2

DESCRIBING AND EXAMINING DATA

Information is the currency of research. However, unlike real money, sometimes we are overwhelmed by too much information. In most cases, data must be summarized to be useful.

The most common method of summarizing data is with descriptive statistics and graphs.

Even if you're planning to analyze a set of data using a statistical technique such as a *t*-test, analysis of variance, or logistic regression, you should always begin by examining the data. This preliminary step helps you determine which statistical analysis techniques should be used to answer your research questions.

In fact, this process of examining your data often reveals information that will surprise or inform you. You may discover unusually high or low values in your data. Perhaps these "outliers" are caused by incorrectly coded data, or they may reveal information about your data (or subjects) that you have not anticipated. You might observe that some values are not normally distributed. You might notice that a histogram of observations shows two distinct peaks, causing you to realize that your data show a difference between genders. Insights such as these often result from the proper use of descriptive techniques.

The following sections of this chapter discuss the most commonly used tactics for understanding, describing, or preparing your data for further analysis. The two major topics discussed in this chapter are as follows:

- Describing quantitative data using statistics
- Describing categorical data using statistics

Each section includes, where appropriate, methods for reporting results in a standard manner for a report or journal article using APA guidelines (2009). Chapter 3: Creating and Using Graphs continues with the concept of describing data but with an emphasis on graphs.

SIDEBAR

Recall that the purpose of *Quick Guide to IBM SPSS* is to "cut to the chase" and discuss only material that is typically important in performing a well-planned and analyzed experiment. Therefore, some statistics, tests, and graphs that are presented in standard computer output are not discussed.

EXAMPLE DATA FILES

Before continuing, there is one bit of housekeeping we need to cover; we assume that you've installed the sample data sets on your computer, as described in Chapter 1 in the section "Downloading Sample SPSS Data Files." What? You haven't installed the sample data? Go back. "Do not pass Go. Do not collect $200."

You've installed the sample data files now? Great! Welcome back. Now that you have the sample data stored on your computer, you are ready to proceed.

DESCRIBING QUANTITATIVE DATA

Quantitative data are numeric data on which computations such as addition and subtraction make sense. In SPSS terminology, quantitative data are called "scale" data. Statisticians usually use the word *quantitative* instead of *scale* to refer to this type of numeric data, and in this description, we use them interchangeably. Furthermore, there are additional delineations of quantitative data that SPSS does not mention or use. For example, quantitative data are often characterized as the following:

- *Continuous values*: Data values fall along a continuum. Examples are weight or length.
- *Discrete values*: Data are defined as a countable number of possible outcomes. Examples are number of children in your family, number of years of schooling, and so on.
- *Interval values*: Data where the difference between two observed values is meaningful. For example, in the Fahrenheit temperature scale, the

difference between 0 and 10 degrees is the same "distance" as the difference between 90 and 100 degrees.

- *Ratio Values*: Interval data with the additional restriction that it has a natural zero (0) value. For example, the weight of an object has a natural value of 0, meaning no weight (absence of any weight). (Whereas a 0 temperature, in Celsius or Fahrenheit, does not mean the absence of temperature.)

SIDEBAR

The determination of whether arithmetic computation for the data makes sense is a key component in the decision regarding whether your data are quantitative or not.

Examples of the types of data that would fall into the quantitative (scale) category include the following:

- Rainfall (continuous, ratio)
- Child's temperature recoded upon arrival at an emergency room (continuous, interval)
- Number of peaches harvested from each tree (discrete, ratio)
- Years of smoking (discrete, ratio)

It makes sense to talk about the average value of each of these variables (even if there is no such thing as a part of a peach growing on a tree).

Observe the Distribution of Your Data

Since many common statistical procedures assume normally distributed data, you may want to examine your data set to determine whether it fits this criterion. There are a number of ways to check the normality of a set of observed values. These include both numerical and graphical techniques. This chapter discusses several numerical methods used to understand your data's distribution, and graphical techniques are discussed in Chapter 3: Creating and Using Graphs.

Testing for Normality

SPSS provides the Kolmogorov-Smirnov and Shapiro-Wilk tests in the Explore procedure to test the hypothesis that the distribution of a set of data is normal. The hypotheses used in testing for normality are as follows:

H_0: The data follow a normal distribution.

H_a: The data do not follow a normal distribution.

If a test does not reject normality, this suggests that a parametric procedure that assumes normality (e.g., a *t*-test) can be safely used. However, it is always a good idea to also examine data graphically in addition to the formal tests for normality.

SIDEBAR

The term *normal distribution* used here and in the remainder of the book indicates that histograms of data sampled from this distribution will approximate a bell-shaped curve. The normal distribution is also referred to as a Gaussian distribution (after the mathematician Karl Friedrich Gauss). We will refer to data from a normal distribution as "normal data." Figure 2.6, shown in a later example in this chapter, shows a bell-shaped curve fitted to data.

Tips and Caveats for Quantitative Data

How to Use the Information About Normality

Given the fact that there is a normality assumption associated with many statistical procedures (e.g., the *t*-test), you'd think it was a "make-or-break" criterion for your analysis. This is not necessarily the case. In fact, true normality is usually a myth. What is important is to ascertain whether your data show a serious departure from normality. Data showing a moderate departure from normality can usually be used in parametric procedures without loss of integrity. However, if your data are not close to normal and your sample sizes are small, you should consider using a nonparametric statistical test that does not assume normality. Nonparametric tests are discussed in Chapter 8: Nonparametric Analysis Procedures. For further discussion of the normality assumption in the context of the *t*-test, see Chapter 4: Comparing One or Two Means Using the *t*-Test. Other reasons you might choose a nonparametric procedure are given by Fay and Proschan (2010) and Sawilowsky (2005).

If Data Are Not Normally Distributed, Don't Report the Mean

Describe distinctly nonnormal data with the median and range or interquartile range (Lang & Secic, 2006). Another way to report this type of data is by using a Tukey five-number summary consisting of the minimum, 25th percentile, 50th percentile, 75th percentile, and maximum. This five-number summary is the basis for the boxplot (illustrated in Chapter 3: Creating and Using Graphs).

When in Doubt, Report the *SD* Rather Than the *SEM*

When reporting descriptive statistics, there is sometimes a dilemma regarding whether to report the standard deviation (*SD*) or standard error of the mean (*SEM*), which is the *SD* divided by the square root of the sample size. You should report the *SD* if you are describing the variability of the data and the *SEM* if you are reporting the variability of the mean. Some texts and journals recommend that you always report the *SD* since the *SEM* can be calculated easily from the *SD* and sample size, and the *SEM* may give an uninformed reader a false impression about the variability of the data. If you are unsure, the safe bet is to report the *SD*. (In either case, you should make it clear which statistic you are reporting.)

Use Tables and Figures to Report Many Descriptive Statistics

If you are reporting two or three descriptive measurements in a report or article, we recommend that you include the statistics in the text. However, if you have more than two or three measurements, consider using a table or graph.

Break Down Descriptive Statistics by Group

Descriptive statistics should be broken down by group (i.e., calculated separately for each group) for populations composed of distinct groups rather than looking at aggregate data. For example, a mixture of normal subpopulations will usually not have a normal appearance, and an overall mean or standard deviation may be meaningless.

Quantitative Data Description Examples

The following examples illustrate techniques for describing quantitative data. The results include both statistics and graphs (when the techniques covered include a graph or graphs). More detailed information on creating graphs is included in the upcoming Chapter 3: Creating and Using Graphs.

EXAMPLE 2.1

Quantitative Data With an Unusual Value

Describing the Problem

Hypothetical data from several branch banks in Southern California contain information on how many IRAs (individual retirement accounts) were set up in 19 locations during a 3-month period. The variable is called *IRASetup* (labeled IRA Setup). These data are counts and are appropriately classified as quantitative data since, for example, it makes sense to calculate a mean number of accounts per bank. Before calculating and reporting the mean or other parametric measures of these values, you may want to assess the normality of the data. One way to do that is to perform a statistical test.

SPSS Step-by-Step. EXAMPLE 2.1: Quantitative Data With an Unusual Value

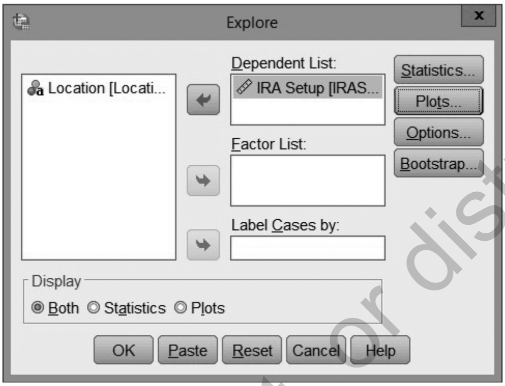
Use the following steps to obtain the output for the IRA data analysis:

1. Open the data set IRA.SAV (**File/Open**). The data for this example are shown in Table 2.1.
2. Select **Analyze/Descriptive Statistics/Explore**. . . .
3. Select the *IRASetup* variable from the dependent list by clicking on the variable name *IRASetup* and then clicking on the arrow next to the Dependent List box (or click on the variable name, *IRASetup*, and drag it into the Dependent List box). (See Figure 2.1.)

TABLE 2.1 ⬢ Data in the IRA.SAV Data Set

Location	IRASetup
U	10
U	12
U	15
R	14
U	16
U	14
R	13
U	12
U	16
R	12
U	13
R	14
R	15
R	17
R	5
R	14
U	13
R	14
U	11

FIGURE 2.1 ⬢ Select IRASetup for the Dependent List



Explore

Location [Locati...

Dependent List:
IRA Setup [IRAS...

Factor List:

Label Cases by:

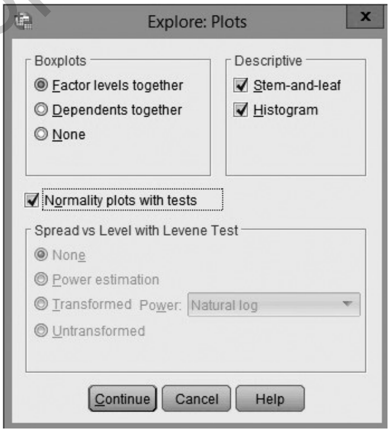
Statistics... Plots... Options... Bootstrap...

Display
● Both ○ Statistics ○ Plots

OK Paste Reset Cancel Help

4. Click on the Plots button and select the “Normality plots with tests” checkbox and “Histogram” and click Continue. (See Figure 2.2.)

FIGURE 2.2 ⬢ Completed Dialog Box for Normality Tests



Explore: Plots

Boxplots
● Factor levels together
○ Dependents together
○ None

Descriptive
☑ Stem-and-leaf
☑ Histogram

☑ Normality plots with tests

Spread vs Level with Levene Test
● None
○ Power estimation
○ Transformed Power: Natural log
○ Untransformed

Continue Cancel Help

(Continued)

(Continued)

5. Click OK. Table 2.2 is displayed (along with other output).

TABLE 2.2 ⬢ Test for Normality on IRA Data

	Tests of Normality					
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
IRA Setup	.173	19	.136	.878	19	.019

[a] Lilliefors Significance Correction

The "Sig." values in the "Tests of Normality" table are the *p*-values based on testing the null hypothesis that the data are normally distributed. Both tests are designed to determine whether the observed data closely fit the shape of a normal curve. The Shapiro-Wilk test result is significant ($p = 0.019$), which suggests that the data are not normal, while the Kolmogorov-Smirnov test result is nonsignificant ($p = 0.136$). This leaves you without a convincing argument one way or another.

To further examine these data (and perhaps understand the reasons for the discrepancy), you can visualize the distribution of the data using graphical displays such as a histogram, a normal Q-Q plot, a detrended normal Q-Q plot, and a box-and-whiskers plot. (See Figure 2.3.)

Here is a brief explanation of how to interpret each of these plots in the context of normality:

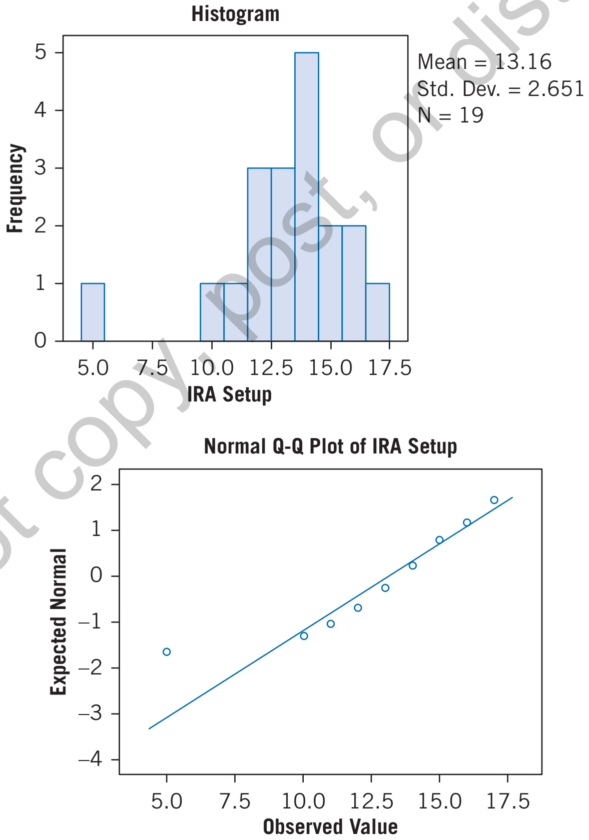
- *Histogram* (upper left). When a histogram's shape approximates a bell curve, it suggests that the data may have come from a normal population.

- *Q-Q plot* (upper right). A quantile-quantile (Q-Q plot) is a graph used to display the degree to which the quantiles of a reference (known) distribution (in this case, the normal distribution) differ from the sample quantiles of the data. When the data fit the reference distribution, then the points will lie in a tight random scatter around the reference line. For the IRA data, the curvature of the points in the plot indicates a possible departure from normality, and the point lying outside the overall pattern of points indicates a possible outlier.

- *Detrended normal Q-Q plot* (lower left). A version of the Q-Q plot that shows the differences between the observed normal and observed values under the assumption of normality. When the data are normal, the points cluster in a random horizontal band around zero. As in the Q-Q plot, the curve and the one outlying point may be of concern when assessing normality for these data.

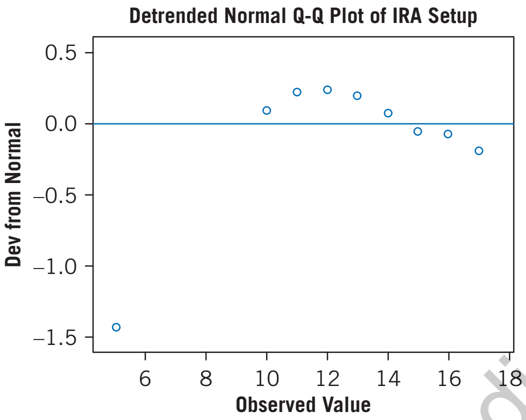
- *Boxplot* (lower right). A boxplot that is symmetric with the median line in

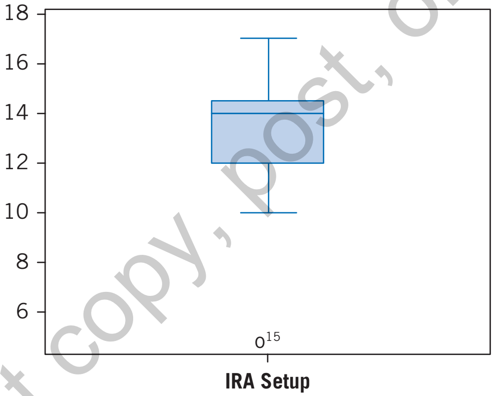
FIGURE 2.3 ⬢ Plots Used to Assess Normality Using *IRA Setup* Data



(Continued)

(Continued)



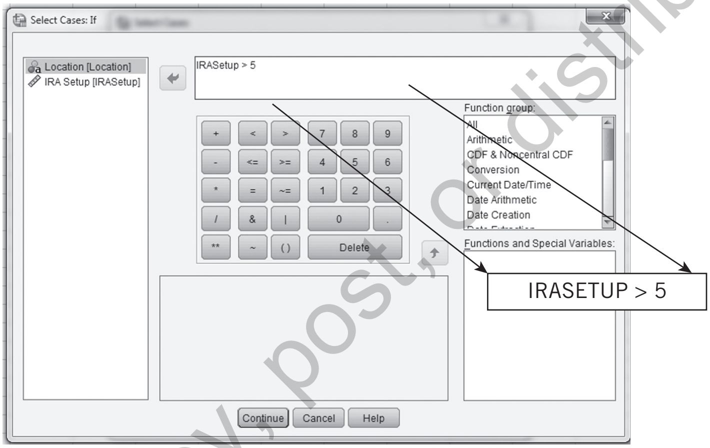


approximately the center of the box and with symmetric whiskers somewhat longer than the subsections of the center box suggests that the data may have come from a normal distribution.

All four of these plots for *IRASetup* indicate something unusual. There is one small value (Case 15, as indicated by the boxplot) that is a potential outlier (unusually small). Suppose that after carefully examining the source documentation for these data, you discover that the branch location for Case 15 was closed for 21 days because of localized wildfires. After this type of discovery, and observing that *IRASetup* = 5 for that branch, you might be justified in excluding this branch location from your analysis.

6. To eliminate the outlying value (*IRASetup* = 5), return to the data editor by clicking Windows and IRA.SAV or choosing the IRA.SAV window using the SPSS icon at the bottom of your screen and select **Data/Select Cases** . . . and select the option "If condition is satisfied. . . ." Click on the "If . . ." button. In the formula text box, enter the expression "*IRASetup* > 5" (without quotes). (See Figure 2.4.)

FIGURE 2.4 ● Select if Condition Is Satisfied



IRASetup > 5

IRASETUP > 5

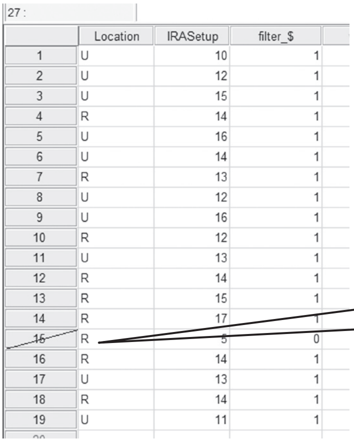
Click Continue and OK. A slash appears in the IRA data file next to record 15, indicating that the record will not be included in subsequent analyses. (See Figure 2.5.) (For more information about filtering cases, see "Transforming, Recoding, and Categorizing Your Data" in Appendix A.)

7. To display the output based on the revised data select **Analyze/Descriptive Statistics/Explore** . . . and OK. (SPSS remembers your previous selection of the variable and options.) Locate the histogram in the output and double-click on it. A Chart Editor will appear. From this editor, select the menu options **Elements/Show Distribution Curve**, and Close. Select **File/Close** to close the Chart Editor. The revised histogram with a superimposed normal curve is displayed in Figure 2.6.

(Continued)

(Continued)

FIGURE 2.5 ⬢ Slash Indicates Item 15 Will Not Be Used in Analyses

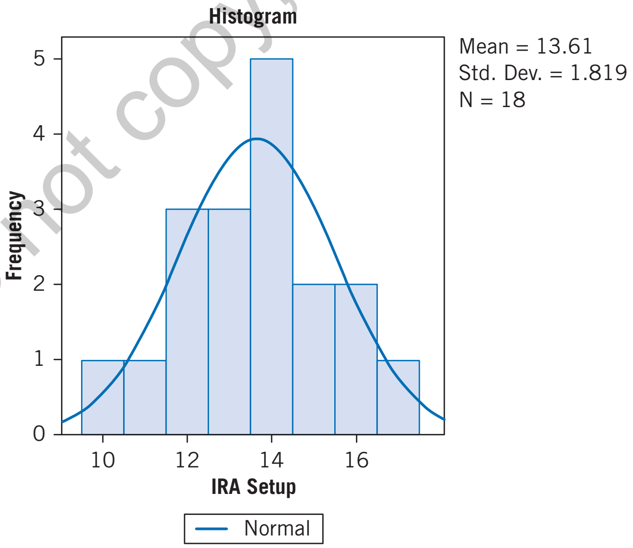


27 :

	Location	IRASetup	filter_$
1	U	10	1
2	U	12	1
3	U	15	1
4	R	14	1
5	U	16	1
6	U	14	1
7	R	13	1
8	U	12	1
9	U	16	1
10	R	12	1
11	U	13	1
12	R	14	1
13	R	15	1
14	R	17	1
~~15~~	R	5	0
16	R	14	1
17	U	13	1
18	R	14	1
19	U	11	1

Note the slash on record 15 indicating that it will be ignored during subsequent analyses.

FIGURE 2.6 ⬢ Revised Histogram With Normal Curve to Assess Normality in the *IRA Setup* Data



When the outlying data value is excluded from the data set and the data are reanalyzed, the Shapiro-Wilk test yields a p-value of 0.73, and the Kolmogorov-Smirnov p-value is greater than 0.2, indicating that there is no reason to be concerned about the normality assumption. Furthermore, after the removal of the extreme data value, the revised histogram looks considerably more normal (as does the boxplot). The histogram for the revised data is shown in Figure 2.6. The superimposed normal curve helps you assess the normality assumption. Although not a perfect fit, the histogram suggests that it is reasonable to assume that the data are from a normal population. The boxplot, not shown here, is relatively symmetric and typical of normally distributed data containing no outliers or extreme values. The other plots also suggest normality. The point of this example is to show that it is important to not only look at statistics and tests but also to look at graphical displays based on your data.

8. To remove the "Select cases" criterion select **Data/Select Cases** . . . and select the option "All cases" and OK

Once you are satisfied that a normality assumption makes sense for your data, you can use statistics such as means, standard deviations, and so on to describe your data. In the example here, note that we were justified in removing the extreme value. In addition to reporting the sample mean of the data, you may also want to report a confidence interval. For the modified dataset in this example, a 95% CI (confidence interval) on the mean is given in the Descriptives table as (lower bound, upper bound) or [12.71, 14.52]. Alternatively, you may also choose to report 13.6 ± 1.8 (mean ± *SD*). Why should you choose one over the other? If you wish to report the *precision* of your estimate, you should report the 95% CI. The interpretation of the 95% CI is that if you repeated this experiment many times, the true mean would fall within the calculated endpoints approximately 95% of the time. If you want to describe the variability of your data, you would use the expression mean ± *SD* (see Lang & Secic, 1997).

SIDEBAR

This example does not imply that you should always remove extreme values from your data set. You should carefully consider any unusual values and determine whether they are valid observations before removing them from your analysis. (It is a good practice to report any data values excluded from an analysis in your write-up and to justify your actions.)

(Continued)

(Continued)

Reporting Results for EXAMPLE 2.1: Quantitative Data With an Unusual Value

The results for the analyses on the IRA data could be reported in the following ways:

Narrative for the Methods Section

> "One value in the data for IRA setups was eliminated because the bank was closed for 21 days during the evaluation period. Descriptive statistics were calculated on the remaining 18 values."

Narrative for the Results Section

> "IRA setups averaged 13.6 per branch (*SD* = 1.8, *N* = 18)."

or

> "The mean (± *SD*) IRA setups was 13.6 (± 1.8)."

If you are reporting the precision of your estimate, you could state the following:

> "The mean was 13.6 (95% CI = 12.71 to 14.52) IRA setups per branch."

If you decide that you want to analyze the data *without* removing the extreme value and therefore not make a normality assumption, you could report your findings using the median and interquartile range. For example,

> "The number of IRA setups per branch ranged from 5 to 17 setups with a median (interquartile range) of 14 (3)."

EXAMPLE 2.2

Quantitative Data by Groups

Describing the Problem

A survey was administered to 79 clinic patients to measure their satisfaction with clinic services. Two versions of the survey were randomly assigned to the participants. The following demographic information (by group) is shown in Table 2.3. This table is not created in SPSS but is created from information gathered from multiple SPSS analyses. The upcoming example will show how to compile the information needed to create this table.

The purpose of this table is to compare the respondents to the two types of surveys on several important demographic variables. The *p*-values refer to a test of the null hypothesis that the means are equal (see the two-sample *t*-test examples in Chapter 4: Comparing One or Two Means Using the *t*-Test). In this table, it appears that the number of years of schooling was significantly higher for those who took the old survey.

Researchers disagree over whether to include the "*p*-Value" column in tables of this type. The controversy arises in part over the known problem of performing multiple tests within the same experiment. For a discussion of *p*-values, see Chapter 1.

As described in Chapter 1, when multiple *p*-values are used in an analysis, it is good practice to use Bonferroni-adjusted significance criteria. In this case, the adjustment would entail using the *p*-value 0.05/4 = 0.0125 as the rejection criterion for these tests. Thus, the only significant result in this table would be for the schooling variable (where the reported *p*-value is 0.003).

TABLE 2.3 ⬢ Table Reporting Group Statistics: Baseline Characteristics of Patients in Study by Group

	Mean (*SD*)		
	Old Survey	New Survey	
Characteristic	*N* = 34	*N* = 45	*p*-Value
Age	34.3 (11.6)	30.2 (9.21)	0.08
Schooling	12.2 (2.43)	10.5 (2.33)	0.003
Minutes	532 (337)	429 (309)	0.16

SPSS Step-by-Step. EXAMPLE 2.2: Quantitative Data by Groups

To calculate descriptive statistics needed to compile the information in Table 2.3, follow these steps in SPSS.

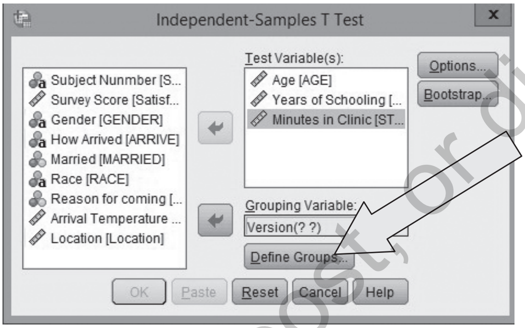
1. Open the data set SURVEY.SAV and select **Analyze/Descriptive Statistics/Explore**. . . .
2. Select *Age, Years of Schooling (Edu)*, and *Minutes in Clinic* (*Stayminutes*) for the dependent variables and *Survey Version* for the Factor List. Click OK to produce output that includes the means and standard deviations. This output is pretty messy and hard to read.

(Continued)

(Continued)

3. To calculate the *p*-values needed for the table, select **Analyze/Compare Means/ Independent-Samples T Test** and select *Age*, *Years of Schooling (Edu)*, and *Minutes in Clinic* (*Stayminutes*) as the "Test Variables" and *Survey Version* as the "Grouping Variable." This dialog box is shown in Figure 2.7.

FIGURE 2.7 ● Select the Define Groups Button

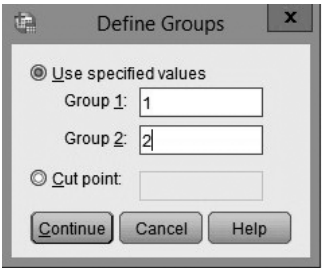


Test Variable(s):
Age [AGE]
Years of Schooling [...
Minutes in Clinic [ST...

Grouping Variable:
Version(? ?)

Click on the "Define groups" button and enter 1 and 2 for the group values as shown in Figure 2.8. (Because the coded values for Survey Type is 1 and 2.)

FIGURE 2.8 ● Define Groups



● Use specified values
Group 1: 1
Group 2: 2
○ Cut point:

Click Continue and OK. In the resulting "Independent Samples" table, the *p*-values for each comparison are listed in the "Equal variances assumed" row in the "Sig. 2-tailed" column. For example, the *p*-value for the comparison of mean ages by group, $p = 0.084$ (reported as 0.08 in Table 2.3) is shown in Table 2.4.

More about the *t*-test and this table is discussed in Chapter 4: Comparing One or Two Means Using the *t*-Test.

TABLE 2.4 ⬢ Results of *t*-Tests (Partial output table shown here)

Independent Samples Test						
		Levene's Test for Equality of Variances				
		F	Sig.	t	df	Sig. (2-tailed)
Age	Equal variances assumed	3.543	.064	1.749	77	.084
	Equal variances not assumed			1.694	61.455	.095
Years of Schooling	Equal variances assumed	2.566	.113	3.045	77	.003
	Equal variances not assumed			3.028	69.606	.003
Minutes in Clinic	Equal variances assumed	.014	.907	1.411	77	.162
	Equal variances not assumed			1.394	67.630	.168

(Continued)

(Continued)

Reporting the Results for EXAMPLE 2.2: Quantitative Data by Groups

Narrative for the Methods Section

> "The two versions of the survey were randomly assigned to patients as they registered at the clinic."

Narrative for the Results Section

When there are several means to report, it is often clearer to the reader if you report the results in a table such as the one shown in Table 2.3. If you use a Bonferroni-adjusted *p*-value for your rejection criterion, you should include a statement such as the following:

> "To maintain the a = .05 significance level for the table comparisons using a Bonferroni adjustment, a *p*-value must be less than *p* = 0.0125 (i.e., 0.05/4) to be considered statistically significant."

In either case, the *schooling* difference remains the only significant comparison. We assume that the observed difference of 1.7 years of schooling is a meaningful difference. For more information about using effect size to interpret this difference, see Chapter 4.

DESCRIBING CATEGORICAL DATA

Categorical variables record a characteristic about a subject or object such as race, gender, presence of a disease, or the color of a car model. Think of the word *categories* when you're analyzing categorical data. This data type is sometimes called *qualitative* and is further divided by the terms *nominal* (order not important) and *ordinal* (having order). We will assume categorical data to be nominal unless specified. In this section, two types of descriptive analyses are illustrated for categorical data. They are as follows:

- Frequency tables
- Crosstabulations

Examples illustrate methods for examining categorical data and reporting your results. (Categories can be defined using number or character codes.)

Categorical variables include the following:

- Presence of a disease (1 = yes, 0 = no)
- Method of delivery (USPS, UPS, FEDEX, OTHER)
- Marital status (1 = married, 2 = single never married, 3 = single divorced, 4 = single widowed)
- Stage of a disease (1, 2, 3, or 4)

SIDEBAR

When creating a table for publication, you should check the specifications for the particular publication to which you will be submitting your article.

Each of these variables is an observation that places the subject or entity into two or more categories. When we observe summaries of these variables, they are typically given as counts (the number of subjects placed into each category) and/or the corresponding percentages.

Considerations for Examining Categorical Data

Notice in the examples above that some categories have order and some do not. For example, in the marital status variable above, there are four unordered categories, and although these categories are recorded in the data set as numbers (1 to 4), these numbers are simply codes and are not meaningful numerically. They should not be interpreted as implying an ordering of the categories, nor should they be used in arithmetic calculations (e.g., mean, etc.). For the purposes of analysis, these categorical data are reported as counts, frequencies, or percentages of subjects falling into various categories. In the marital status example, the categories have no order and are thus nominal variables. On the other hand, stage of cancer (from 1 to 4), rank in the military, and finish order in a race (first place, second place, etc.) are ordinal categorical variables since these categories have a logical order.

Tips and Caveats

When Should Categorical Variables Be Treated as Quantitative Data?

Sometimes, ordinal data such as responses using a Likert scale (1 = *strongly disagree*, 2 = *disagree*, 3 = *neutral*, 4 = *agree*, 5 = *strongly agree*) are incorrectly treated as quantitative data. As a general guideline, in order for the mean and other arithmetically obtained quantities (e.g., standard deviation) to make sense, it must be reasonable to assume that the differences between any two categories are equal. For example, if the categories are 1, 2, 3, 4, 5, then it must be reasonable to consider the difference between categories 2 and 3 to be the same as the difference between 4 and 5, and so on. Therefore, we recommend that it is rarely the case that you treat categorical data as quantitative since the assumption of equal distance between categories cannot usually be assumed.

Describing Categorical Data Examples

The following examples illustrate techniques for assessing and describing categorical data using statistics and graphs.

EXAMPLE 2.3

Quantitative Data With Unusual Values

Describing the Problem

Suppose you are interested in exploring the variables in a data set from the U.S. Department of Energy (2014) containing fuel economy information on 2014 model year automobiles. Your first strategy might be to look for unusually large or small data values by finding the minimum and maximum for each quantitative variable of interest.

SPSS Step-by-Step. EXAMPLE 2.3: Quantitative Data With Unusual Values

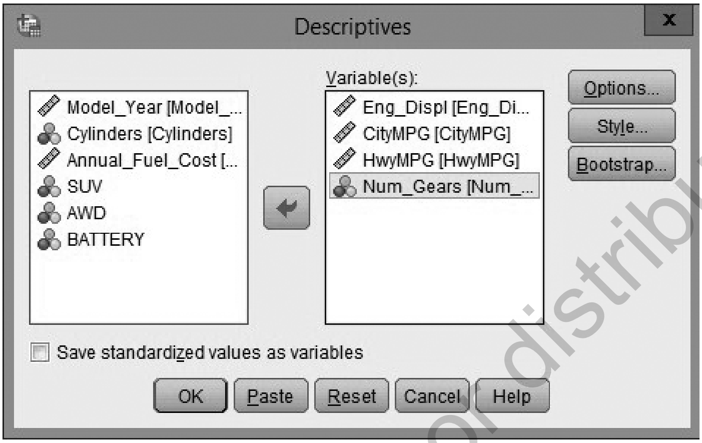
1. Open the data set CARS2014.SAV and select **Analyze/ Descriptive Statistics/ Descriptives**. . . .

2. Select the variables *Eng_Displ, CityMPG*, and *HwyMPG*, and *Num_Gears*. (See Figure 2.9.)

3. Click OK, and the table in Table 2.5 is displayed.

The results are shown in Table 2.5 for four of the variables. In this output, the minimum value for the *Num_Gears* (*Number of Gears*) variable is 1. This seems odd that a car would have only one gear, but by examining the data, it can be seen that several cars have Continuously Variable Transmission (CVT), which does not have typical gears like most cars. The data are correct, since CVT transmissions are different from conventional transmissions. That information might be handled differently when looking at the transmission data. In fact, we might want to (at least temporarily) label the cars with CVT transmission as "missing" when calculating statistics

FIGURE 2.9 ⬢ Select Variables to Analysis in Descriptives Procedure



Model_Year [Model_...
Cylinders [Cylinders]
Annual_Fuel_Cost [...
SUV
AWD
BATTERY

Variable(s):
Eng_Displ [Eng_Di...
CityMPG [CityMPG]
HwyMPG [HwyMPG]
Num_Gears [Num_...

□ Save standardized values as variables

TABLE 2.5 ⬢ Searching for Unusual Values

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
Eng_Displ	1155	1.00	8.40	3.3497	1.39504
CityMPG	1155	8	53	19.85	5.678
HwyMPG	1155	13	48	27.48	6.360
Num_ Gears	1155	1	9	6.17	1.441
Valid N (listwise)	1155				

Source: Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation

(Continued)

(Continued)

SIDEBAR

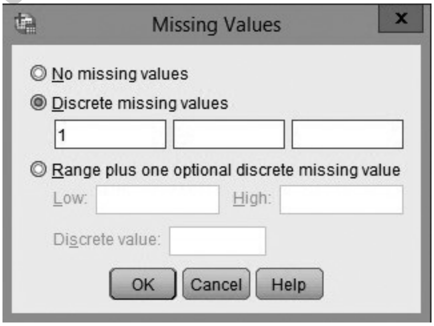
Checks for extreme values will not find a slight miscoding of a data value. For example, if *CityMPG* were entered as 32 instead of 23, that mistake would not be evident in the listing of minimum and maximum values.

involving an analysis concerning the number of transmission gears. If your study deals only with vehicles with conventional geared transmissions, you may want to exclude the CVT cars from your analysis. (Assigning missing values and selecting cases in SPSS are discussed in Appendix A: A Brief Tutorial for Using IBM SPSS for Windows.)

Continuing with the automobile data example, suppose you want to compare *CityMPG* between SUVs and non-SUVs in the 2014 automobile data set (excluding hybrids and CVT transmission vehicles), and you want to investigate whether the normality assumption makes sense for each group. The following shows how to look at these data:

4. To set the *Num_Gears* missing value at 1, make sure you are in the CARS2014 data window, and click on the Variable View tab at the bottom left. When the Variable View is displayed, click on the cell for *Num_Gears* in the Missing column. Click on the ellipses (. . .), click the Discrete missing values option, and enter 1 as a discrete missing value as shown in Figure 2.10, and click OK.

FIGURE 2.10 Indicate Missing Value



Missing Values

○ No missing values
● Discrete missing values
1
○ Range plus one optional discrete missing value
Low: High:
Discrete value:

5. As in Step 1, select **Analyze/Descriptive Statistics/Descriptives** . . . and OK, and examine the Minimum value for *Num_Gears*. It is now 4 instead of 1, and the number of nonmissing cases for that variable is now 1,108 instead of 1,155 because all cars where *Num_Gears* = 1 have been left out of that calculation as shown in Table 2.6. (Setting this missing value for Num_Gears did not change the means of the other variables.)

TABLE 2.6 Revised Analysis of Car Data

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Eng_Displ	1155	1.00	8.40	3.3497	1.39504
CityMPG	1155	8	53	19.85	5.678
HwyMPG	1155	13	48	27.48	6.360
Num_Gears	1108	4	9	6.39	.990
Valid N (listwise)	1108				

6. To examine the data further, create side-by-side boxplots of the car data comparing SUVs and non-SUVs. Select **Graphs/Legacy Dialogs/Boxplot** . . . , select the "Simple" option, and choose the radio button for "Summaries for groups of cases." (Other graph options are covered in Chapter 3: Creating and Using Graphs.)

7. Click the Define button. Select *CityMPG* as the "Variable," *SUV* as the "Category Axis," and Carline for "Label Cases by" as shown in Figure 2.11.

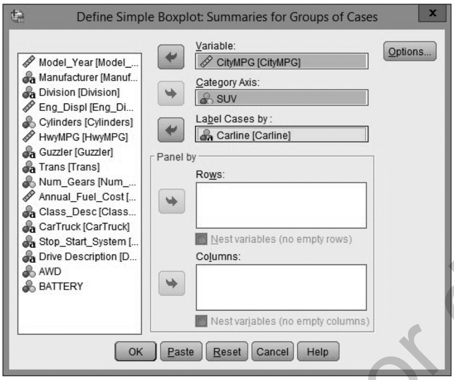
8. Click OK to produce the comparative boxplots in Figure 2.12.

The graph in Figure 2.12 shows side-by-side boxplots that indicate that some observations (on the high end of *CityMPG*) are indicated as outliers (marked as an "o") and some as extreme values (marked as an "*"). An outlier is defined

(Continued)

(Continued)

FIGURE 2.11 ● Select Variables for (Legacy) Comparative Boxplot



Define Simple Boxplot: Summaries for Groups of Cases

- Model_Year [Model_...
- Manufacturer [Manuf...
- Division [Division]
- Eng_Displ [Eng_Di...
- Cylinders [Cylinders]
- HwyMPG [HwyMPG]
- Guzzler [Guzzler]
- Trans [Trans]
- Num_Gears [Num_...
- Annual_Fuel_Cost [...
- Class_Desc [Class...
- CarTruck [CarTruck]
- Stop_Start_System [...
- Drive Description [D...
- AWD
- BATTERY

Variable: CityMPG [CityMPG]

Category Axis: SUV

Label Cases by: Carline [Carline]

Panel by

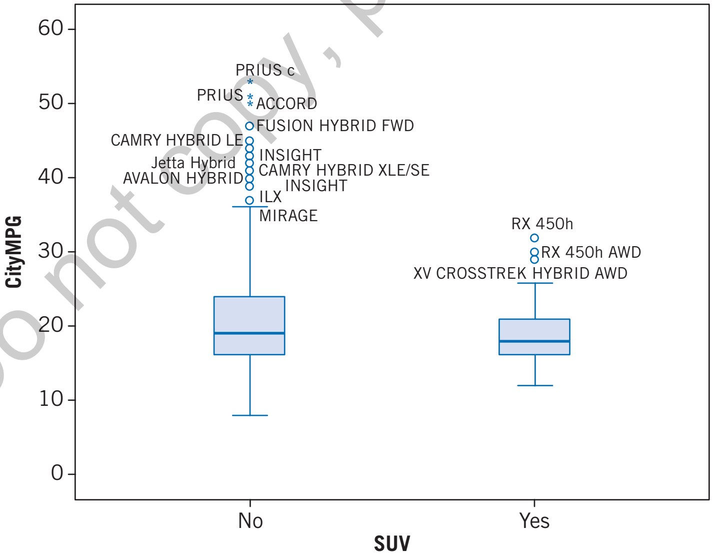
Rows:

□ Nest variables (no empty rows)

Columns:

□ Nest variables (no empty columns)

FIGURE 2.12 ● Side-by-Side Boxplots Showing Outliers and Extreme Values by Group



(in SPSS) as a value from 1.5 to 3 interquartile ranges (IQRs) beyond the 75th (or below the 25th) percentile, and an extreme value is greater than 3 IQRs beyond the 75th (or below the 25th) percentile. Figure 2.12 shows that there are a number of outliers for non-SUV automobiles, but the data for SUVs are less variable, with only a few extreme (high *CityMPG*) values. (Note that the outliers on the chart are hybrid models.)

Another way to display this boxplot plot is by using the Chart Builder. We'll show a boxplot example using Chart Builder in Chapter 3.

EXAMPLE 2.4

Frequency Table for Categorical Data

Describing the Problem

In a survey of 79 patients at a clinic, information was collected on how they arrived (car, bus, or walked). You may want to examine the data using a frequency table to report the number and percentage of patients who arrived using each travel method along with a bar chart showing a visualization of these percentages. To display this information, follow the steps in this example:

SPSS Step-by-Step. EXAMPLE 2.4: Frequency Table for Categorical Data

1. Open the data set SURVEY.SAV and select **Analyze/ Descriptive Statistics/ Frequencies**. . . .
2. Select *How Arrived* (*ARRIVED*) as the variable.
3. Click on the Charts button, and select **Bar Chart, Percentages, Continue**, and OK. See Figure 2.13.

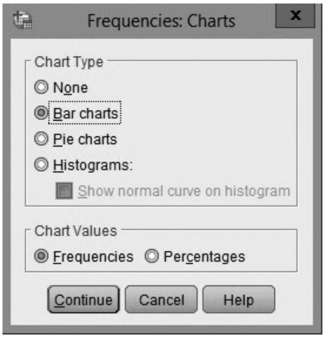
Table 2.7 displays the frequency table produced by SPSS for this data set, and Figure 2.14 shows the associated bar chart. Note that the bar chart information is displayed in percentages rather than counts. We could optionally have produced a similar chart using frequencies (counts).

In Table 2.7, the frequency is the number of patients who arrived using each method of transportation. The percentage (and valid

(Continued)

(Continued)

FIGURE 2.13 ⬢ Choose Options for Bar Chart



Frequencies: Charts

Chart Type
○ None
● Bar charts
○ Pie charts
○ Histograms:
□ Show normal curve on histogram

Chart Values
● Frequencies ○ Percentages

TABLE 2.7 ⬢ Frequency Table for *How Arrived*

How Arrived					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	BUS	11	13.9	13.9	13.9
	CAR	66	83.5	83.5	97.5
	WALK	2	2.5	2.5	100.0
	Total	79	100.0	100.0	

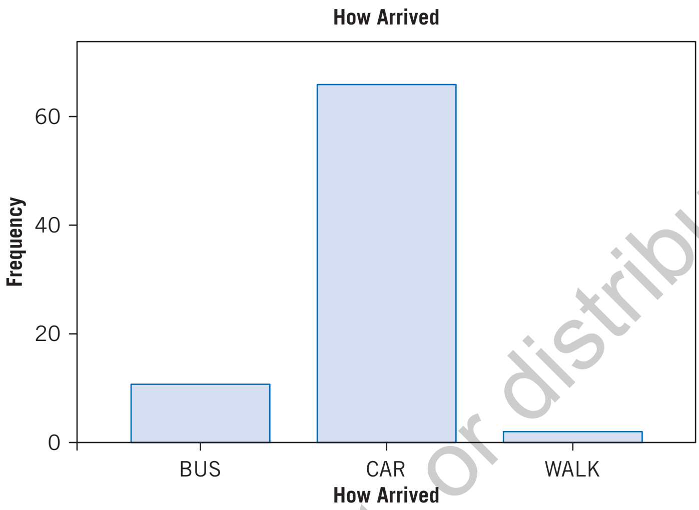
percentage, which is the percentage after removing missing values) tells you the percentage of patients arriving by each method. It is clear both from the frequency table and the bar chart that most patients arrived by car.

Reporting Results for Frequency Data

Narrative for the Methods Section

> “Arrival methods were examined by finding the number of patients arriving at

FIGURE 2.14 ⬢ Bar Chart for the *How Arrived* Variable



the clinic using a car, bus, or by walking."

Narrative for the Results Section

"Patients arrived by car 83.5% of the time (66 of 79), 13.9% by bus (11 of 79), and 2.5% by walking (2 of 79). (Round-off error makes the total slightly under 100%.)"

EXAMPLE 2.5

Crosstabulation of Categorical Variables

Describing the Problem

Using the 2014 automobile data, suppose you want to crosstabulate two variables. Table 2.8 contains a crosstabulation of the variables *SUV* (sport utility vehicle) and *AWD*

(Continued)

(Continued)

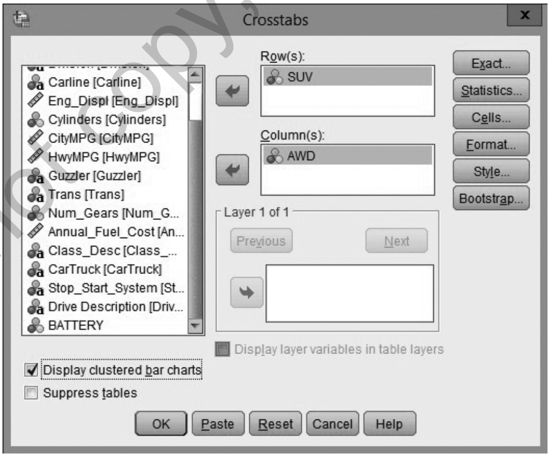
(all-wheel drive). Specific statistical tests to analyze this type of table are discussed in Chapter 6: Analysis of Categorical Data. In this example, we are only interested in the descriptive information contained in the table.

SPSS Step-by-Step. EXAMPLE 2.5: Crosstabulation of Categorical Variables

Follow these steps to create the crosstabulation output from the CARS2014 data set:

1. Open the data set CARS2014.SAV, and select **Analyze/Descriptive Statistics/Crosstabs**. . . .
2. From the dialog box, select the variable named *SUV* as the row variable and *AWD* as the column variable.
3. To display the bar chart, select the “Display clustered bar charts” checkbox (as shown in Figure 2.15).
4. Click on the Cells button and select **Row, Column, and Total Percentages** as shown in Figure 2.16. Click Continue.
5. Click OK to display the results as shown in Table 2.8.

FIGURE 2.15 ⬢ Select Variables for Crosstabulation Analysis



Crosstabs

Carline [Carline]
Eng_Displ [Eng_Displ]
Cylinders [Cylinders]
CityMPG [CityMPG]
HwyMPG [HwyMPG]
Guzzler [Guzzler]
Trans [Trans]
Num_Gears [Num_G...
Annual_Fuel_Cost [An...
Class_Desc [Class_...
CarTruck [CarTruck]
Stop_Start_System [St...
Drive Description [Driv...
BATTERY

Row(s): SUV

Column(s): AWD

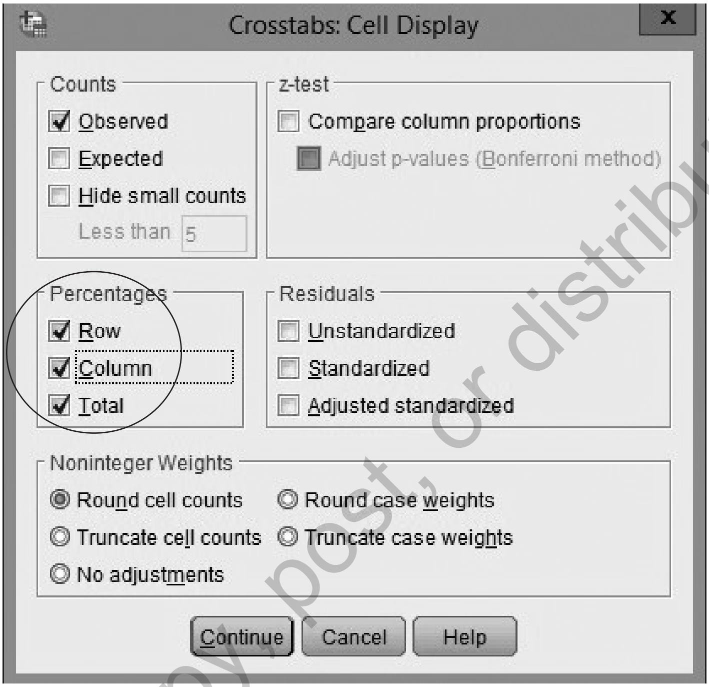
Layer 1 of 1

☐ Display layer variables in table layers

☑ Display clustered bar charts

☐ Suppress tables

FIGURE 2.16 ● Select Cell Options for Crosstabulation Example



Crosstabs: Cell Display

Counts
☑ Observed
□ Expected
□ Hide small counts
Less than 5

z-test
□ Compare column proportions
□ Adjust p-values (Bonferroni method)

Percentages
☑ Row
☑ Column
☑ Total

Residuals
□ Unstandardized
□ Standardized
□ Adjusted standardized

Noninteger Weights
● Round cell counts
○ Round case weights
○ Truncate cell counts
○ Truncate case weights
○ No adjustments

Observe that 33.0% (95 of 288) of the SUVs in this data set have all-wheel drive. Also note that 35.1% (95 of 271) of all-wheel-drive vehicles are SUVs and that 8.2% (95 of 1,155) of all vehicles are SUVs with all-wheel drive. A bar chart for these data is shown in Figure 2.17.

This bar chart allows you to visualize the relationship between SUVs and whether they have all-wheel drive. In this chart, it is visually clear that most non-SUVs do not have all-wheel drive, while a larger percentage of the vehicles classified as SUVs have all-wheel drive. You can also see that there are many more non-SUVs than there are SUVs in this data set.

(Continued)

(Continued)

TABLE 2.8 ● Output for Crosstabulation Example

SUV * AWD Crosstabulation					
			AWD		
			No	Yes	Total
SUV	No	Count	691	176	867
		% within SUV	79.7%	20.3%	100.0%
		% within AWD	78.2%	64.9%	75.1%
		% of Total	59.8%	15.2%	75.1%
	Yes	Count	193	95	288
		% within SUV	67.0%	33.0%	100.0%
		% within AWD	21.8%	35.1%	24.9%
		% of Total	16.7%	8.2%	24.9%
Total		Count	884	271	1155
		% within SUV	76.5%	23.5%	100.0%
		% within AWD	100.0%	100.0%	100.0%
		% of Total	76.5%	23.5%	100.0%

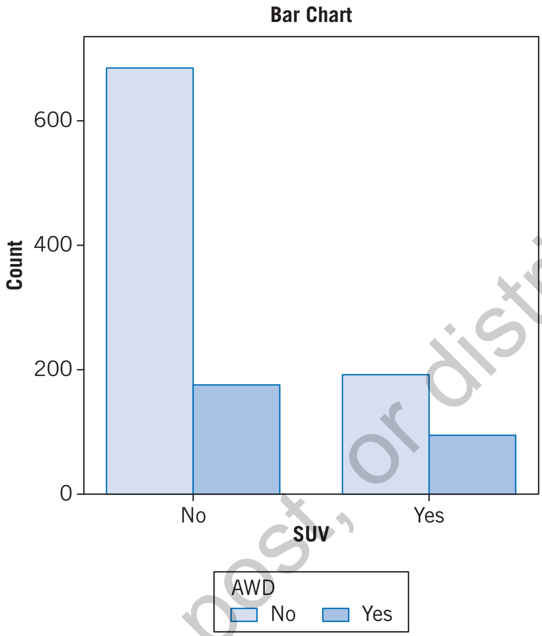
Reporting Crosstabulation Results

To describe the information in a frequency table or crosstabulation in your report or article, you should always include counts along with percentages (Lang & Secic, 1997). For example, your description (**from the information in the crosstabulation above**) might be as follows:

Narrative for the Methods Section

> "The relationship between model type and all-wheel drive was examined using crosstabulation."

FIGURE 2.17 ⬢ Clustered Bart Chart for SUV Data



Narrative for the Results Section

"This table shows that 33.0% (95 of 288) of SUV models and 20.3% (176 of 867) of non-SUV models have all-wheel drive."

Summary

Understanding your data is the first step in any data analysis. This chapter explains how to use descriptive statistics and graphs to understand and report information about your data. The next chapter continues with this subject, going deeper into how to display your data using SPSS graphs.

References

American Psychological Association (APA). (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Fay, M. P., & Proschan, M. A. (2010). Wilcoxon-Mann-Whitney or *t*-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, *4*, 1.

Lang, T. A., & Secic, M. (1997). *How to report statistics in medicine*. Philadelphia, PA: American College of Physicians.

Lang, T. A., & Secic, M. (2006). *How to report statistics in medicine*. Philadelphia, PA: American College of Physicians.

Sawilowsky, S. S. (2005). Misconceptions leading to choosing the *t* test over the Wilcoxon Mann-Whitney U test for Shift in Location Parameter. *Journal of Modern Applied Statistical Methods*, 4(2), 598–600.

U.S. Department of Energy. (2014). *Model 2014 model year fuel economy data*. Retrieved from http://www.fueleconomy.gov.