choose whether they prefer Oprah Winfrey or David Letterman:

| | Oprah Winfrey | David Letterman | Totals |
|---|---|---|---|
| Men | 150 | 350 | 500 |
| Women | 380 | 120 | 500 |
| | 530 | 470 | 1,000 |

Or, consider the 3-by-3 contingency table:

| | Oprah Winfrey | Bill O'Reilly | Jon Stewart | Totals |
|---|---|---|---|---|
| Conservatives | 60 | 260 | 20 | 340 |
| Liberals | 100 | 20 | 200 | 320 |
| Moderates | 210 | 70 | 60 | 340 |
| | 370 | 350 | 280 | 1,000 |

In either case, opinions about the preferred celebrity seem to be contingent, or to depend on some characteristic. Results like these would suggest that although Oprah is preferred overall, men prefer David Letterman over Oprah. Also, one's favorite celebrity depends on political orientation. Thus, managers should consider the contingencies when trying to identify preferred celebrities.

© D VAN/UPI/LANDOV

Sources: Erdogan, B. Zater, Michael J. Baker, and Stephen Tagg, "Selecting Celebrity Endorsers: The Practitioner's Perspective," *Journal of Advertising Research* 41 (May/June 2001), 39–48; Goetzl, David and Wayne Friedman, "What We're Talking about," *Advertising Age* 73 (December 2, 2002), 51–57; "Harris Poll: Oprah Again Tops America's List of Favorite TV Personalities," *Wall Street Journal Online*, (February 3, 2006), http://online.wsj.com/article/SB113889692780763347.html; "IWC Schaffhausen Appoints Tony Parker as New Friend of the Brand," *PR Newswire* (June 28 2007); Flannery, Russell, "Forbes China Celebrity List," Forbes.com (March 18, 2009).

shop at Target? In part C, we are considering Target shoppers as the base—what percentage of Target shoppers are men? Selecting either the row percentages or the column percentages will emphasize a particular comparison or distribution. The nature of the problem the researcher wishes to answer will determine which marginal total will serve as a base for computing percentages.

Fortunately, a conventional rule determines the direction of percentages. The rule depends on which variable is identified as an independent variable and which is a dependent variable. Simply put, *independent variables should form the rows* in a contingency table. The marginal total of the independent variable should be used as the base for computing the percentages. Although survey research does not establish cause-and-effect evidence, one might argue that it would be logical to assume that a variable such as biological sex might predict beverage preference. This makes more sense than thinking that beverage preference would determine biological sex!

**TO THE POINT**

*The more we study, the more we discover our ignorance.*

—Percy Bysshe Shelley

## 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**
An analysis of the basic cross-tabulation for each level of a variable not previously considered, such as subgroups of the sample.

**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. If the researcher has cross-tabulated shopping preference by sex (see Exhibit 20.3) and wishes to investigate another variable (say, marital status), a more elaborate analysis may be conducted. 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.

EXHIBIT 20.4
**Cross-Tabulation of Marital Status, Sex, and Responses to the Question "Do You Shop at Target?"**

| | 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.

**moderator variable**
A third variable that changes the nature of a relationship between the original independent and dependent variables.

In other situations the addition of a third variable to the analysis may lead us to reject the original conclusion about the relationship. When this occurs, the elaboration analysis suggests the relationship between the original variables is spurious (see Chapter 3).

The chapter vignette described data suggesting a relationship between the type of store in which a consumer shops and beverage preference. Convenience store shoppers seem to choose beer over wine, while grocery store shoppers choose wine over beer. Does store type drive drinking preference? Perhaps a third variable, age, determines both the type of store consumers choose to buy in and their preference for adult beverages. Younger consumers both disproportionately shop in convenience stores and drink beer.

## How Many Cross-Tabulations?

Surveys may ask dozens of questions and hundreds of categorical variables can be stored in a data warehouse. Using computer programs, business researchers could "fish" for relationships by cross-tabulating every categorical variable with every other categorical variable. Thus, every possible response becomes a possible explanatory variable. A researcher addressing an exploratory research question may find some benefit in such a fishing expedition. Software exists that can automatically search through volumes of cross-tabulations. These may even provide some insight into the business questions under investigation. Alternatively, the program may flag the cross-tabulations suggesting the strongest relationship. CHAID (chi-square automatic interaction detection) software exemplifies software that makes searches through large numbers of variables possible.[7] Data-mining can be conducted in a similar fashion and may suggest relationships that are worth considering further.

However, outside of exploratory research, researchers should conduct cross-tabulations that address specific research questions or hypotheses. When hypotheses involve relationships among two categorical variables, cross-tabulations are the right tool for the job.

## Quadrant Analysis

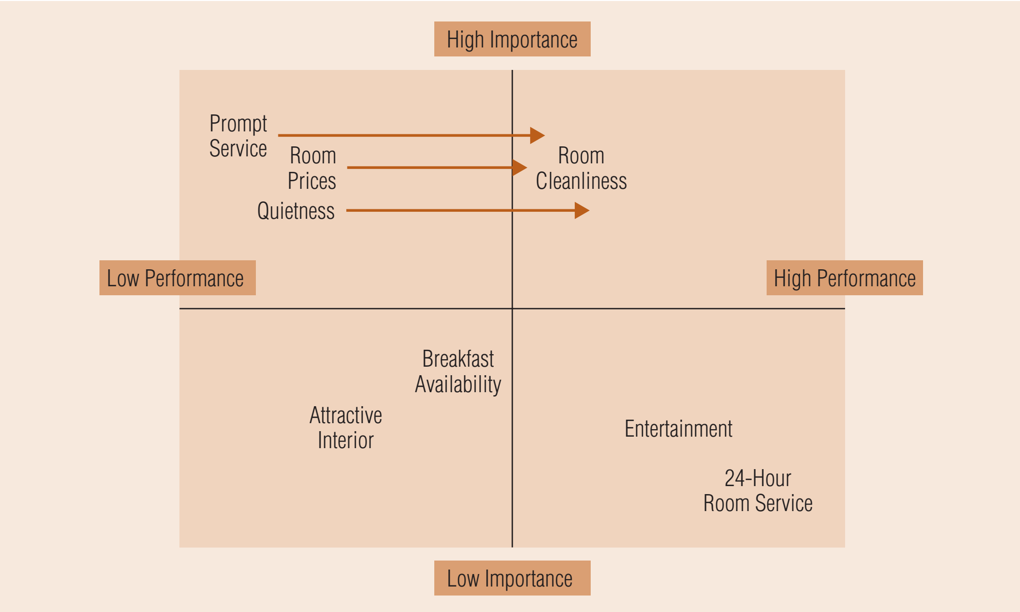
**Quadrant analysis** is a variation of cross-tabulation in which responses to two rating scale questions are plotted in four quadrants of a two-dimensional table. A common quadrant analysis in business research portrays or plots relationships between average responses about a product attribute's importance 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.

**quadrant analysis**
An extension of cross-tabulation in which responses to two rating-scale questions are plotted in four quadrants of a two-dimensional table.

**importance-performance analysis**
Another name for quadrant analysis.

Exhibit 20.5 illustrates a quadrant analysis for an international, mid-priced hotel chain.[8] 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.

EXHIBIT 20.5
**An Importance-Performance or Quadrant Analysis of Hotels**




# 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.

**data transformation**
Process of changing the data from their original form to a format suitable for performing a data analysis addressing research objectives.

**TO THE POINT**

*All that we do is done with an eye to something else.*

—Aristotle

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:

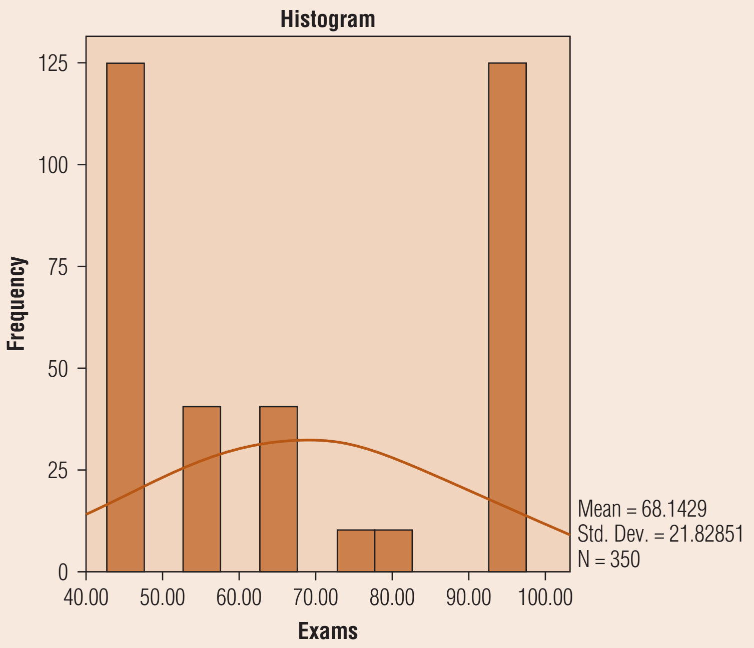
| | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
|---|---|---|---|---|---|
| I am satisfied with my college experience at this university | ☐ | ☐ | ☐ | ☐ | ☐ |

The following frequency table describes results for this survey item:

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

The distribution of responses suggests the responses are bimodal. That is, two "peaks" exist in the distribution, one at either end of the scale. Exhibit 20.6 shows an example of a bimodal distribution. Since the vast majority of respondents [80 percent = (110 + 210)/400] indicated either strongly disagree or strongly agree, the variable closely resembles a categorical variable. In general, customers either strongly disagreed or strongly agreed with the statement. So, the research may wish to collapse the responses into two categories. While multiple ways exist to accomplish this, the researcher may assign the value of one to all respondents who either strongly disagreed or disagreed and the value two to all respondents who either agreed or strongly agreed. Respondents marking neutral would be deleted from analysis. In this case, we would end up with 140 (110 + 30) respondents that disagree with this statement and 245 (210 + 35) that agreed.

EXHIBIT 20.6
**Bimodal Distributions Are Consistent with Transformations into Categorical Values**




Adapted from 1987 Nielsen Television Report.

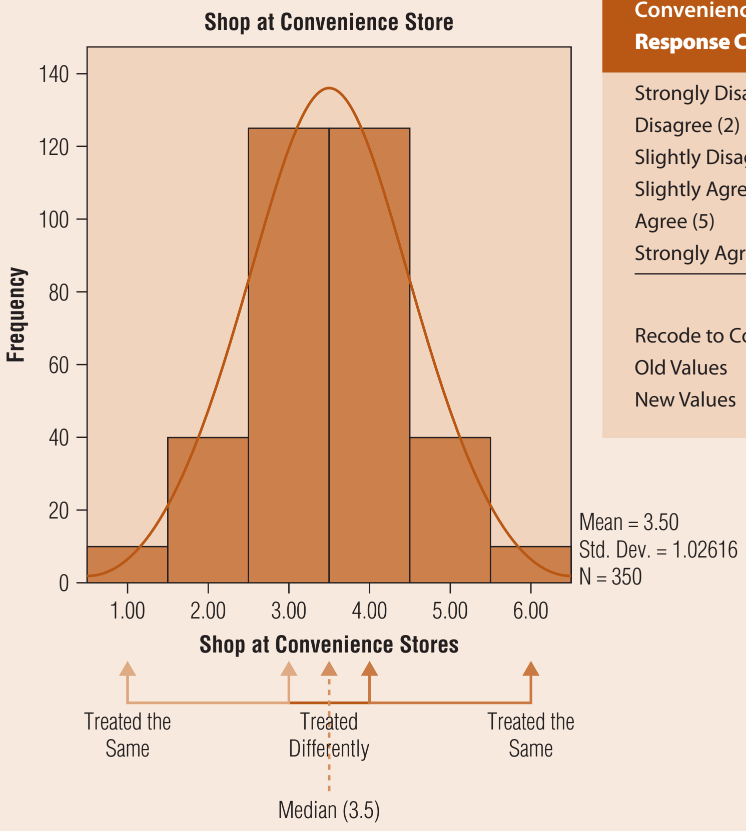
**median split**
Dividing a data set into two categories by placing respondents below the median in one category and respondents above the median in another.

## 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.

Exhibit 20.7 illustrates this problem. Clearly, most respondents either slightly agree or slightly disagree with this statement. The central tendency could be represented by the median of 3.5, a mean of 3.5, and modes of 3 and 4 (3 and 4 each have the same number of responses). The "outliers," if any, appear to be those not indicating something other than slight agreement/disagreement. A case can be made that the respondents indicating slight disagreement are more similar to those indicating slight agreement than they are to those respondents indicating strong disagreement. Yet we can see the recode places values 1 and 3 in the same new category, but places values 3 and 4 in a different category (see the recoding scheme in Exhibit 20.7). The data distribution does not support a median split into two categories and so a transformation collapsing these values into agreement and disagreement is inappropriate.

**EXHIBIT 20.7 The Problem with Median Splits with Unimodal Data**




**Frequency Distribution: $X_1$ = I Do Most of My Shopping at Convenience Stores.**

| Response Category (Code) | Counts | Cumulative Percentage |
|---|---|---|
| Strongly Disagree (1) | 10 | 2.86% |
| Disagree (2) | 40 | 14.29% |
| Slightly Disagree (3) | 125 | 50.00% |
| Slightly Agree (4) | 125 | 85.71% |
| Agree (5) | 40 | 97.14% |
| Strongly Agree (6) | 10 | 100.00% |

Median = 3.5

Recode to Complete Data Transformation:

| | | | | | | |
|---|---|---|---|---|---|---|
| Old Values | 1 | 2 | 3 | 4 | 5 | 6 |
| New Values | 1 | 1 | 1 | 2 | 2 | 2 |

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, we have more appropriate analytical techniques that are discussed in the chapters which follow.

## Index Numbers

**index numbers**

Scores or observations recalibrated to indicate how they relate to a base number.

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:

$$\text{1 person hh:} \quad \frac{41:01}{52:36} = 0.7832 \times 100 = 78.32$$

$$\text{2 person hh:} \quad \frac{47:58}{52:36} = 0.9087 \times 100 = 90.87$$

$$\text{3+ person hh:} \quad \frac{60:49}{52:36} = 1.1553 \times 100 = 115.53$$

$$\text{Total U.S. average:} \quad \frac{52:36}{52:36} = 1.0000 \times 100 = 100.00$$

EXHIBIT 20.8
**Hours of Television Usage per Week**

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

Adapted from 1987 Nielsen Television Report.

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. Managers may often chart consumption in some category over time. Relating back to the chapter vignette, grocers may wish to chart the U.S. wine consumption index. Using 1968 as a base year, the current U.S. wine consumption index is just over 2.0, meaning that the typical American consumer drinks about twice as much wine today as in 1968, which is just over 8.7 liters of wine per year.[9] The Research Snapshot on the next page shows another application of data transformation and index creation.

## 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.

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

# RESEARCH SNAPSHOT

## Twitter and the *Re*Tweetability Index

Twitter is one of the fastest growing social networks. A privately funded organization in San Francisco, Twitter's first prototype was developed in March of 2006 and launched publicly five months later. Since then, Twitter has evolved into a real-time messaging service compatible with several different networks and multiple devices:

*Simplicity has played an important role in Twitter's success. People are eager to connect with other people and Twitter makes that simple. Twitter asks one question, "What are you doing?" Answers must be under 140 characters in length and can be sent via mobile texting, instant message, or the web.*

*Twitter's core technology is a device agnostic message routing system with rudimentary social networking features. By accepting messages from sms, web, mobile web, instant message, or from third party API projects, Twitter makes it easy for folks to stay connected.*

If you are not familiar with Twitter, a basic understanding of the terminology is necessary. After signing up for a Twitter account, you can *tweet* your 140 character message. *Followers* are people who have signed up to receive someone's Twitter messages. A *retweet* (or RT) occurs when a follower takes a tweet and then tweets that message to everyone in their own Twitter network. Encouraging other Twitter users to retweet your messages is the key in spreading your message across the Twittersphere. Wow!

Dan Zarrella, a self-proclaimed viral marketing scientist, has developed an index to assess the most influential Twitter users. While several sites rank users by their number of followers, and others report the number of RTs, Zarrella has combined these figures with the daily number of tweets to calculate the *Re*Tweetability Index:

(Retweets per Day / Tweets per Day) / Followers

The index is intended to provide a score and ranking of Twitter users based on the power of their tweets. The higher the number, the more influential Twitter you are!

Sources: "About Twitter," Twitter, http://twitter.com/about; Dan Zarrella's *Re*Tweetability Index, http://www.retweetability.com; Saric, Marko, "Make Your Blog Go Viral with Twitter ReTweets," How to Make My Blog.com (January 13, 2009), http://www.howtomakemyblog.com/twitter/make-your-blog-go-viral-with-twitter-retweets.

© GEORGE DOYLE & CIARAN GRIFFIN

COURTESY OF TWITTER.COM

EXHIBIT 20.9
**Executive Rankings of Potential Conference Destinations**

| 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 |

EXHIBIT 20.10
**Frequencies of Conference Destination Rankings**

| | Preference Rankings | | | |
|---|---|---|---|---|
| Destination | 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 |

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:

| | |
|---|---|
| Hawaii: | $(3 \times 1) + (5 \times 2) + (1 \times 3) + (1 \times 4) = 20$ |
| Paris: | $(3 \times 1) + (1 \times 2) + (3 \times 3) + (3 \times 4) = 26$ |
| Greece: | $(2 \times 1) + (2 \times 2) + (4 \times 3) + (2 \times 4) = 26$ |
| Hong Kong: | $(2 \times 1) + (2 \times 2) + (2 \times 3) + (4 \times 4) = 28$ |

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!

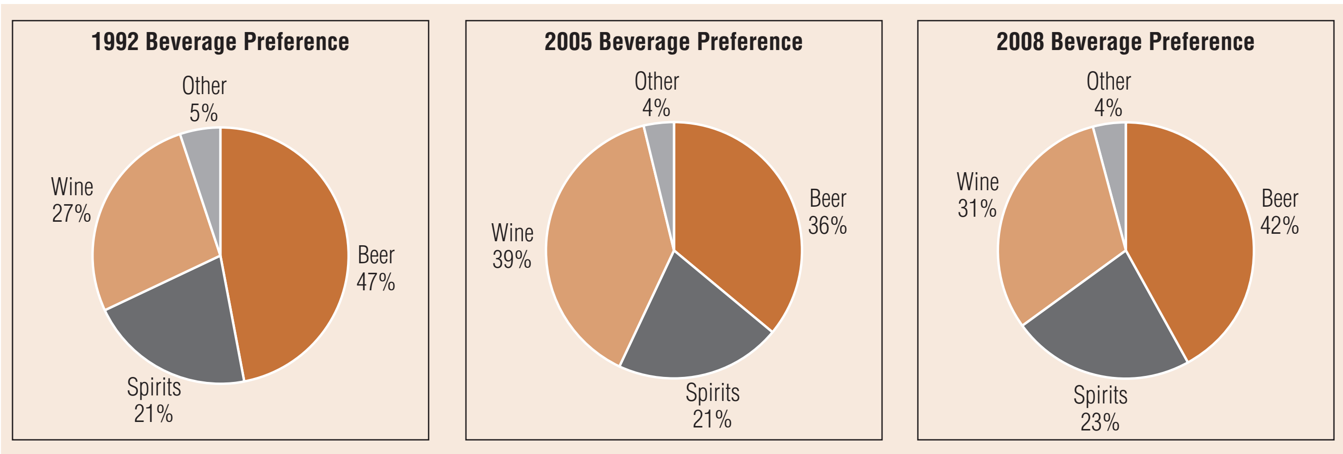
## Tabular and Graphic Methods of Displaying Data

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 communicates

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



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.

Chapter 25 discusses how these and other graphic aids may improve the communication value of a written report or oral presentation.

# 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.[10] 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.

Exhibit 20.12 on the next page shows a printout of descriptive statistics generated by SAS for two variables: EMP (number of employees working in an MSA, or Metropolitan Statistical Area) and SALES (sales volume in dollars in an MSA) for 10 MSAs. The number of data elements (*N*), mean, standard deviation, and other descriptive statistics are displayed. SAS output is generally simple and easy to read.

**EXHIBIT 20.12 SAS Computer Output of Descriptive Statistics**

| State = NY Variable | N | Mean | Standard Deviation | Minimum Value | Maximum Value | Std. Error of Mean | Sum | Variance | C.V. |
|---|---|---|---|---|---|---|---|---|---|
| EMP | 10 | 142.930 | 232.665 | 12.800 | 788.800 | 73.575 | 1429.300 | 54133.0 | 162.782 |
| SALES | 10 | 5807.800 | 11905.127 | 307.000 | 39401.000 | 3764.732 | 58078.000 | 141732049.1 | 204.985 |

Key: EMP = number of employees (000) SALES = Sales (000)

As an example of SPSS output, the histograms shown in Exhibits 20.6 and 20.7 were created by SPSS. By clicking on "charts" in the SPSS tool menu, one can see the variety of charts that can be created. The key place to click to generate statistical results in tabular form is "analyze." Here, one can see the many types of analysis that can be created. In this chapter, the choices found by clicking on "analyze" and then "descriptive statistics" are particularly relevant.

Exhibit 20.13 shows an SPSS cross-tabulation of two variables, class status and smoking behavior. The data come from a sample intercepted on an urban university campus. It addresses the research question, "Does smoking on campus vary across groups?" More nonsmokers than smokers are found. However, the results show that graduate students, and to a lesser extent instructors, smoke more than the norm. The SPSS user can ask for any number of statistics and percentages to be included with this output by clicking on the corresponding options.

**EXHIBIT 20.13**
**Examples of SPSS Output for Cross-Tabulation**

**CLASS * SMOKING Cross-Tabulation**

| | | Smoking | | |
|---|---|---|---|---|
| Count | | Smoker | Non-Smoker | Total |
| Class | high school | 7 | 9 | 16 |
| | undergraduate | 9 | 22 | 31 |
| | graduate | 15 | 10 | 25 |
| | career | 6 | 6 | 12 |
| Total | | 37 | 47 | 84 |

## 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. As mentioned in Chapter 2, 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.

**TO THE POINT**

*The thing to do is to supply light.*

—Woodrow Wilson

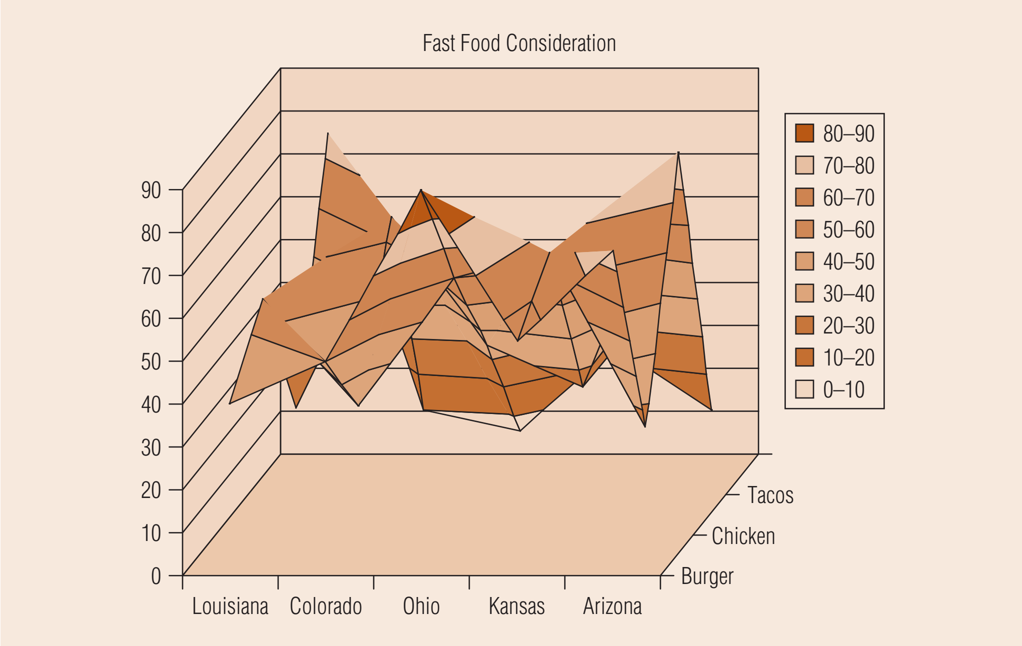



**EXHIBIT 20.14**
**A 3-D Graph Showing Fast-Food Consumption Patterns around the United States**

Many computer programs can draw **box and whisker plots**, which provide graphic representations of central tendencies, percentiles, variabilities, and the shapes of frequency distributions. Exhibit 20.15 shows a computer-drawn box and whisker plot for 100 responses to a question measured on a 10-point scale. The response categories are shown on the vertical axis. The small box inside the plot represents responses for half of all respondents. Thus, half of respondents marked 4, 5, or 6. This gives a measure of variability called the **interquartile range**, but the term *midspread* is less complex and more descriptive. The location of the line within the box indicates the median. The dashed lines that extend from the top and bottom of the box are the whiskers. Each whisker extends either the length of the box (the midspread in our example is 2 scale points) or to the most extreme observation in that direction.

**box and whisker plots**
Graphic representations of central tendencies, percentiles, variabilities, and the shapes of frequency distributions.

**interquartile range**
A measure of variability.

An **outlier** is a value that lies outside the normal range of the data. In Exhibit 20.15 on the next page outliers are indicated by either a 0 or an asterisk. Box and whisker plots are particularly useful for spotting outliers or comparing group categories (e.g., men versus women).

**outlier**
A value that lies outside the normal range of the data.

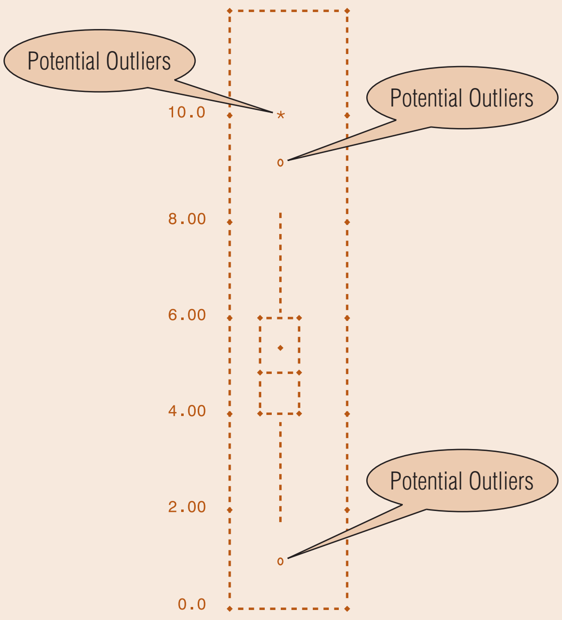
# Interpretation

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. This

**interpretation**
The process of drawing inferences from the analysis results.

**EXHIBIT 20.15**
**Computer Drawn Box and Whisker Plot**




| | |
|---|---|
| Mean | 5.40 |
| Median | 5.00 |
| 75th percentile | 6.00 |
| 25th percentile | 4.00 |
| Standard deviation | 1.62 |

From "Graphic Displays of Data: Box and Whisker Plots," Reports No. 17, Market Facts, Inc.

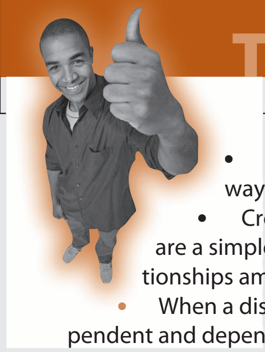
interpretation of the statistical analysis may lead to a realization that certain health plans are better suited for different family situations.

From a management perspective, however, the qualitative meaning of the data and their managerial implications are an important aspect of the interpretation. Consider the crucial role played by interpretation of research results in investigating one new product, a lip stain that could color the lips a desired shade semi-permanently and last for about a month at a time:

> *The lip stain idea, among lipstick wearers, received very high scores on a rating scale ranging from "excellent" to "poor," presumably because it would not wear off. However, it appeared that even among routine wearers of lipstick the idea was being rated highly more for its interesting, even ingenious, nature than for its practical appeal to the consumer's personality. They liked the idea, but for someone else, not themselves. . . . [Careful interpretation of the data] revealed that not being able to remove the stain for that length of time caused most women to consider the idea irrelevant in relation to their own personal needs and desires. Use of the product seems to represent more of a "permanent commitment" than is usually associated with the use of a particular cosmetic. In fact, women attached overtly negative meaning to the product concept, often comparing it with hair dyes instead of a long-lasting lipstick.*[11]

This example shows that interpretation is crucial. However, the process is difficult to explain in a textbook because there is no one best way to interpret data. Many possible interpretations of data may be derived from a number of thought processes. Experience with selected cases will help you develop your own interpretative ability.

Data are sometimes merely reported and not interpreted. Research firms may provide reams of computer output that do not state what the data mean. At the other extreme, some researchers tend to analyze every possible relationship between each and every variable in the study. Such an



© GEORGE DOYLE & CIARAN GRIFFIN

## TIPS OF THE TRADE

- A frequency table can be a very useful way to depict basic tabulations.
- Cross-tabulation and contingency tables are a simple and effective way to examine relationships among less than interval variables.
  - When a distinction can be made between independent and dependent variables (that are nominal or ordinal), the convention is rows are independent variables and columns are dependent variables.
- Importance-performance charts are a good way to illustrate market positioning by showing where brands are strong or weak on important variables. A weakness on an important variable is a call to action.
- A continuous variable that displays a bimodal distribution is appropriate for a median split.
  - Median splits on continuous variables displaying a normal distribution are typically not appropriate and result in a loss of information. If necessary, these should only be performed after deleting one-fourth to one-third of the responses around the median to help prevent logically inconsistent classifications.
- Box and whisker plots can reveal outliers.
  - Outliers can distort statistical analysis. Therefore, they become candidates for deletion.

approach is a sign that the research problem was not adequately defined prior to beginning the research and the researcher really doesn't know what business decision the research is addressing. Researchers who have a clear sense of the purpose of the research do not request statistical analysis of data that have little or nothing to do with the primary purpose of the research.

## Summary

**1. Know what descriptive statistics are and why they are used.** Descriptive analyses provide descriptive statistics including measures of central tendency and variation. Statistics such as the mean, mode, median, range, variance, and standard deviation are all descriptive statistics. These statistics provide a summary describing the basic properties of a variable.

**2. Create and interpret simple tabulation tables.** Statistical tabulation is another way of saying that we count the number of observations in each possible response category. In other words, tabulation is the same as tallying. Tabulation is an appropriate descriptive analysis for less-than interval variables. Frequency tables and histograms are used to display tabulation results.

**3. Understand how cross-tabulations can reveal relationships.** Cross-tabulation is when we combine two or more less-than interval variables to display the relationship. For example, a cross-tabulation of respondent gender with adult beverage preference (i.e., beer, spirits, wine) would give us two rows (male and female) and three columns (beer, spirits, wine), which would show the preferred beverage for each gender. The key to interpreting a cross-tabulation result is to compare actual observed values with hypothetical values that would result from pure chance. When observed results vary from these values, a relationship is indicated.

**4. Perform basic data transformations.** Data transformations are often needed to assist in data analysis and involve changing the mathematical form of data in some systematic way. Basic data transformations include reverse coding, summating scales, creating index numbers, and collapsing a variable based on a median split.

**5. List different computer software products designed for descriptive statistical analysis.** While spreadsheets have improved with respect to their ability to conduct basic statistical analyses, business researchers still rely heavily on specialized statistical software. SAS and SPSS are two of the best known statistical packages. Each is available for even the most basic modern PC and can be used with a drop-down window interface, practically eliminating the need for writing computer code.

**6. Understand a researcher's role in interpreting the data.** The interpretation process explains the meaning of the data. Interpretation is drawing inferences from the analysis results; providing meaning for the figures which are observed. Inferences drawn from interpretations lead to managerial implications.

# Key Terms and Concepts



# Questions for Review and Critical Thinking

1. What are five descriptive statistics used to describe the basic properties of variables?
2. What is a *histogram?* What is the advantage of overlaying a normal distribution over a histogram?
3. A survey asks respondents to respond to the statement "My work is interesting." Interpret the frequency distribution shown here (taken from an SPSS output):
   a. My work is interesting:

| Category Label | Code | Abs. Freq. | Rel. Freq. (Pct.) | Adj. Freq. (Pct.) | Cum. Freq. (Pct.) |
|---|---|---|---|---|---|
| Very true | 1 | 650 | 23.9 | 62.4 | 62.4 |
| Somewhat true | 2 | 303 | 11.2 | 29.1 | 91.5 |
| Not very true | 3 | 61 | 2.2 | 5.9 | 97.3 |
| Not at all true | 4 | 28 | 1.0 | 2.7 | 100.0 |
| | • | 1,673 | 61.6 | Missing | |
| | Total | 2,715 | 100.0 | 100.0 | |
| Valid cases | 1,042 | | Missing cases | 1,673 | |

4. Use the data in the following table to
   a. Prepare a frequency distribution of the respondents' ages
   b. Cross-tabulate the respondents' genders with cola preference
   c. Identify any outliers

| Individual | Gender | Age | Cola Preference | Weekly Unit Purchases |
|---|---|---|---|---|
| James | M | 19 | Coke | 2 |
| Parker | M | 17 | Pepsi | 5 |
| Bill | M | 20 | Pepsi | 7 |
| Laurie | F | 20 | Coke | 2 |
| Jim | M | 18 | Coke | 4 |
| Jil | F | 16 | Coke | 4 |
| Tom | M | 17 | Pepsi | 12 |
| Julia | F | 22 | Pepsi | 6 |
| Amie | F | 20 | Pepsi | 2 |
| Dawn | F | 19 | Pepsi | 3 |

5. Data on the average size of a soda (in ounces) at all 30 major league baseball parks are as follows: 14, 18, 20, 16, 16, 12, 14, 16, 14, 16, 16, 16, 14, 32, 16, 20, 12, 16, 20, 12, 16, 16, 24, 16, 16, 14, 14, 12, 14, 20. Compute descriptive statistics for this variable including a box and whisker plot. Comment on the results.
6. The following computer output shows a cross-tabulation of frequencies and provides frequency number N) and row R) percentages.
   a. Interpret this output including a conclusion about whether or not the row and column variables are related.

b. Critique the way the analysis is presented.
c. Draw a pie chart indicating percentages for having read a book in the past three month for those with and those without high school diplomas.

| Have You Read a Book in Past 3 Months? | Have High School Diploma? Yes | No | Total |
|---|---|---|---|
| Yes | 489<br>73.8 | 174<br>26.2 | 663 |
| No | 473<br>55.6 | 378<br>44.4 | 851 |
| | ...... | ...... | ...... |
| TOTAL | 962 | 552 | 1,514 |

7. List and describe at least three basic data transformations.
8. What conditions suggest that a ratio variable should be transformed (recoded) into a dichotomous (two group) variable?
9. A data processing analyst for a research supplier finds that preliminary computer runs of survey results show that consumers love a client's new product. The employee buys a large block of the client's stock. Is this ethical?

# Research Activities

1. **'NET** Go the Web site for the Chicago Cubs baseball team (**http://chicago.cubs.mlb.com**). Use either the schedule listing or the stats information to find their record in the most recent season. Create a data file with a variable indicating whether each game was won or lost and a variable indicating whether the game was played at home in Wrigley Field or away from home. Using computerized software like SPSS or SAS,
   a. Compute a frequency table and histogram for each variable.
   b. Use cross-tabulations to examine whether a relationship exists between where the game is played (home or away) and winning.
   c. Extra analysis: Repeat the analyses for the Houston Astros baseball team (**http://houston.astros.mlb.com**). What does this suggest for the relationship between playing at home and winning?
2. **'NET** Go to **http://www.spss.com** and click on Industries and Market Research. What services does the company provide?

## Case 20.1 Body on Tap


© GETTY IMAGES/ PHOTODISC GREEN

A few years ago Vidal Sassoon, Inc., took legal action against Bristol-Myers over a series of TV commercials and print ads for a shampoo that had been named Body on Tap because of its beer content.[12] The prototype commercial featured a well-known high fashion model saying, "In shampoo tests with over 900 women like me, Body on Tap got higher ratings than Prell for body. Higher than Flex for conditioning. Higher than Sassoon for strong, healthy-looking hair."

The evidence showed that several groups of approximately 200 women each tested just one shampoo. They rated it on a six-step qualitative scale, from "outstanding" to "poor," for 27 separate attributes, such as body and conditioning. It became clear that 900 women did not, after trying both shampoos, make product-to-product comparisons between Body on Tap and Sassoon or between Body on Tap and any of the other brands mentioned. In fact, no woman in the tests tried more than one shampoo.

The claim that the women preferred Body on Tap to Sassoon for "strong, healthy-looking hair" was based on combining the data for the "outstanding" and "excellent" ratings and discarding the lower four ratings on the scale. The figures then were 36 percent for Body on Tap and 24 percent (of a separate group of women) for Sassoon. When the "very good" and "good" ratings were combined with the "outstanding" and "excellent" ratings, however, there was only a difference of 1 percent between the two products in the category of "strong, healthy-looking hair."

The research was conducted for Bristol-Myers by Marketing Information Systems, Inc. (MISI), using a technique known as blind monadic testing. The president of MISI testified that this method typically is employed when what is wanted is an absolute response to a product "without reference to another specific product." Although he testified that blind monadic testing was used in connection with comparative advertising, that was not the purpose for which Bristol-Myers retained MISI. Rather, Bristol-Myers wished to determine consumer reaction to the introduction of Body on Tap. Sassoon's in-house research expert stated flatly that blind monadic testing cannot support comparative advertising claims.

### Question

Comment on the professionalism of the procedures used to make the advertising claim. Why do you believe the researchers performed the data transformations described?

## Case 20.2 Downy-Q Quilt


© GETTY IMAGES/ PHOTODISC GREEN

The research for Downy-Q is an example of a commercial test that was conducted when an advertising campaign for an established brand had run its course.[13] The revised campaign, "Fighting the Cold," emphasized that Downy-Q was an "extra-warm quilt"; previous research had demonstrated that extra warmth was an important and deliverable product quality. The commercial test was requested to measure the campaign's ability to generate purchase interest.

The marketing department had recommended this revised advertising campaign and was now anxious to know how effectively this commercial would perform. The test concluded that "Fighting the Cold" was a persuasive commercial. It also demonstrated that the new campaign would have greater appeal to specific market segments.

### Method

Brand choices for the same individuals were obtained before and after viewing the commercial. The commercial was tested in 30-second, color-moving, storyboard form in a theater test. Invited viewers were shown programming with commercial inserts. Qualified respondents were women who had bought quilts in outlets that carried Downy-Q. The results are shown in Case Exhibits 20.2–1 through 20.2–4.

### Question

Interpret the data in these tables. What recommendations and conclusions would you offer to Downy-Q management?

**CASE EXHIBIT 20.2–1 Shifts in Brand Choice before and after Showing of Downy-Q Quilt Commercial**

Question: We are going to give away a sample of fabric softener. You can select the brand you most prefer. Which brand would you chose?

| | Brand Choice after Commercial (%) | |
|---|---|---|
| Brand Choice before Commercial | Downy-Q ($n$ = 23) | Other Brand ($n$ = 237) |
| Downy-Q | 78 | 19 |
| Other brand | 22 | 81 |

**CASE EXHIBIT 20.2–2 Pre/Post Increment in Choice of Downy-Q**

Improvement in score based on exposure to commercial.

| | "Fighting the Cold" | | Norm: All Quilt Commercials | |
|---|---|---|---|---|
| Demographic Group | Base | Score | Average | Range |
| Total audience | (260) | +15 | +10 | 6–19 |
| By marital status | | | | |
| Married | (130) | +17 | | |
| Not married | (130) | +12 | | |
| By age | | | | |
| Under 35 | (130) | +14 | | |
| 35 and over | (130) | +15 | | |
| By employment status | | | | |
| Not employed | (90) | +13 | | |
| Employed | (170) | +18 | | |

**CASE EXHIBIT 20.2–3 Adjective Checklist for Downy-Q Quilt Commercial**

Question: Which of these words do you feel come closest to describing the commercial you've just seen? (Check all the apply.)

| Adjective | "Fighting the Cold" (%) | Norm: All Quilt Commercials (%) |
|---|---|---|
| Positive | | |
| Appealing | 18 | 24 |
| Clever | 11 | 40 |
| Convincing | 20 | 14 |
| Effective | 19 | 23 |
| Entertaining | 5 | 24 |
| Fast moving | 12 | 21 |
| Genuine | 7 | 4 |
| Imaginative | 7 | 21 |
| Informative | 24 | 18 |
| Interesting | 13 | 17 |
| Original | 7 | 20 |
| Realistic | 8 | 3 |
| Unusual | 3 | 8 |
| Negative | | |
| Amateurish | 9 | 11 |
| Bad Taste | 4 | 4 |
| Dull | 33 | 20 |
| Repetitious | 17 | 16 |
| Silly | 8 | 19 |
| Slow | 8 | 7 |
| Unbelievable | 3 | 5 |
| Unclear | 3 | 2 |
| Unimportant | 14 | 14 |
| Uninteresting | 32 | 19 |

**CASE EXHIBIT 20.2–4 Product Attribute Checklist for Downy-Q**

Question: Which of the following statements do you feel apply to Downy-Q? (Mark as many or as few as you feel apply.)

| Attributes | "Fighting the Cold" (%) |
|---|---|
| Extra warm | 56 |
| Lightweight | 48 |
| Pretty designs | 45 |
| Durable fabrics | 28 |
| Nice fabrics | 27 |
| Good construction | 27 |