IN THIS CHAPTER

- Defining levels of measurement
- Looking at frequencies for categorical and continuous variables
- Using the descriptives procedure to summarize continuous variables

Chapter 14
Using Descriptive Statistics

Summaries of individual variables provide the basis for more complex analyses (as you see in the next few chapters). They also help establish base rates, answer important questions (for example, the percent of satisfied customers), allow users to check sample size and the data for unusual cases or errors, and provide insights into ways in which you may combine different groups. Ideally, you want to obtain as much information as possible from your data. In practice, given the measurement level of the variables, only some summary statistics are meaningful.

In this chapter, we begin by discussing level of measurement. Next, you run the frequencies procedure to obtain summary statistics for both categorical and continuous variables. Finally, you use the descriptives procedure to summarize continuous variables.

Looking at Levels of Measurement

The level of measurement of a variable determines the appropriate summary statistics and graphs that can be used to describe the data. For example, if you have a variable such as marital status, it wouldn't make sense to ask for the mean

of the variable; instead, you might ask for the percentages associated with the different categories. In addition, level of measurement determines the kind of research questions you can answer, so determining the level of measurement is a critical step in the research process.

REMEMBER

Choosing the appropriate summary statistic for each level of measurement is not the same as setting up the metadata in variable view.

The term *levels of measurement* refers to the coding scheme or the meaning of the numbers associated with each variable. Many statistical techniques are appropriate only for data measured at particular levels or combinations of levels of measurement. Different statistical measures are appropriate for different types of variables, and the statistical summaries depend on the level of measurement.

Defining the four levels of measurement

Introductory statistics textbooks present four levels of measurement, each defined by certain properties. Each successive level contains the properties of the preceding types as well as additional characteristics. The four levels of measurement are as follows:

- **Nominal:** For nominal data, each value represents a category, but the categories have no inherent order. For example, the variable eye color may be coded as 0 (blue), 1 (brown), 2 (black), and 3 (green), but these values tell you only that you have distinct categories, *not* that one category has more or less, or is better or worse than the other.
- **Ordinal:** For ordinal data, each value is a category and the categories have a meaningful order or rank, but there is no measurable distance between categories. For example, if you're measuring the outcome of a foot race, you can determine which contestant came in first, second, third, and so on. However, based on the ranking, you can't tell how much faster each competitor was compared to the others, nor can you say that the difference between first and second place is the same as the difference between second and third place. Other examples of ordinal variables are attitudinal questions with categories, such as strongly disagree (1), disagree (2), neutral (3), agree (4), and strongly agree (5), or variables such as income coded into categories representing ranges of values.
- **Interval:** Interval data has all the properties of ordinal variables, and in addition, a one-unit change in the numeric value represents the same change in quantity regardless of where it occurs on the scale. For example, for a variable such as

temperature measured in Fahrenheit, the difference between 20 degrees and 21 degrees (1 unit) is equal to the difference between 50 degrees and 51 degrees. In other words, these variables have equal intervals between points on the scale.

- **Ratio:** Ratio data has all the properties of interval variables with the addition of a true zero point, representing the absence of the property being measured. For example, temperature measured in Fahrenheit is measured on an interval scale because zero degrees does not represent the absence of temperature. However, a variable such as number of purchases is a ratio variable because a value of zero indicates no purchases. Ratios can then be calculated (for example, eight purchases represents twice as many purchases as four purchases).

TECHNICAL STUFF

These four levels of measurement are often combined into two main types:

- **Categorical:** Nominal and ordinal measurement levels where different values represent qualitative differences
- **Continuous (or scale):** Interval and ratio measurement levels where different values represent quantitative differences

REMEMBER

SPSS uses three levels of measurement: nominal, ordinal, and scale. *Scale variables* are interval- or ratio-level variables. SPSS does not differentiate between interval- and ratio-level variables because, from a statistical standpoint, no difference exists in the summary statistics you can obtain and in the statistical procedures you can perform for either variable.

Defining summary statistics

The most common way to summarize variables is to use measures of central tendency and variability:

- **Central tendency:** One number that is often used to summarize the distribution of a variable. Typically, we think of *central tendency* as referring to the average value. The three main measures of central tendency follow:
  - **Mode:** The category or value that contains the most cases — in other words, the most common value. This measure is typically used on nominal or ordinal data and can easily be determined by examining a frequency table.
  - **Median:** The midpoint of a distribution — in other words, the 50th percentile. If all the cases for a variable are arranged in order according to their value from lowest to highest, the median is the value that splits the data into two

equally sized groups. The median is most useful when there are extreme values, such as when analyzing home sales prices.

- **Mean:** The mathematical average of all the values in the distribution — that is, the sum of the values of all cases divided by the total number of cases. The mean is the most common measure of central tendency in statistical tests.
- **5% Trimmed Mean:** The mathematical average of all the values in the distribution after the upper 5 percent and the lower 5 percent of data has been removed. Thus, the 5% trimmed mean is calculated on the middle 90 percent of the distribution. The 5% trimmed mean is most useful for continuous data with outliers.

» **Variability:** The amount of spread or dispersion around the measure of central tendency. Several measures of variability are available:

- **Maximum:** The highest value for a variable.
- **Minimum:** The lowest value in the distribution.
- **Range:** The difference between the maximum and minimum values. It provides a broad sense of the distribution, but it is affected by outliers.
- **Interquartile Range:** The difference between the 75th and 25th percentile values. It is the range for the middle 50 percent of the sample and it is not affected by outliers.
- **Variance:** Provides information about the amount of spread around the mean value. It's an overall measure of how clustered data values are around the mean. The variance is calculated by summing the square of the difference between each value and the mean and dividing this quantity by the number of cases minus one. In general terms, the larger the variance, the more spread there is in the data; the smaller the variance, the more the data values are clustered around the mean.
- **Standard deviation:** The square root of the variance. The variance is expressed in the units of the variable squared. Thus, if you were looking at the variability of the number of apples sold at a supermarket from day to day, the units of the variation would be apples^2. This squared unit is difficult to interpret, so the standard deviation restores the unit of variability to the units of measurement of the original variable.

In conclusion, we care about level of measurement because it determines which summary statistics and graphs we should use to describe the data. Table 14-1 summarizes the most common summary statistics and graphs for each of the measurement levels used by SPSS.

TABLE 14-1 Level of Measurement and Descriptive Statistics

	Nominal	Ordinal	Scale
Definition	Unordered categories	Ordered categories	Numeric values
Examples	Gender, geographic location, job category	Satisfaction ratings, income groups, ranking of preferences	Number of purchases, cholesterol level, age
Measures of central tendency	Mode, count, percentages	Mode, count, percentages, median	Mode, median, mean, trimmed mean
Measures of dispersion	None	Minimum, maximum, range, interquartile range	Minimum, maximum, range, standard deviation, variance
Graph	Pie or bar	Pie or bar	Histogram, box and whisker plot

Focusing on Frequencies for Categorical Variables

The most common technique for describing categorical data — nominal and ordinal levels of measurement — is to request a *frequency table*, which provides a summary showing the number and percentage of cases falling into each category of a variable. Users can also request additional summary statistics such as the mode or median, among others.

Here's how to run the *frequencies procedure* so you can create a frequency table that will allow you to obtain summary statistics for categorical variables:

1. **From the main menu, choose File ➪ Open ➪ Data and load the Merchandise.sav data file.**

   The file is not in the SPSS installation directory. You have to download it from this book's companion website at `www.dummies.com/go/spss4e`. The file contains customer purchase history and has 16 variables and 3,338 cases.

2. **Choose Analyze ➪ Descriptive Statistics ➪ Frequencies.**

   The Frequencies dialog appears.

   In this example, you want to study the distribution of the variables Payment_ Method (Auto Pay, Check, or Credit Card), Premier (Yes or No), and Status (Current or Churned). You can place these variables in the Variable(s) box and each will be analyzed separately.

3. **Select the Payment_Method, Premier, and Status variables, and place them in the Variable(s) box, as shown in Figure 14-1.**

   If you were to run the frequencies procedure now, you would get three tables, each showing the distribution of one variable. It's customary, though, to request additional summary statistics.

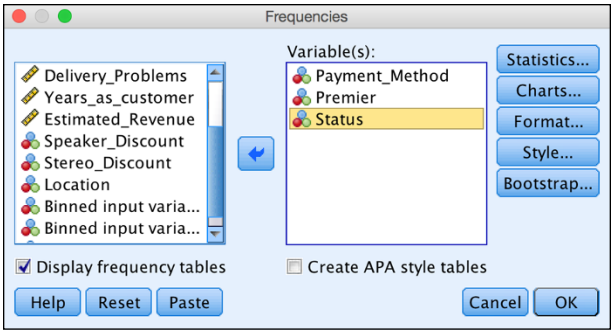

FIGURE 14-1: Place the variables in the Variable(s) box.

4. **Click the Statistics button.**

   The Frequencies: Statistics dialog appears.

5. **In the Central Tendency section, select the Mode check box, as shown in Figure 14-2.**

WARNING

This dialog provides many statistics, but it's critical that you request only those appropriate for the level of measurement of the variables you placed in the Variable(s) box. For nominal variables, the only suitable statistic is mode.

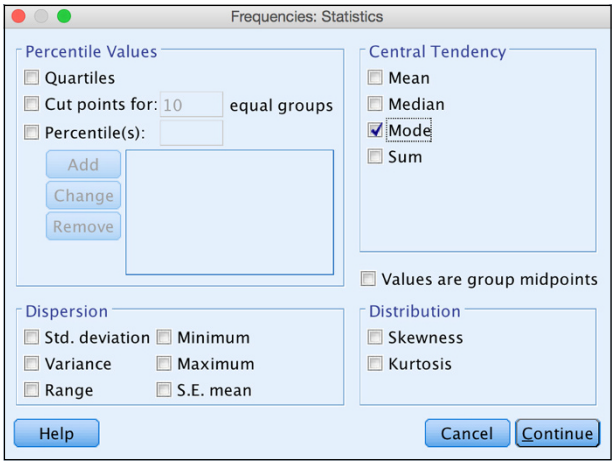

FIGURE 14-2: The Select the Mode check box.

6. **Click Continue.**

   Requesting a graph can be useful, so you can have a visual display of the data. That's what you'll do now.

7. **Click the Charts button.**

   The Frequencies: Charts dialog appears.

8. **In the Chart Type section, select the Bar Charts radio button; in the Chart Values section, select the Percentages radio button (see Figure 14-3).**

   This dialog has options for pie charts and bar charts. Either type of chart is acceptable for a nominal variable. Charts can be built using either counts or percentages, but normally percentages are a better choice.

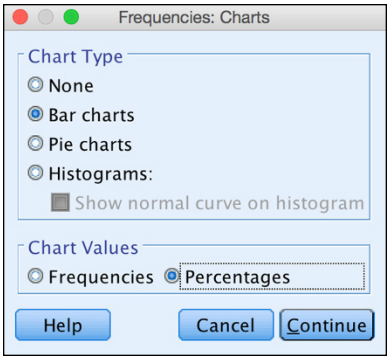

FIGURE 14-3: Select Bar Charts and Percentages.

9. **Click Continue, and then click OK.**

   SPSS runs the frequencies procedure and calculates the summary statistics, frequency table, and bar chart you requested.

The statistics table shown in Figure 14-4 displays the number of valid and missing cases for each variable requested in the frequencies procedure.

REMEMBER

Be sure to review this table to check the number of missing cases. In this example, you have 3,338 valid cases and no missing data.

The statistics table also displays any additional statistics that were requested. You asked only for the mode, the category that has the highest frequency, so only the mode is shown for each of the variables. In this example, the mode is represented by values of 3, 1, and 2, respectively, and denotes the category of Credit Card for Payment_Method, No for Premier, and Current group for Status.

Statistics

		Payment_Method	Premier	Status
N	Valid	3338	3338	3338
	Missing	0	0	0
Mode		3	1	2

FIGURE 14-4: The Statistics table for three variables.

The frequency table, which is shown in Figure 14-5, displays the distribution of the Payment_Method variable. (In this case, you focus on the Payment_Method variable because all other frequency tables will have similar information.) The information in the frequency table is comprised of counts and percentages.

Payment_Method

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Auto Pay	669	20.0	20.0	20.0
	Check	743	22.3	22.3	42.3
	Credit Card	1926	57.7	57.7	100.0
	Total	3338	100.0	100.0	

FIGURE 14-5: The frequency table for the Payment_Method variable.

The Frequency column contains *counts*, or the number of occurrences of each data value. So, for the Payment_Method variable, it's easy to see why the category Credit Card was the mode — 1,926 customers made purchases this way. The Percent column shows the percentage of cases in each category relative to the number of cases in the entire dataset, including those with missing values. In the example, those 1,926 customers who paid via credit card account for 57.7 percent of all customers. The Valid Percent column contains the percentage of cases in each category relative to the number of valid (nonmissing) cases. Because there is no missing data, the percentages in the Percent column and in the Valid Percent column are identical. The Cumulative Percent column contains the percentage of cases whose values are less than or equal to the indicated value. Cumulative percent is useful only for variables that are ordinal or scale.

TIP

Depending on your research question, the Percent column or the Valid Percent column may be useful when you have a lot of missing data or a variable was not applicable to a large percentage of people.

Bar charts, like the one in Figure 14-6, summarize the distribution that was observed in the frequency table and allow you to see the distribution. For the Payment_Method variable, more than half of the people are in the Credit Card category.

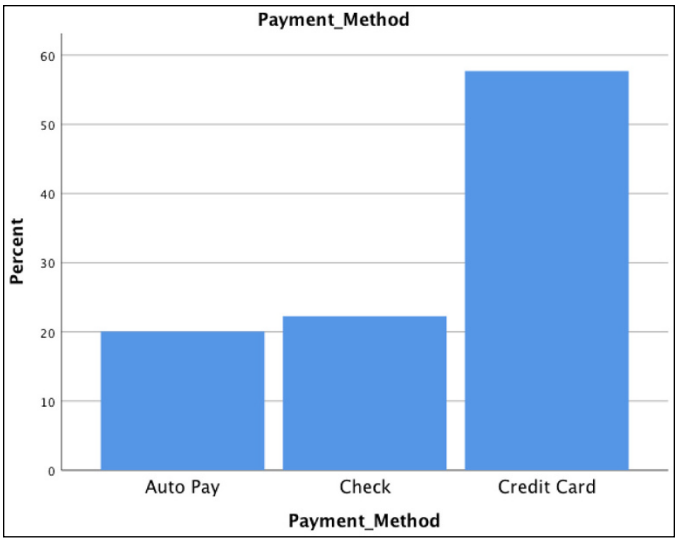

FIGURE 14-6: A bar chart for the Payment_Method variable.

Understanding Frequencies for Continuous Variables

As you have seen, frequency tables show counts and percentages, which are extremely useful when working with categorical variables. However, for continuous variables, which can have many values, frequency tables become less useful. For example, if you were working with a variable such as income, it wouldn't be very useful to know that only one person in the dataset made $22,222 last year. It's likely that each response would have a different value, so the frequency table would be very, very large and not useful as a summary of the variable.

Instead, if the variables of interest are continuous, the frequencies procedure can be useful because of the summary statistics it can produce. To run frequencies for continuous variables, follow these steps:

1. **From the main menu, choose File ➪ Open ➪ Data and load the Merchandise.sav data file.**

   Download the file from this book's companion website, at `www.dummies.com/go/spss4e`.

2. **Choose Analyze ➪ Descriptive Statistics ➪ Frequencies.**

3. **Select the Stereos, TVs, Speakers, Delivery_Problems, Years_as_customer, and Estimated_Revenue variables, and place them in the Variable(s) box.**

4. **Deselect the Display Frequency Tables check box, as shown in Figure 14-7.**

   A warning dialog appears saying, "You have turned off all output. Unless you select any Output Options this procedure will not be run." You receive this warning because at the moment nothing is selected. This is okay because you will now select the summary statistics you want to display.

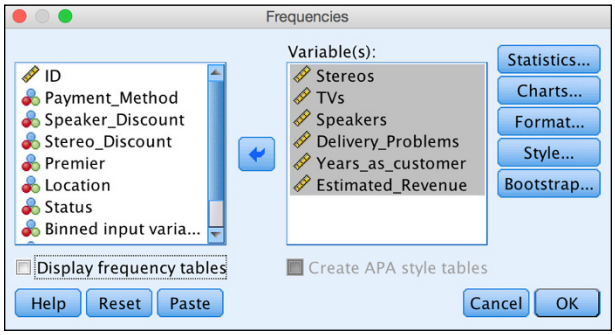

FIGURE 14-7: The Frequencies dialog.

5. **Click the Statistics button.**

   The Frequencies: Statistics dialog appears.

   All these summary statistics are appropriate for scale variables. The statistics can be divided into those summarizing the central tendency, those measuring the amount of variation (dispersion) in the data, different percentile values you can request, and statistics assessing the shape of the distribution.

6. **In the Central Tendency section, select the Mean, Median, and Mode check boxes. In the Dispersion section, select the Std. Deviation, Minimum, and Maximum check boxes.**

   These selections are shown in Figure 14-8.

7. **Click Continue.**

8. **Click the Charts button.**

   The Frequencies: Charts dialog appears.

9. **Select the Histograms radio button and select the Show Normal Curve on Histogram check box, as shown in Figure 14-9.**

10. **Click Continue, and then click OK.**

    SPSS runs the frequencies procedure and calculates the summary statistics and the histogram you requested.

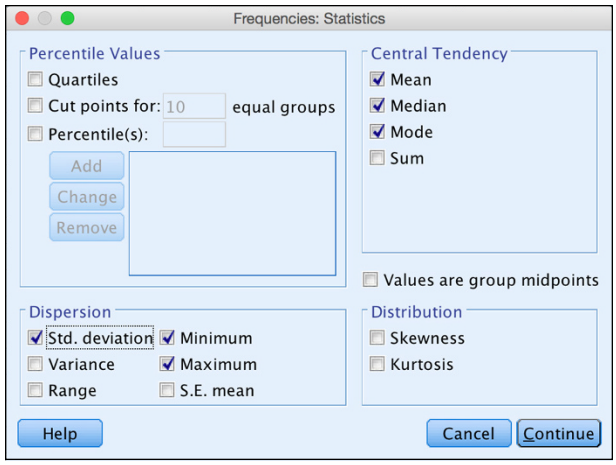

FIGURE 14-8: The Frequencies: Statistics dialog.

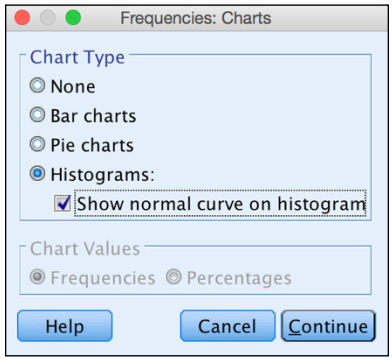

FIGURE 14-9: The Frequencies: Charts dialog.

The statistics table in Figure 14-10 shows that you have 3,338 valid cases and don't have any missing data. The statistics table contains the requested statistics. For example, for the Speakers variable, the minimum value is 0 and the maximum value is 451. This range of values seems to be very large, so it would be useful to double-check the data to make sure there are no errors.

Statistics

		Stereos	TVs	Speakers	Delivery_Problems	Years_as_customer	Estimated_Revenue
N	Valid	3338	3338	3338	3338	3338	3338
	Missing	0	0	0	0	0	0
Mean		13.71	.83	51.30	.13	6.38	5034510.94
Median		14.00	.00	36.00	.00	6.00	5029070.00
Mode		0	0	4	0	7	5029070
Std. Deviation		9.417	2.228	54.104	.434	2.565	2828800.41
Minimum		0	0	0	0	2	11028
Maximum		30	10	451	4	11	9983290

FIGURE 14-10: The Statistics table for six variables.

Likewise, in an ideal world, you would like the mean, median, and mode to be similar, because they're all measures of central tendency. In this case, note that for the Speakers variable, the mean (51.3), median (36), and mode (4) are very different from each other, which is an indication that this variable is probably not normally distributed. (You see why this is important in later chapters.)

You can visually check the distribution of these variables with a histogram, as shown in Figure 14-11. A histogram has bars, but unlike a bar chart, the bars are plotted along an equal interval scale. The height of each bar is the count of values falling within the interval. Note that the lower range of values is truncated at 0 and the number of speakers is greatest down toward the lower end of the distribution, although there are some extreme values. The distribution is not normal.

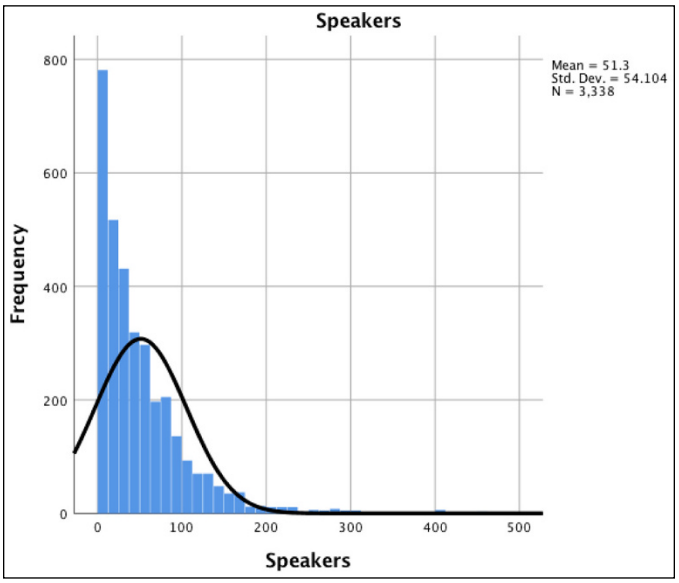

FIGURE 14-11: A histogram for the Speakers variable.

Summarizing Continuous Variables with the Descriptives Procedure

The *descriptive procedure* is an alternative to the frequencies procedure (see the preceding section) when the objective is to summarize continuous variables. The descriptives procedure provides a succinct summary of various statistics and the number of cases with valid values for each variable included in the table.

To use the descriptives procedure, follow these steps:

1. **From the main menu, choose File ➪ Open ➪ Data and load the Merchandise.sav data file.**

   Download the file at `www.dummies.com/go/spss4e`.

2. **Choose Analyze ➪ Descriptive Statistics ➪ Descriptives.**

   The Descriptives dialog appears.

3. **Select the Stereos, TVs, Speakers, Delivery_Problems, Years_as_customer, and Estimated_Revenue variables, and place them in the Variable(s) box, as shown in Figure 14-12.**

4. **Click OK.**

   SPSS runs the descriptives procedure and calculates the summary statistics.

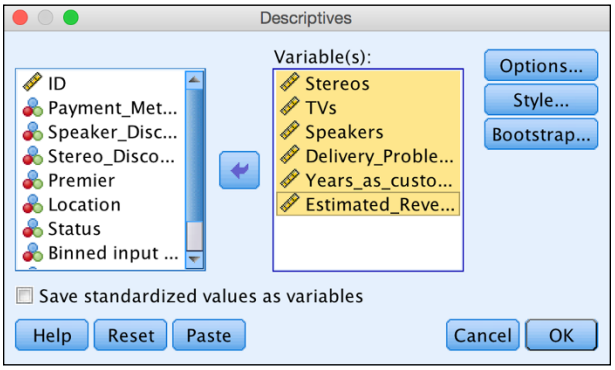

FIGURE 14-12: Place the variables in the Variable(s) box.

The minimum and maximum values provide an efficient way to check for values outside the expected range (see Figure 14-13). If you see a value that is too low or too high, you might have data errors or potential outliers. Likewise, it's always important to investigate the mean and determine whether the value makes sense. Ask yourself, "Is this what I was expecting?" Sometimes a mean may be lower or higher than expected, which can indicate a problem related to how the data was coded or collected.

It is important also to check the standard deviation. For example, if you have a standard deviation of zero, every person in the dataset provided the same value. This information might be useful from a business perspective if everyone loved your product. From a statistical perspective, however, a variable that doesn't vary isn't useful.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Stereos	3338	0	30	13.71	9.417
TVs	3338	0	10	.83	2.228
Speakers	3338	0	451	51.30	54.104
Delivery_Problems	3338	0	4	.13	.434
Years_as_customer	3338	2	11	6.38	2.565
Estimated_Revenue	3338	11028	9983290	5034510.94	2828800.41
Valid N (listwise)	3338				

FIGURE 14-13: The Descriptive Statistics table.

Finally, the last row in the table, Valid N (listwise), is the number of cases that have a valid value for all the variables appearing in the table. In this example, you have no missing data, so this number is at least partially useful because it shows that all the data entries are in a sense complete. Valid N (listwise) is useful for a set of variables that you intended to use for a *multivariate analysis* (an analysis looking at the relationships between many variables).