Chapter 4

Entering and Defining Data

In This Chapter

- Considering your choices when defining a variable
- Defining variables
- Entering numbers
- Making sure that you're using the right measurement type

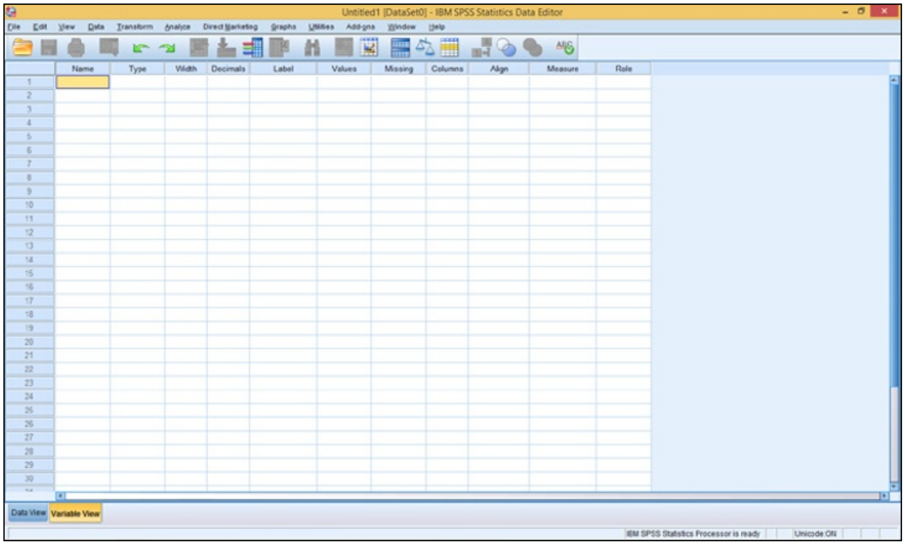
To process your data, you have to get it into the computer. Entering data has been a problem with computers since the beginning. No matter how you decide to get your numbers into SPSS, at some point someone has to type them (unless they come from some form of automatic monitoring). These days, it feels like we spend half of our time entering data into online forms, which saves some analyst from typing on the other end. SPSS can read data from other places. You can also type directly into SPSS — and, if you want, copy the data to places other than SPSS later.

Entering data into SPSS is a two-step process: First, you define what sort of data you'll be entering. Then you enter the actual numbers. This may sound difficult, but it isn't so bad. When you see how data entry works in SPSS, you'll discover you have some pretty nifty software to help you.

You organize your data into cases. Each case is made up of a collection of variables. First, you define the characteristics of the variables that make up a case, and then you enter the data into the variables to make up the contents of the cases. This chapter shows you how to work with this technique of getting data into your system.

Entering Variable Definitions on the Variable View Tab

You use the Variable View tab of the Data Editor window, shown in Figure 4-1, to define the names and characteristics of variables. This is where you always start if you plan on entering data into SPSS. As you can see in Figure 4-1, every


Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role

Figure 4-1: You use the Variable View tab to define the characteristics of variables.

characteristic you can define about your variables is named at the top of the window. All you have to do is enter something in each column for each variable.

The predefined set of 11 characteristics are the only ones needed to completely specify all the attributes of any variable. The characteristics are all known to the internal SPSS processing. When you add a new variable, you'll find that reasonable defaults appear for most characteristics.

The Variable View tab is just for defining the variables. The entry of the actual numbers comes later (see "Entering and Viewing Data Items on the Data View Tab," later in this chapter).

Each variable characteristic has a default, so if you don't specify a characteristic, SPSS fills one in for you. However, what it selects may not be what you want, so let's look at all the possibilities.

Name

The cell on the far left is where you enter the name of the variable. Just click the cell and type a short descriptor, such as **age**, **income**, **sex**, or **odor**. (A longer descriptor, called a *label,* comes later.) You can type longer names here, but you should keep them short because they'll be used in named lists and as identifier tags on the data graphs and such — where the format can be a bit crowded. Names that are too long can cause the output from SPSS to be garbled or truncated.

If the name you assigned turns out to be too long or is misspelled, you can always change it on the Variable View tab. One of the nice things about SPSS is that you can correct mistakes quickly.

Here are some handy hints about names:

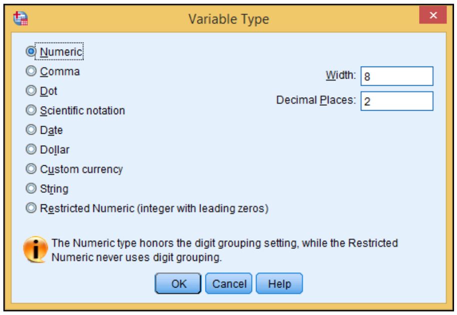
- **You can use some bizarre characters in a name, such as @, #, and $, as well as the underscore character (_) and numbers.** But if you use screwy characters in a name, you may live to regret it. For one thing, you can't start a variable name with these characters. Plus, they'll remind expert users of special variables in some advanced features. An underscore in the middle of a name is a great way to make a name more readable, but otherwise, it's best to keep your names simple.
- **Be sure to start every name with an uppercase or lowercase letter.**
- **You can't include blanks anywhere in a name, but an underscore is a good substitute.**

If you want to export data to another application, make sure the names you use are in a form acceptable to that application. Watch out for special characters.

Type

Most data you enter will be just regular numbers. Some, however, will be a special type, such as currency, and some will be displayed in a special format. Other data, such as dates, will require special procedures for calculation. You simply specify what type you have, and SPSS takes care of those other details for you. This is a comprehensive look at all the types. (We give you more advice about some special types in Chapter 7.)

Click the cell in the Type column you want to fill in, and a button with three dots appears on its right. Click that button, and the Variable Type dialog box, shown in Figure 4-2, appears.



Variable Type

● Numeric
○ Comma
○ Dot
○ Scientific notation
○ Date
○ Dollar
○ Custom currency
○ String
○ Restricted Numeric (integer with leading zeros)

Width: 8

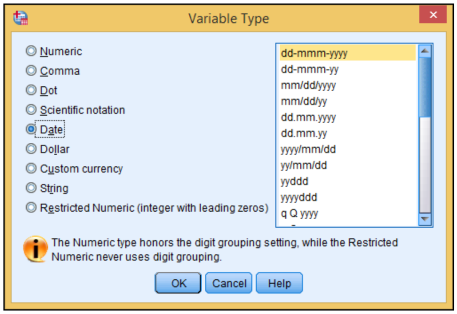
Decimal Places: 2

The Numeric type honors the digit grouping setting, while the Restricted Numeric never uses digit grouping.

Figure 4-2: The Variable Type dialog box allows you to specify the type of variable you're defining.

You can choose from the following predefined types of variables:

- **Numeric:** Standard numbers in any recognizable form. The values are entered and displayed in the standard form, with or without decimal points. Values can be formatted in standard scientific notation, with an embedded *E* to represent the start of the exponent. The Width value is the total number of all characters in a number — including any positive or negative signs and the exponent indicator. The Decimal Places value specifies the number of digits displayed to the right of the decimal point, not including the exponent.
- **Comma:** This type specifies numeric values with commas inserted between three-digit groups. The format includes a period as a decimal point. The Width value is the total width of the number, including all commas and the decimal point. The Decimal Places value specifies the number of digits to the right of the decimal point. You may enter data without the commas, but SPSS will insert them when it displays the value. Commas are never placed to the right of the decimal point.
- **Dot:** Same as Comma, except a period is used to group the digits into threes, and a comma is used for the decimal point.
- **Scientific Notation:** A numeric variable that always includes the *E* to designate the power-of-ten exponent. The *base* (the part of the number to the left of the *E*) may or may not contain a decimal point. The *exponent* (the part of the number to the right of the *E,* which also may or may not contain a decimal) indicates how many times 10 multiplies itself, after which it's multiplied by the base to produce the actual number. You may enter *D* or *E* to mark the exponent, but SPSS always displays the number using *E.* For example, the number 5,286 can be written as 5.286E3. To represent a small number, the exponent can be negative. For example, the number 0.0005 can be written as 5E–4. This format is useful for very large or very small numbers.
- **Date:** A variable that can include the year, month, day, hour, minute, and second. When you select Date, the available format choices appear in a list on the right side of the dialog box, as shown in Figure 4-3. Choose the format that best fits your data. Your selection determines how SPSS will format the contents of the variable for display. This format also determines, to some extent, the form in which you enter the data. You can enter the data using slashes, colons, spaces, or other characters. The rules are loose — if SPSS doesn't understand what you enter, it tells you, and you can re-enter it another way. For example, if you select a format with a two-digit year, SPSS accepts and displays the year that way, but it will use four digits to perform calculations. The first two digits (the number of the century) will be selected according to the configuration you set by choosing Edit➪Options and then clicking the Data tab.


Variable Type

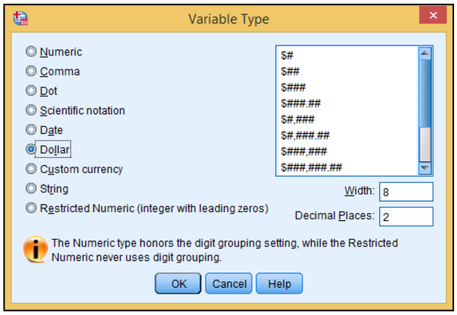
○ Numeric
○ Comma
○ Dot
○ Scientific notation
● Date
○ Dollar
○ Custom currency
○ String
○ Restricted Numeric (integer with leading zeros)

dd-mmm-yyyy
dd-mmm-yy
mm/dd/yyyy
mm/dd/yy
dd.mm.yyyy
dd.mm.yy
yyyy/mm/dd
yy/mm/dd
yyddd
yyyyddd
q Q yyyy

The Numeric type honors the digit grouping setting, while the Restricted Numeric never uses digit grouping.

Figure 4-3: Selecting a date format also selects which items are included.

- **Dollar:** When you select Dollar, the available format choices appear in a list on the right side of the dialog box (see Figure 4-4). Dollar values are always displayed with a leading dollar sign and a period for a decimal point; for large values, they include commas to collect the digits in groups of threes. You select the format and its Width and Decimal Places values. The format choices are similar, but it's important that you choose one that's compatible with your other dollar-variable definitions so they line up when you print and display monetary values in output tables. The Width and Decimal Places settings help with vertical alignment in the output, no matter how many digits you include in the format itself. No matter what format you choose, you can enter the values without the dollar sign and the commas; SPSS inserts those for you.


Variable Type

○ Numeric
○ Comma
○ Dot
○ Scientific notation
○ Date
● Dollar
○ Custom currency
○ String
○ Restricted Numeric (integer with leading zeros)

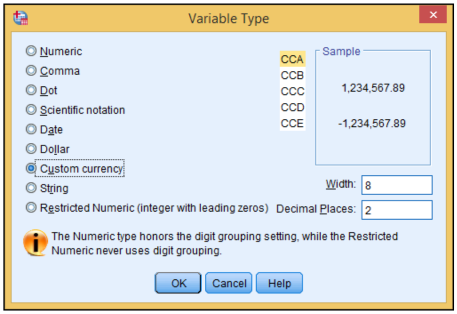
$#
$##
$###
$###.##
$#,###
$#,###.##
$###,###
$###,###.##

Width: 8
Decimal Places: 2

The Numeric type honors the digit grouping setting, while the Restricted Numeric never uses digit grouping.

Figure 4-4: The different dollar formats mostly specify the number of digits to be included.

- **Custom Currency:** The five custom formats for currency are named CCA, CCB, CCC, CCD, and CCE, as shown in Figure 4-5. You can view and modify the details of these formats by choosing Edit➪Options and then clicking the Currency tab. Fortunately, you can modify the definitions of these custom formats as often as you like without fear of damaging your data. As with the Dollar format, the Width and Decimal Places settings are primarily for aligning the data when you're printing a report.



Variable Type

○ Numeric
○ Comma
○ Dot
○ Scientific notation
○ Date
○ Dollar
● Custom currency
○ String
○ Restricted Numeric (integer with leading zeros)

CCA
CCB
CCC
CCD
CCE

Sample
1,234,567.89
-1,234,567.89

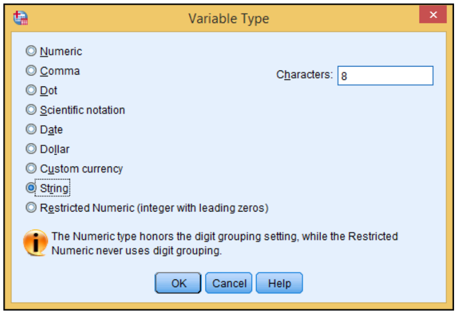
Width: 8
Decimal Places: 2

The Numeric type honors the digit grouping setting, while the Restricted Numeric never uses digit grouping.

Figure 4-5: Five custom currency formats are available.

- **String:** A freeform non-numeric item (see Figure 4-6). The only good time to use string is when it truly is a string, like an address, a proper name, or a product code (SKU). Avoid using the String type when it really should be labeled Numeric. Something like favorite color, sex, or state should *not* be a string because it has a finite list of possibilities that are known in advance. (See the "Values" section later in this chapter.)

  SPSS allows a very large number for the size of the string — so large that you could fit a paragraph, which is exactly what you would do if you were doing text mining. Open-ended response items in a survey would also be an example of a string.



Variable Type

○ Numeric
○ Comma
○ Dot
○ Scientific notation
○ Date
○ Dollar
○ Custom currency
● String
○ Restricted Numeric (integer with leading zeros)

Characters: 8

The Numeric type honors the digit grouping setting, while the Restricted Numeric never uses digit grouping.

Figure 4-6: Strings are text like addresses, names, and open-ended responses.

- **Restricted Numeric:** A relatively new choice, so you may not see it mentioned in older books about SPSS. This is perfect for numbers that sometimes have leading zeros like zip codes and Social Security numbers. They aren't really numbers because you don't perform arithmetic on them. Back in the day, these types of numbers had to be declared as strings.

Width

The width setting in the definition of a variable determines the number of characters used to display the value. If the value to be displayed is not large enough to fill the space, the output will be padded with blanks. If it's larger than you specify, it will either be reformatted to fit or asterisks will be displayed.

TIP

Certain type definitions allow you to set a width value. The width value you enter as the width definition is the same as the one you enter when you define the type. If you make a change to the value in one place, SPSS changes the value in the other place automatically. The two values are the same.

At this point, you can do one of three things:

- Skip this cell and accept the default (or the number you entered previously under Type).
- Enter a number and move on.
- Use the up and down arrows that appear in the cell to select a numeric value.

Decimals

The number of decimals is the number of digits that appear to the right of the decimal point when the value appears onscreen. This is the same number that you may have specified as the Decimal Places value when you defined the variable type. If you entered a number there, it appears here as the default. If you enter a number here, it changes the one you entered for the type. They're the same.

Now you can do one of three things:

- Skip this cell and accept the default (or the number you entered earlier under Type).
- Enter a number and move on.
- Use the up and down arrows that appear in the cell to select a numeric value.

Label

The name and the label serve the same basic purpose: They're descriptors that identify the variable. The difference is that the *name* is the short identifier and the *label* is the long one. You need one of each because some output formats work fine with a long identifier and other formats need the short form.

You can use just about anything for the label. What you choose has to do with how you expect to use your data and what you want your output to look like. For example, the variable name may be "Sex" and the longer label may be "Boys and Girls," "Men and Women," or simply "Gender."

The length of the label isn't determined by some sort of software requirement. However, output looks better if you use short names and somewhat longer labels. Each one should make sense standing alone. After you produce some output, you may find that your label is lousy for your purposes. That's okay; it's easy to change. Just pop back to the Variable View tab and make the change. The next time you produce output, the new label will be used.

You can also just skip defining a label. If you don't have a label defined for a variable, SPSS will use the name you defined for everything.

Values

The Values column is where you assign labels to all the possible values of a variable. If you select a cell in the Values column, a button with three dots appears. Clicking that button displays the dialog box shown in Figure 4-7.

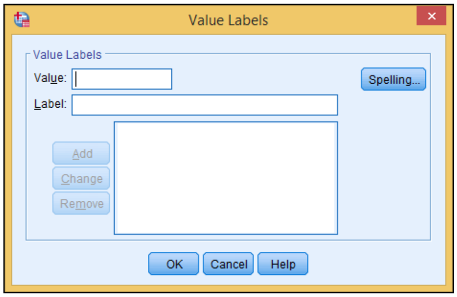


Figure 4-7: You can assign a name to each possible value of a variable.

Normally, you make one entry for each possible value that a variable can assume. For example, for a variable named `Sex` you could have the value `1` assigned the label "Male" and `2` assigned the label "Female." Or, for a variable named `Committed` you could have `0` for "No," `1` for "Yes," and `2` for "Undecided." If you have labels defined, when SPSS displays output, it can show the labels instead of the values.

To define a label for a value:

1. **In the Value box, enter the value.**
2. **In the Label box, enter a label.**
3. **Click the Add button.**

   The value and label appear in the large text block.
4. **To change or remove a definition, simply select it in the text block and make your changes; then click the Change button.**
5. **Repeat Steps 1–4 as needed.**
6. **Click OK to save the value labels and close the dialog box.**

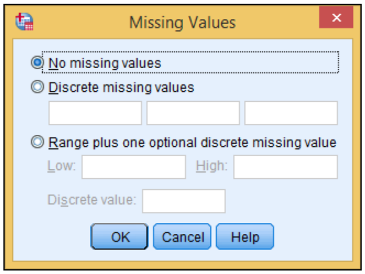
You can always come back and change the definitions using the same process you used to enter them. The dialog box will reappear, filled in with all the definitions; then you can update the list.

Sometimes you have a whole bunch of strings and you really don't want to make them all values because it seems like it'll be a lot of work. A variable like college major is a good example. If you dread setting up `1` as "Astrophysics," `2` as "Biology," `3` as "Chemistry," and so on, you can use a special dialog box called Automatic Recode (under the Transform menu) and it'll do all the work for you.

Missing

You can specify what is to be entered for a value that is missing for a variable in a case. In other words, when you have values for all variables in a case except one, you can specify a placeholder for the missing value. Select a cell in the Missing column. Click the button with three dots and the Missing Values dialog box, shown in Figure 4-8, appears.

For example, say you're entering responses to questions, and one of the questions is, "How many cars do you own?" The normal answer to this question is a number, so you define the variable type as a number. If someone chooses to ignore this question, this variable won't have a value. However,



Missing Values

● No missing values

○ Discrete missing values

○ Range plus one optional discrete missing value

Low: High:

Discrete value:

Figure 4-8: You can specify exactly what is entered for a missing value.

you can specify a placeholder value. Perhaps `0` seems like a good choice for a placeholder here, but it's not really — lots of people don't have cars. Instead, a less likely value — like, say, `–1` — makes a better choice. A very popular choice among SPSS users is `–9`, but this will depend on the values of the original variable.

You can even specify unique values to represent different reasons for a value being missing. In the previous example, you could define `–1` as the value entered when the answer is, "I don't remember," and `–2` could be used when the answer is, "None of your business." If you specify that a value is representing a missing value, that value is not included in general calculations. During your analysis, however, you can determine how many values are missing for each of the different reasons. You can specify up to three specific values (called *discrete values*) to represent missing data, or you can specify a range of numbers along with one discrete value, all to be considered missing. The only reason you would need to specify a range of values is if you have lots of reasons why data is missing and want to track them all.

One of the many reasons you don't want to abuse the string type is that it makes a mess of missing data or incorrect data. If Female and Male are strings, you can get entries like "m," "M," "Male," and even crazy unexpected ones like "H" and "mail." You're better off doing what all experienced users do: Use numeric codes with values!

Columns

The Columns column is where you specify the width of the column you'll use to enter the data. The folks at SPSS could have used the word *Width* to describe it, but they already used that term for the width of the data itself. A better name may have been the two words *Column Width,* but that would have been too long to display nicely in this window, so they just called it *Columns.* To specify the number of columns, select a cell and enter the number.

Align

The Align column determines the position of the data in its allocated space, whenever the data is displayed for input or output. The data can be left-aligned, right-aligned, or centered. You've defined the width of the data and the size of the column in which the data will be displayed; the alignment determines what is done with any space left over.

When you select a cell in the Align column, a list appears and you can choose one of the three alignment possibilities, as shown in Figure 4-9. Aligning to the left means inserting all blanks on the right; aligning to the right inserts all the extra spaces on the left; centering the data splits the spaces evenly on each side — we don't know what it does if an odd space is left over. (We also worry about things like the number of seeds in a tomato and where the clouds go at night.)

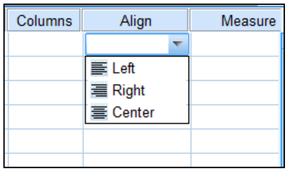


Figure 4-9: Values can be justified right or left, or positioned in the center.

Measure

Your value here specifies the measure of something in one of three ways. When you click a cell in the Measure column, you can select one of these choices (see Figure 4-10):

- **Ordinal:** These numbers specify the position (order) of something in a list. For example, *first, second,* and *third* are ordinal numbers.
- **Nominal:** Numbers that specify categories or types of things. You can have `0` represent "Disapprove" and 1 represent "Approve." Or you can use `1` to mean "Fast" and `2` to mean "Slow."
- **Scale:** A number that specifies a magnitude. It can be distance, weight, age, or a count of something.

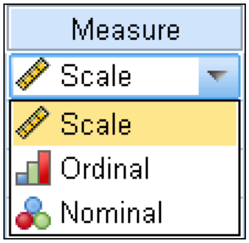


Figure 4-10: The type of measurement being made by the values in this variable.

Role

Some of the SPSS dialog boxes select variables according to their role and include them as defaults. You don't need to worry about this characteristic. It can be handy when you have some experience with SPSS and understand how defaults are chosen.

Be on the lookout for Analyze menus that let you use predefined roles. They use this feature. There aren't that many of them, but the number grows with each new version of SPSS.

These predefined roles allow greater capability with SPSS Modeler, which is a kind of sibling product to SPSS Statistics.

When you click a cell in the Role column, you can select one of six choices (see Figure 4-11):

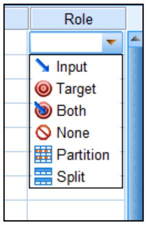


Figure 4-11: The role assumed by this variable in certain SPSS dialog boxes.

- **Input:** This variable is used for input. This is the default role. Definition of roles was introduced to Version 18 of SPSS, and all data imported from earlier versions will be assigned this role.
- **Target:** This variable is used as output by SPSS procedures.
- **Both:** This variable is used as both input and output.

- **None:** This variable has no role assignment.
- **Partition:** This variable is used to partition the data into separate samples for training, testing, and validation.
- **Split:** This variable is used to build separate models for each possible value of the variable. This capability should not be confused with file splitting (see Chapter 8).

Entering and Viewing Data Items on the Data View Tab

After you've defined all the variables for each case, click the Data View tab of the Data Editor window so you can begin typing the data. At the top of the columns in Figure 4-12, you can see some names we chose for variables. Switching to the Data View tab makes the window ready to receive entered data — and to verify that what's entered matches the specified format and type of the data.

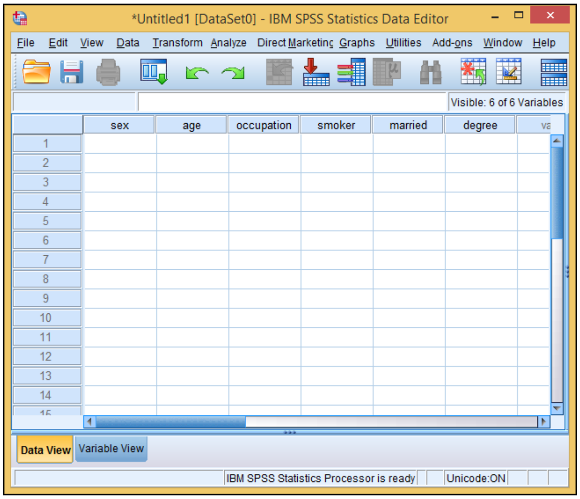

*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 6 of 6 Variables
sex
age
occupation
smoker
married
degree
va
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

Figure 4-12: The Data View tab, ready to accept new data.

Entering data into one of these cells is straightforward: You simply click the cell and start typing.

If something is already in a cell and you want to change it instead of just typing over it, look up toward the top of the window, just underneath the toolbar: You'll see the name of the variable and the currently selected value. Click the value in the field at the top, and you can edit it right there. You can do all the normal mouse and keyboard stuff there, too — you can use the Backspace key to erase characters, or select the entire value and type right over it.

TIP

If you feel like a lousy (or inexperienced) mouse driver, take some time to experiment and figure out how to edit data. Lots of software use these same editing techniques, so becoming proficient now will pay you dividends later.

If your data is already in a file, you may be able to avoid typing it in again by reading that file directly into SPSS. For more information, see Chapter 5.

WARNING!

Don't take chances. As soon as you type a few values, save your data to a file by choosing File➪Save As. Then choose File➪Save throughout the process of entering data, and you won't be ruined if the computer crashes unexpectedly.

We all have to go back and refine our variable definitions from time to time. That's normal. When you come across something that doesn't do what you want it to, just switch back to the Variable View tab and correct it. Nobody but you and SPSS will ever know about it, and SPSS never talks.

Filling In Missed Categorical Values

Now that you've defined your variables and entered your data, you may want to check that you have names defined for all your actual ordinal and nominal values, and that you have defined the correct measures for them. SPSS can help by scanning your data, finding values for which you don't have definitions, and pointing them out in a friendly way.

The following steps use an existing file to walk through a demonstration:

1. **Choose File➪Open➪Data to load the file named `car_sales.sav`.**

   This file came with your installation of SPSS and is found, along with a number of other files, in the same directory in which you installed SPSS. You can load any of these data files, but `car_sales.sav` is the one used in this demonstration. If you load this file while you already have some other data showing in the window, SPSS will open a new Data Editor window to display the new information; your existing data will not be lost.

   When you open this data file — or any data file, for that matter — SPSS opens the SPSS Statistics Viewer window to tell you that it has opened a file (or the information could be displayed in the SPSS Statistics Viewer window that's already open). You won't need this information for what you're doing here, so you can just close the window.

2. **Choose Data➪Define Variable Properties.**

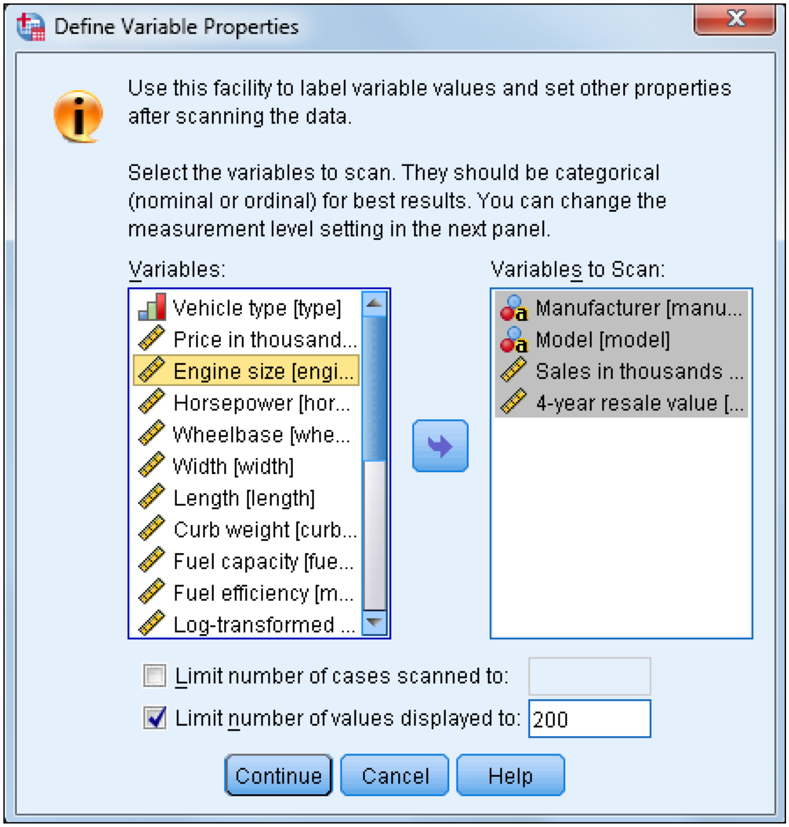
   The Define Variable Properties dialog box appears.

3. **On the left, select all the names of the variables you want to check, and then click the arrow in the center of the dialog box to move them to the right, as shown in Figure 4-13.**

4. **When you're done, click Continue.**

5. **Select one of the variable names in the list on the left.**

   Its different values appear in the center of the dialog box, as shown in Figure 4-14. (In this example, every value has a name assigned to it.)



Define Variable Properties

Use this facility to label variable values and set other properties after scanning the data.

Select the variables to scan. They should be categorical (nominal or ordinal) for best results. You can change the measurement level setting in the next panel.

Variables:

- Vehicle type [type]
- Price in thousand...
- Engine size [engi...
- Horsepower [hor...
- Wheelbase [whe...
- Width [width]
- Length [length]
- Curb weight [curb...
- Fuel capacity [fue...
- Fuel efficiency [m...
- Log-transformed ...

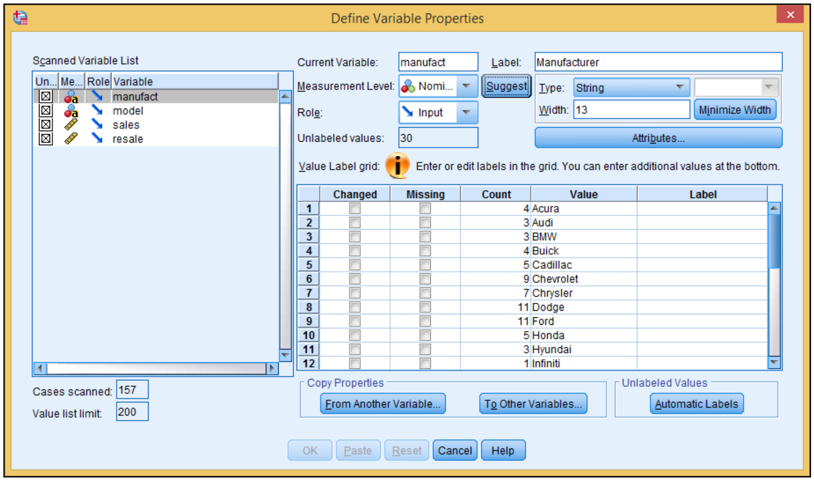
Variables to Scan:

- Manufacturer [manu...
- Model [model]
- Sales in thousands ...
- 4-year resale value [...

☐ Limit number of cases scanned to:

☑ Limit number of values displayed to: 200

Figure 4-13: Selecting variables to check their properties.



Define Variable Properties

Scanned Variable List

Un...	Me...	Role	Variable
☒			manufact
☒			model
☒			sales
☒			resale

Current Variable: manufact

Label: Manufacturer

Measurement Level: Nomi...

Type: String

Width: 13

Role: Input

Unlabeled values: 30

Value Label grid: Enter or edit labels in the grid. You can enter additional values at the bottom.

	Changed	Missing	Count	Value	Label
1	□	□	4	Acura	
2	□	□	3	Audi	
3	□	□	3	BMW	
4	□	□	4	Buick	
5	□	□	5	Cadillac	
6	□	□	9	Chevrolet	
7	□	□	7	Chrysler	
8	□	□	11	Dodge	
9	□	□	11	Ford	
10	□	□	5	Honda	
11	□	□	3	Hyundai	
12	□	□	1	Infiniti	

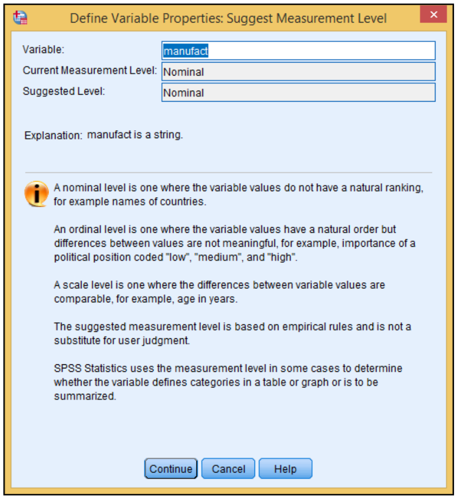
Cases scanned: 157

Value list limit: 200

Figure 4-14: The values of the selected variable.

6. **Ask SPSS to suggest a new type for this variable by clicking the Suggest button in the top center of the dialog box.**

   The dialog box in Figure 4-15 appears, telling you what SPSS concludes about this variable and its values. This same window, with different text, appears for each variable you test. Sometimes the text suggests changes in the variable definition, and sometimes it doesn't.



Define Variable Properties: Suggest Measurement Level

Variable: manufact

Current Measurement Level: Nominal

Suggested Level: Nominal

Explanation: manufact is a string.

A nominal level is one where the variable values do not have a natural ranking, for example names of countries.

An ordinal level is one where the variable values have a natural order but differences between values are not meaningful, for example, importance of a political position coded "low", "medium", and "high".

A scale level is one where the differences between variable values are comparable, for example, age in years.

The suggested measurement level is based on empirical rules and is not a substitute for user judgment.

SPSS Statistics uses the measurement level in some cases to determine whether the variable defines categories in a table or graph or is to be summarized.

Figure 4-15: From the pattern of values, SPSS concludes whether you may have chosen the wrong measurement.

7. **To apply any changes, click Continue.**

   You return to the window shown in Figure 4-14, where you can select another variable.

You won't want to make changes to all your variables, but SPSS helps you find the ones that you do need to change. Values defined as "missing" are not included in the computations. The text in the window always explains the criteria used to reach a conclusion, and SPSS allows you to make the final decision.

Chapter 5

Opening Data Files

In This Chapter

- Understanding the SPSS file format
- Reading a simple text file into SPSS
- Transferring data from another application into SPSS
- Saving SPSS data in various formats

You don't need to put your data into the computer more than once. If you've entered your data in another program, you can copy it from there into SPSS — because every program worth using has some form of output that can serve as input to SPSS. This chapter discusses ways to transfer data into and out of SPSS.

Getting Acquainted with the SPSS File Format

SPSS has its own format for storing data and writes files with the `.sav` extension. This file format contains special codes and usually can't be used to export your data to another application. It's used only for saving SPSS data that you want to read back into SPSS at a later time. Several example files in this format are copied to your computer as part of the normal SPSS installation. These files can be found in the same directory as your SPSS installation. You can load any one of them by choosing File➪Open➪Data and selecting the file to be loaded. When you do so, the variable names and data are loaded and fill your SPSS window.

If you have SPSS filled with data, you can save it to a `.sav` file by choosing File➪Save As and providing a name for the file. Or if you've loaded the information from a file, or you've previously saved a copy of the information to a file, you can simply choose File➪Save to overwrite the previous file with a fresh copy of both variable definitions and data.

It’s easy to be fooled by the way the SPSS documentation uses the word *file*. If you have defined data and variables in your program, the SPSS documentation often refers to it all as a “file,” even though it may have never been written to disk. SPSS also refers to the material written to disk as a file, so watch the context.

When you write your file to disk, if you don’t add the `.sav` extension to the filename, SPSS adds it for you. When you choose File➪Open➪Data to display the list of files, you may or may not see the extension on the filename (it depends on how your Windows system is configured), but it’s there.

Formatting a Text File for Input into SPSS

If your data is in an application that can’t directly create a file of a type that SPSS can read, getting the data into SPSS may be easier than you think. If you can get the information out of your application and into a text file, it’s fairly easy to have SPSS read the text file.

When it comes to writing information to disk, some applications are more obliging than others. Look for an Export menu option — it usually has some options that allow you to organize the output text in a form you want. (Read on for a description of possible organization schemes.)

If the application doesn’t allow you to format text the way you want, look for printer options — maybe you can redirect printer output to a disk file and work from there. If you use the application’s printer output, you may need to use your word processor to clean up the form of the data. We know this multistep operation sounds like a lot of work, but it’s often easier than typing all your data in again by hand.

The data file you output from SPSS doesn’t have to include the variable names, just the values that go into the variables. You can format the data in the file by using spaces, tabs, commas, or semicolons to separate data items. Such dividers are known as *delimiters*. Another method of formatting data avoids delimiters altogether. In that method, you don’t have to separate the individual data items, but you must make each data item a specific length, because you have to tell SPSS exactly how long each one is.

The most intuitive format is to have one case (one row of data) per line of text. That means the data items in your text file are in the same positions they’ll be in when they’re read into SPSS. Alternatively, you can have all your data formatted as one long stream, but you'll have to tell SPSS how many items go into each case.

REMEMBER

Always save this kind of raw data as simple text; the file you store it in should have the .txt extension so SPSS can recognize it for what it is.

Reading Simple Data from a Text File

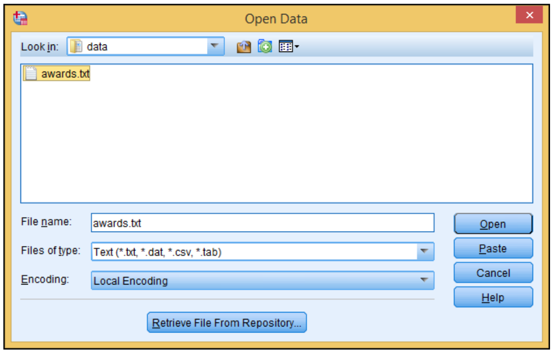
This section contains an example of a procedure you can follow to read data from a simple text file into SPSS. The file is named awards.txt. It contains two cases (rows of data) as two lines of text, with the data items in the two lines separated by spaces. The content of the file is as follows:

```
"Pat" 1 35 3.00 9
"Chris" 1 22 2.4 7
```

The following example reads this text file and inserts it into the cells of SPSS. Along the way, SPSS keeps you informed about what's going on so there won't be any big surprises at the end.

1. **Choose File➪Read Text Data.**

   The Open Data window, shown in Figure 5-1, appears.



Look in: data

File name: awards.txt

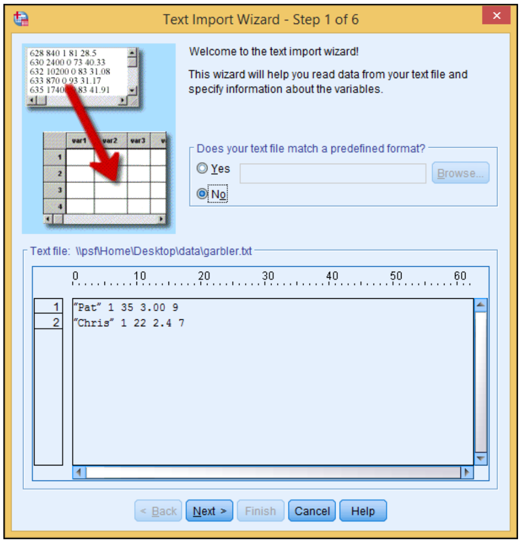
Files of type: Text (*.txt, *.dat, *.csv, *.tab)

Encoding: Local Encoding

Figure 5-1: Locate the file you want to read.

2. **Select the awards.txt file, and then click Open.**

   The Text Import Wizard (the first screen of which is shown in Figure 5-2) appears, allowing you to load and format your data.



Text Import Wizard - Step 1 of 6

Welcome to the text import wizard!

This wizard will help you read data from your text file and specify information about the variables.

Does your text file match a predefined format?

○ Yes

● No

Text file: \\psf\Home\Desktop\data\garbler.txt

```
1  "Pat" 1 35 3.00 9
2  "Chris" 1 22 2.4 7
```

Figure 5-2 Make sure your data looks reasonable.

3. **Examine the input data.**

   The screen lets you peek at the contents of the input file so you can verify that you've chosen the right file. Also, if your file uses a predefined format (which it doesn't, in this example), you can select it here and skip some of the later steps. If your data doesn't show up nicely separated into values the way you want, you may be able to correct it in a later step. Don't panic just yet.

4. **Click Next.**

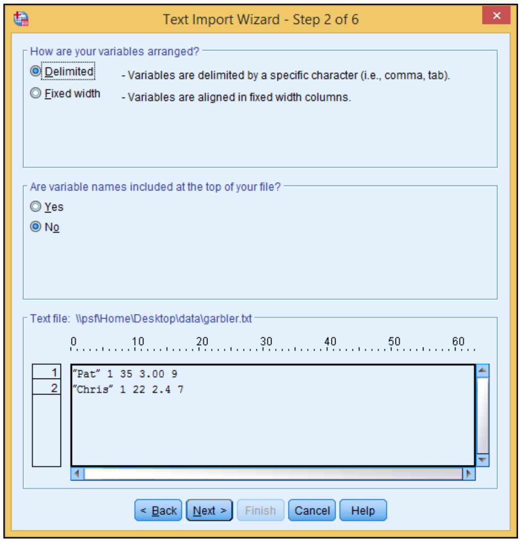
   The screen shown in Figure 5-3 appears.

5. **Specify that the data is delimited and the names are not included.**

   As you can see in this example, SPSS takes a guess, but you can also specify how your data is organized. It can be divided using spaces (as in this example), commas, tabs, semicolons, or some combination. Or your data may not be divided — it may be that all the data items are jammed together and each has a fixed width. If your text file includes the names of the variables (we show you how this works in a minute), you need to tell SPSS.

6. **Click Next.**

   The screen shown in Figure 5-4 appears.



Text Import Wizard - Step 2 of 6

How are your variables arranged?

● Delimited - Variables are delimited by a specific character (i.e., comma, tab).

○ Fixed width - Variables are aligned in fixed width columns.

Are variable names included at the top of your file?

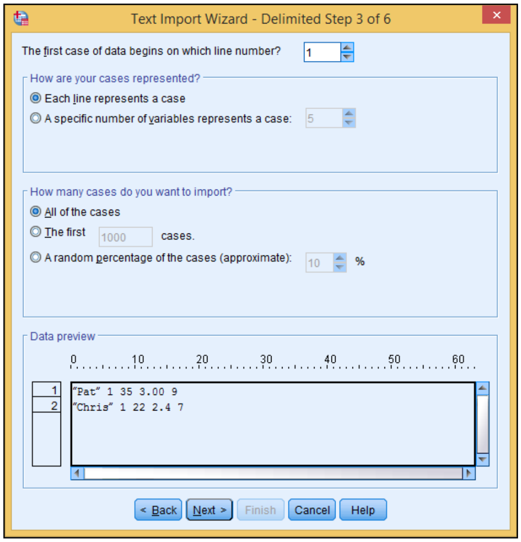
○ Yes

● No

Text file: \\psf\Home\Desktop\data\garbler.txt

```
1 "Pat" 1 35 3.00 9
2 "Chris" 1 22 2.4 7
```

Figure 5-3: Specify whether the fields are delimited and whether the variable names are included.



Text Import Wizard - Delimited Step 3 of 6

The first case of data begins on which line number? 1

How are your cases represented?

● Each line represents a case

○ A specific number of variables represents a case: 5

How many cases do you want to import?

● All of the cases

○ The first 1000 cases.

○ A random percentage of the cases (approximate): 10 %

Data preview

```
1 "Pat" 1 35 3.00 9
2 "Chris" 1 22 2.4 7
```

Figure 5-4: Specify where the data appears in the file.

7. Specify how SPSS is to interpret the text.

For this example, the correct settings are shown in Figure 5-4. You can tell SPSS something about the file and which data you want to read.

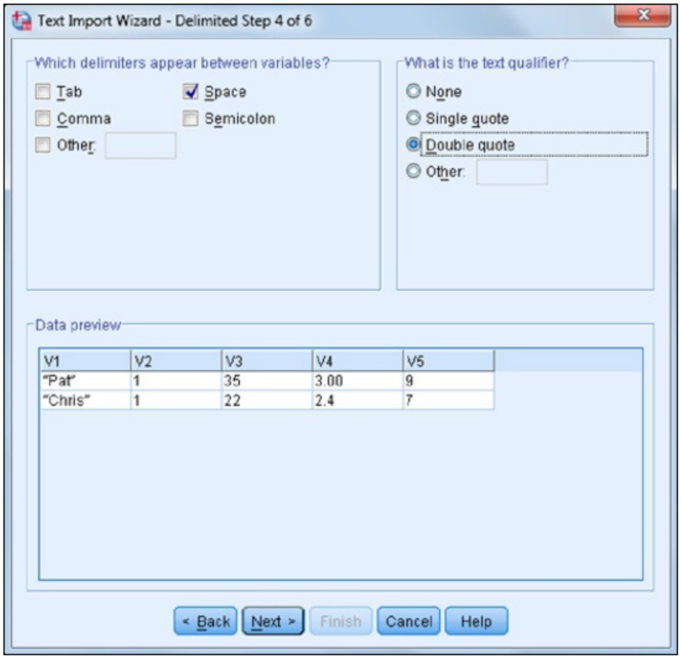
Perhaps some lines at the top of the file should be ignored — this happens when you're reading data from text intended for printing and header information is at the top. By telling SPSS about it, those first lines can be skipped.

Also, you can have one line of text represent one case (one row of data in SPSS), or you can have SPSS count the variables to determine where each row starts.

And you don't have to read the entire file — you can select a maximum number of lines to read starting at the beginning of the file, or you can select a percentage of the total and have lines of text randomly selected throughout the file. Specifying a limited selection can be useful if you have a large file and would like to test parts of it.

8. **Click Next.**

   The screen shown in Figure 5-5 appears.



Text Import Wizard - Delimited Step 4 of 6

Which delimiters appear between variables?

☐ Tab ☑ Space
☐ Comma ☐ Semicolon
☐ Other:

What is the text qualifier?

○ None
○ Single quote
● Double quote
○ Other:

Data preview

V1	V2	V3	V4	V5
"Pat"	1	35	3.00	9
"Chris"	1	22	2.4	7

Figure 5-5: Specify the delimiters that go between data items and which quotes to use for strings.

9. **Specify space as the delimiters and double quotes as text qualifiers.**

   SPSS knows how to use commas, spaces, tabs, and semicolons as delimiting characters. You can even use some other character as a delimiter by selecting Other and then typing the character into the blank. You can also specify whether your text is formatted with quotes (as in our example) and whether you use single or double quotes. Strings must be surrounded in quotes if they contain any of the characters being used as delimiters.

You can specify that a data item is missing in your text file. Simply use two delimiters in a row, without intervening data.

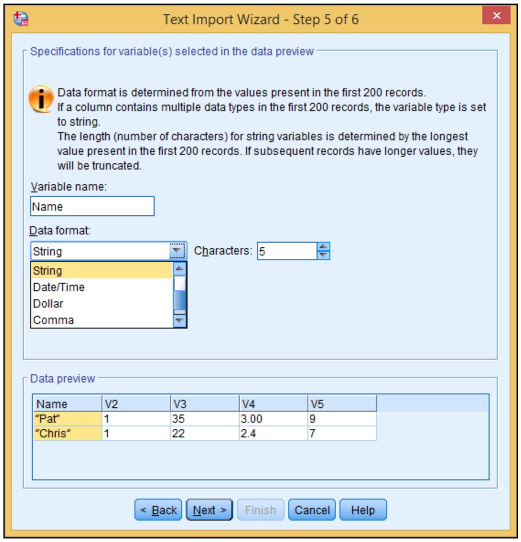
10. **Click Next.**

    The screen shown in Figure 5-6 appears.

11. **Change the variable name and data format (optional).**

    SPSS assigns the variables the names V1, V2, V3, and so on. To change a name, select it in the column heading at the bottom of the window, and then type the new name in the Variable Name field at the top. You can select the format from the Data Format drop-down list, as shown in Figure 5-6.

    This step is optional. If you need to refine your data types, you can do so later in the Variable View tab of the Data Editor window. The point here is to get the data into SPSS.


Text Import Wizard - Step 5 of 6

Specifications for variable(s) selected in the data preview

Data format is determined from the values present in the first 200 records.
If a column contains multiple data types in the first 200 records, the variable type is set to string.
The length (number of characters) for string variables is determined by the longest value present in the first 200 records. If subsequent records have longer values, they will be truncated.

Variable name: Name

Data format: String

Characters: 5

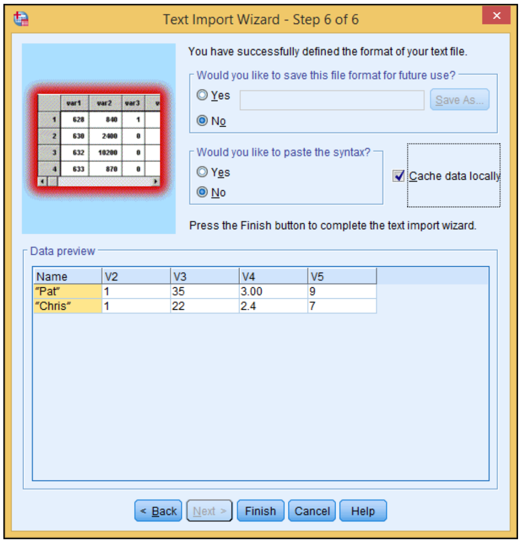
Data preview

Name	V2	V3	V4	V5
"Pat"	1	35	3.00	9
"Chris"	1	22	2.4	7

Figure 5-6: Name your variables and select their data types.

12. **Click Next.**

    The screen shown in Figure 5-7 appears.



You have successfully defined the format of your text file.

Would you like to save this file format for future use?

○ Yes

● No

Would you like to paste the syntax?

○ Yes

● No

☑ Cache data locally

Press the Finish button to complete the text import wizard.

Data preview

Name	V2	V3	V4	V5
"Pat"	1	35	3.00	9
"Chris"	1	22	2.4	7

Figure 5-7: Save the format, grab the syntax, or enable caching.

13. **In the Would You Like to Save This File Format for Future Use? Section, click No.**

    Saving the file format for future use is something you would do if you were loading more files of this same format into SPSS — it reduces the number of questions to answer and the amount of formatting to do next time.

    In the Would You Like to Paste the Syntax? section, you have the chance to grab a copy of the Syntax language instructions that do all this, but unless you know about the Syntax language (as described in Chapters 20 and 21), it's best to pretend that this option doesn't exist. (For that matter, the Cache Data Locally option is a bit odd. We don't know why it's there, unless SPSS has some problem with huge files. SPSS seems to load data faster with it than without it, but it's strictly an internal thing and SPSS works just fine either way.)

14. **Click the Finish button.**

    Depending on the type of data conversions and the amount of formatting, SPSS may take a bit of time to finish. But be patient. The Data View tab of the Data Editor window will eventually display your data.

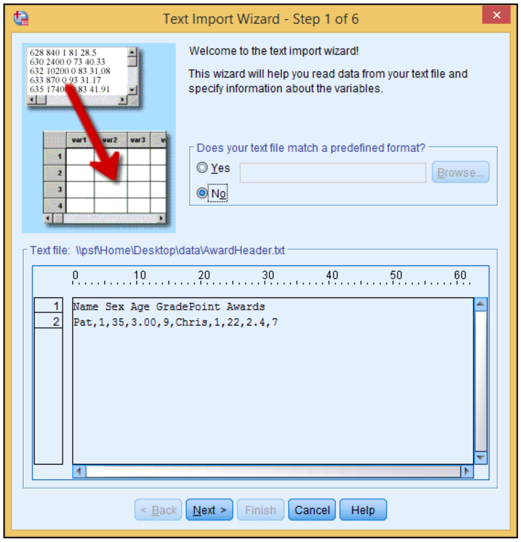
15. **Look at the data. Correct your data types and formats, if necessary. Then save it all to a file by choosing File➪Save As.**

    You're instructed to enter a filename. You can just call it `Awards`. The new file will have the `.sav` extension, which indicates that it's a standard SPSS file.

The SPSS way of reading data is a lot more flexible than this simple example demonstrates. Another example can help show why. Here, a file named `AwardHeader.txt` includes the same data, formatted slightly differently:

```
Name Sex Age GradePoint Awards
Pat,1,35,3.00,9,Chris,1,22,2.4,7
```

This time the data in the file is preceded by the variable names listed on the first line, the data is all in one long line, and the data is separated by commas. To read this into SPSS, you start the same way you did before. However, SPSS can't figure it all out in Step 1 this time (as shown in Figure 5-8). SPSS can't even tell which is header and which is data.



Welcome to the text import wizard!

This wizard will help you read data from your text file and specify information about the variables.

Does your text file match a predefined format?

○ Yes

● No

Text file: \\psf\Home\Desktop\data\AwardHeader.txt

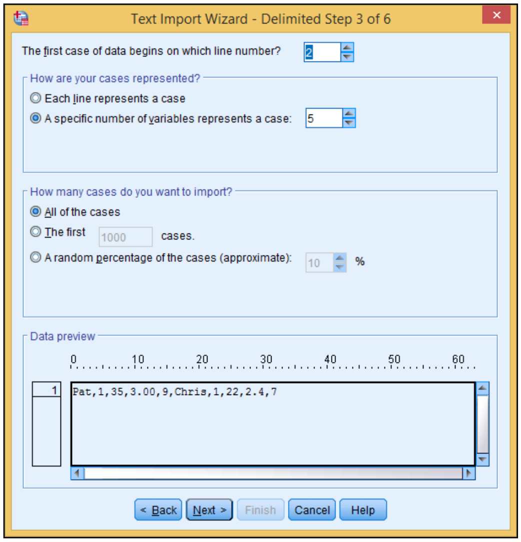
```
1 Name Sex Age GradePoint Awards
2 Pat,1,35,3.00,9,Chris,1,22,2.4,7
```

Figure 5-8: The data remains as a block of text until you explain the parts.

In Step 2 of 6, you select the option that informs SPSS that the variable names appear in the first line of text. Then, in Step 3 of 6 (as shown in Figure 5-9), you specify that the data begins on line 2 of the text file and each case has five data items.

TIP

It's possible for the data to begin several lines down in the input text file, but if variable names are present, they must be on the first line. Also, when you specify variable names, SPSS ignores the beginning and ending of lines, and counts the data values to determine when it has a complete row (case).



The first case of data begins on which line number? 2

How are your cases represented?

○ Each line represents a case

● A specific number of variables represents a case: 5

How many cases do you want to import?

● All of the cases

○ The first 1000 cases.

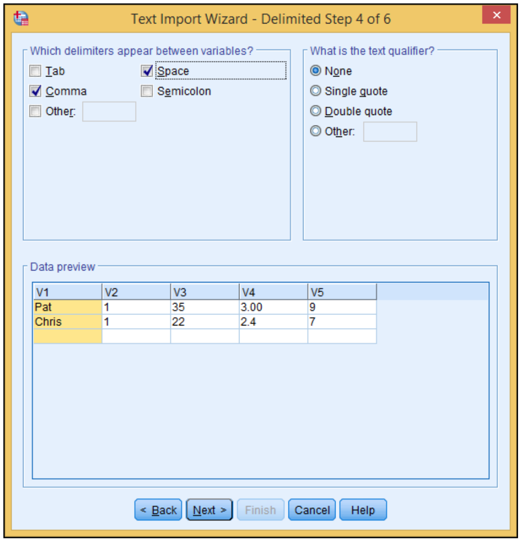
○ A random percentage of the cases (approximate): 10 %

Data preview

```
0        10        20        30        40        50        60
1  Pat,1,35,3.00,9,Chris,1,22,2.4,7
```

Figure 5-9: Specify that the data starts on line 2 and each case has five data items.

In Step 4 of 6 (shown in Figure 5-10), commas and spaces were chosen as delimiters. (Although no spaces appear in the data in this example, it doesn't hurt to include a space delimiter if it may occur somewhere in your data.) Also, None was chosen for the characters surrounding string values. In this example, SPSS figured out the spacing on its own and used these settings for its default. Also, by the time you reach Step 4 of 6, SPSS has started organizing the data according to your definitions. It has already read the variable names and included them as column headers.



Which delimiters appear between variables?

□ Tab ☑ Space

☑ Comma □ Semicolon

□ Other:

What is the text qualifier?

● None

○ Single quote

○ Double quote

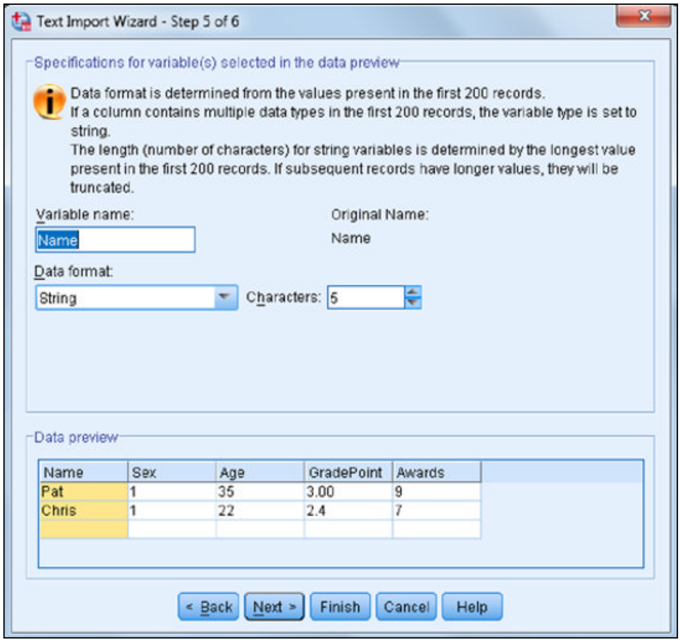
○ Other:

Data preview

V1	V2	V3	V4	V5
Pat	1	35	3.00	9
Chris	1	22	2.4	7

Figure 5-10: Specifying delimiters and quote characters.

In Step 5 of 6 (shown in Figure 5-11), you have the opportunity to change the variable names and specify their types. Here again, you see that SPSS has made a guess for the type of each one.


Text Import Wizard - Step 5 of 6
Specifications for variable(s) selected in the data preview
Data preview
< Back
Next >
Finish
Cancel
Help

Data format is determined from the values present in the first 200 records.
If a column contains multiple data types in the first 200 records, the variable type is set to string.
The length (number of characters) for string variables is determined by the longest value present in the first 200 records. If subsequent records have longer values, they will be truncated.

Variable name: Name

Original Name: Name

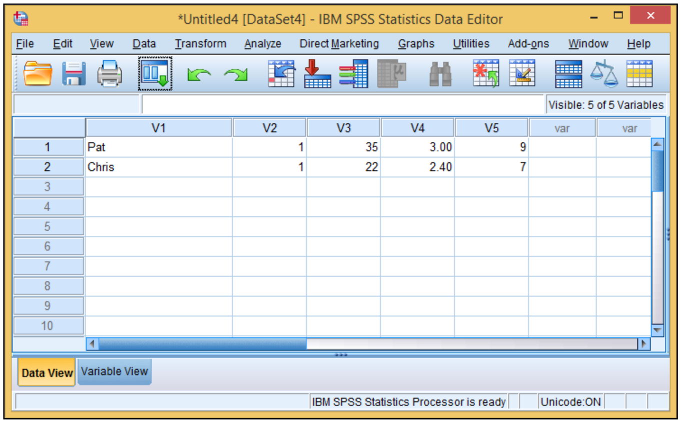
Data format: String

Characters: 5

Name	Sex	Age	GradePoint	Awards
Pat	1	35	3.00	9
Chris	1	22	2.4	7

Figure 5-11: Specifications for variables.

After you complete Step 6 of 6, click the Finish button and wait for the data to load, as shown in Figure 5-12.



Visible: 5 of 5 Variables

	V1	V2	V3	V4	V5	var	var
1	Pat	1	35	3.00	9		
2	Chris	1	22	2.40	7		
3							
4							
5							
6							
7							
8							
9							
10							

Figure 5-12: The data as formatted in SPSS.

You can see how many awards each person has, but you still have a little work to do. For example, click the Variable View tab, change the `sex` variable to a nominal data type, and assign the names "male" and "female" to the values `1` and `2`. (You can't assume anything about sex by the names.) You may want to add some descriptive labels. For example, the `awards` variable could be given the descriptive name "number of awards won during lifetime." See how a good descriptive name can clear up a little mystery?

Transferring Data from Another Program

You can get your data into SPSS from a file created by another program, but it isn't always easy. SPSS knows how to read some file formats, but if you're not careful, you'll find your data stored in an odd file format. Deciphering some file formats can be as confusing as Klingon trigonometry. SPSS can read only from file formats it knows.

SPSS recognizes the file formats of several applications. Here's a complete list:

- **IBM SPSS Statistics (`.sav`):** IBM SPSS Statistics data, and also the format used by the DOS program SPSS/PC+
- **dBase (`.dbf`):** An interactive database system
- **Microsoft Excel (`.xls`):** A spreadsheet for performing calculations on numbers in a grid
- **Portable (`.por`):** A portable format read and written by other versions of SPSS, including other operating systems
- **Lotus (`.w`):** A spreadsheet for performing calculations with numbers in a grid
- **SAS (`.sas7bdat`, `.sdy`, `.sd2`, `.ssd`, and `.xpt`):** Statistical analysis software
- **Stata (`.dta`):** Statistical analysis and graphics software
- **Sylk (`.slk`):** A symbolic link file format for transporting data from one application to another
- **Systat (`.syd` and `.sys`):** Software that produces statistical and graphical results

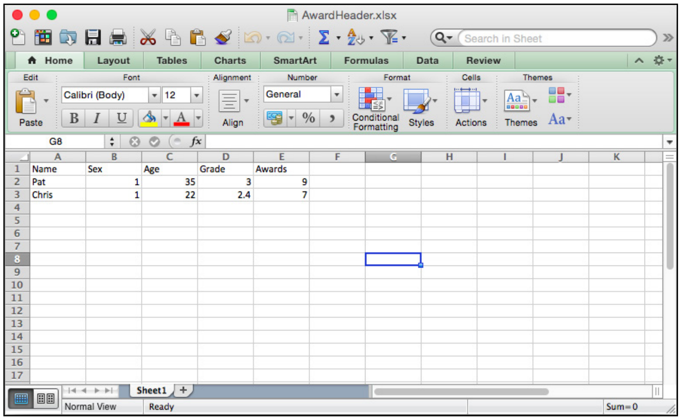
Although SPSS knows how to read any of these formats, you may still need to make a decision from time to time about how SPSS should import your dataset. But you have some advantages:

- You know exactly what you want — the form of data appearing in SPSS is simple, and what you see is what you get.
- SPSS has some reasonable defaults and makes some good guesses along the way.
- You can always fiddle with things after you've loaded them.

You're only reading from the data file, so you can't hurt it. Besides, you have everything safely backed up, don't you? Just go for it. If the process gets hopelessly balled up, you can always call it quits and start over. That's the way we do it — we think of it as a learning process.

Reading an Excel file

SPSS knows how to read Excel files directly. If you want to read the data from an Excel file, we suggest you read the steps in "Reading Simple Data from a Text File," earlier in this chapter, because the two processes are similar. If you understand the decisions you have to make in reading a text file, reading from an Excel file will be duck soup. Figure 5-13 shows the appearance of data displayed by Excel.



	A	B	C	D	E
1	Name	Sex	Age	Grade	Awards
2	Pat	1	35	3	9
3	Chris	1	22	2.4	7

Figure 5-13: A simple example of Excel spreadsheet data.

Do the following to read this data into SPSS:

1. **Save the Excel data to a file.**

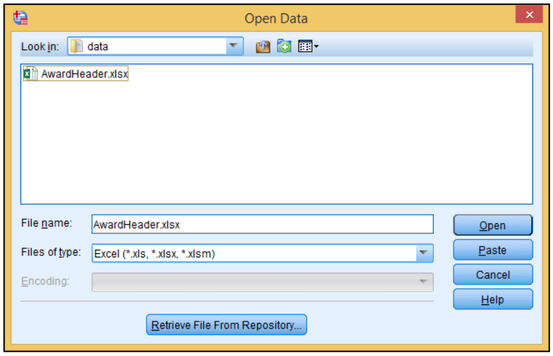
   In this example, the file is called `AwardHeader.xlsx`. If you want to copy only a portion of the spreadsheet, make a note of the cell numbers in the upper-left and lower-right corners of the group you want.

2. **Close Excel.**

   You must stop the Excel program from running before you can access the file from SPSS.

3. **Choose File➪Open➪Data.**

4. **Select the `.xlsx` file type, as shown in Figure 5-14, and then click Open.**



Look in: data

AwardHeader.xlsx

File name: AwardHeader.xlsx

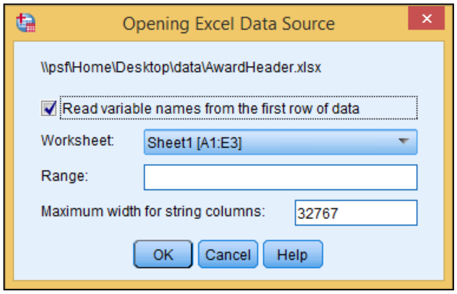
Files of type: Excel (*.xls, *.xlsx, *.xlsm)

Encoding:

Figure 5-14: From the many types of files understood by SPSS, select the Excel spreadsheet type.

5. **Select the data to include.**

   An Excel file can contain more than one worksheet, and you can choose the one you want from the drop-down list, as shown in Figure 5-15. Also, if you've elected to read only part of the data, enter the Excel cell numbers of the upper-left and lower-right corners here. You specify the range of cells the same way you would in Excel — using two cell numbers separated by a colon. Don't worry about the maximum length for strings.



\\psf\Home\Desktop\data\AwardHeader.xlsx

☑ Read variable names from the first row of data

Worksheet: Sheet1 [A1:E3]

Range:

Maximum width for string columns: 32767

Figure 5-15: Select which data in the spreadsheet to include.

6. **Click OK.**

   Your data appears in the SPSS window.

7. **Check your variables and adjust their definitions as necessary.**

   SPSS makes a bunch of assumptions about your data, and it probably makes some wrong ones. Closely examine and adjust your variable definitions by switching to the Variable View tab and making the necessary changes.

8. **Save the file using your chosen SPSS name, and you're off and running.**

Reading from an unknown program type

Often, you can transfer data from another application into SPSS by selecting, copying, and pasting the data you want, but that method has its drawbacks. The places you're copying from and pasting to are usually larger than the screen, so highlighting and selecting can be tricky. You must be ready to choose Edit➪Undo when necessary.

A better method is to write the data to a file in a format understood by SPSS, and then read that file into SPSS. SPSS knows how to read some file formats directly. Using such a file as an intermediary means you have an extra backup copy of your data, and that's never a bad idea.

Saving Data and Images

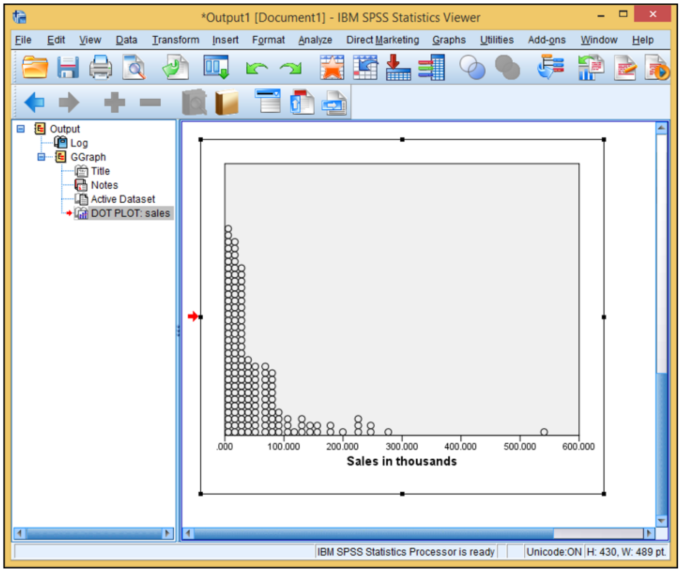
Writing data from SPSS is easier than reading data into SPSS. All you do is choose File➪Save As, select your file type, and then enter a filename. You have lots of file types to choose from. You can write your data not only in two plain-text formats, but also in Excel spreadsheet format, three Lotus formats, three dBase formats, six SAS formats, and six Stata formats.

TIP

If you'll be exporting data from SPSS into another application, find out what kinds of files the other application can read, and then use SPSS to write in one of those formats.

A second form of output from SPSS is an image. If you've generated a graphic that you want to insert into your word processor or place on your website, SPSS is ready to help you do it. (We almost wish it were hard to do so we could look smart showing you how, but it's easy.)

When you go through the steps to produce a graph, as explained in Part IV, you'll be looking at the resulting graphics in the SPSS Statistics Viewer window, which is shown in Figure 5-16.


*Output1 [Document1] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Output
Log
GGraph
Title
Notes
Active Dataset
DOT PLOT: sales
.000
100.000
200.000
300.000
400.000
500.000
600.000
Sales in thousands
Unicode:ON
H: 430, W: 489 pt.

IBM SPSS Statistics Processor is ready

Figure 5-16: SPSS Statistics Viewer displays graphs onscreen.

From SPSS Statistics Viewer, you can export images (and do some other things, too). Here's how:

1. **Produce a graph or table.**

   You can use any of the examples in Part IV to produce a graphic display. The SPSS Statistics Viewer window pops up and displays the output.

2. **Choose File➪Export.**

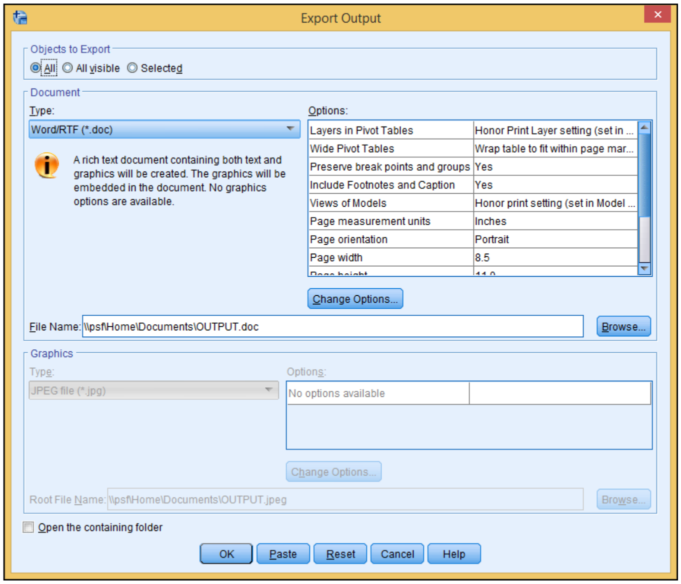
   The Export Output window, shown in Figure 5-17, appears.

3. **In the Objects to Export section, select which items to include in the output.**

   You can elect to have all objects output, all visible objects output, or only the ones you've selected. In Figure 5-16, for example, the panel on the left indicates that the graph is selected (because it's highlighted). The *visibility* of an object refers to whether its name appears in the list — if you collapse the list so a particular name can't be seen, the item is not visible. You can select items by clicking the items themselves, or by selecting their names in the list on the left.

4. **In the Document section, from the Type drop-down list, choose an output format.**

   Your choices vary according to what you decided to output as specified at the top of the window. Here's a list of the possible options:



Objects to Export

● All ○ All visible ○ Selected

Document

Type: Word/RTF (*.doc)

A rich text document containing both text and graphics will be created. The graphics will be embedded in the document. No graphics options are available.

Options:

Layers in Pivot Tables	Honor Print Layer setting (set in ...
Wide Pivot Tables	Wrap table to fit within page mar...
Preserve break points and groups	Yes
Include Footnotes and Caption	Yes
Views of Models	Honor print setting (set in Model ...
Page measurement units	Inches
Page orientation	Portrait
Page width	8.5
Page height	11.0

Change Options...

File Name: \\psf\Home\Documents\OUTPUT.doc Browse...

Graphics

Type: JPEG file (*.jpg)

Options: No options available

Change Options...

Root File Name: \\psf\Home\Documents\OUTPUT.jpeg Browse...

□ Open the containing folder

Figure 5-17: These selections control what gets output and into what format.

- *Excel 97–2004 (*.xls):* Excel files can include text, tables, and graphics, with the graphics embedded in the 97–2004 workbook. The data can create a new file or be added to an existing workbook. No graphic options are available.
- *Excel 2007 and higher (*.xlsx):* Excel files can include text, tables, and graphics, with the graphics embedded in the 2007 and higher workbook. The data can create a new file or be added to an existing workbook. No graphic options are available.
- *Excel 2007 and higher macro enabled (*.xlsm):* Excel files can include text, tables, and graphics, with the graphics embedded in the 2007 and higher macro-enabled workbook. The data can create a new file or be added to an existing workbook. No graphic options are available.
- *HTML (*.htm):* HTML files can be used for text both with and without graphics. If graphics are included, those will be exported separately, and they'll be included as HTML links. The graphic file type must also be chosen.
- *Web Reports (*.htm or *.mht):* Creates an interactive document that is compatible with most browsers, including Cognos Active Report.
- *Portable Document Format (*.pdf):* PDF documents exported will include not only text but also any graphics existing in the original. No graphics options are available.

- *PowerPoint (*.ppt):* PowerPoint documents can be written as text with the graphics embedded in the TIFF format. No graphic options are available.
- *Text-Plain (*.txt):* Text files can be output with graphic references included, and the graphics written to separate files. The reference is the name of the graphic file. The graphic file format is specified by choosing options in the lower section of this window.
- *Text-UTF8 (*.txt):* UTF-8 is Unicode text encoded as a stream of 8-bit characters. Graphics are handled the same as they are for text files.
- *Text-UTF16 (*.txt):* UTF-16 is Unicode text encoded as a stream of 16-bit characters. Graphics are handled the same as they are for text files.
- *Word/RTF (*.doc):* Word documents are written in rich text format (RTF), which can be copied into a Word document. No graphic options are available.
- *None:* When selected, this option means no text is output — only graphic images. The graphic file format is specified by options in the lower section of this window.

5. **In the Graphics section, select the image file format, if one is needed, from the Type drop-down list.**

   You may be asked to select a format for your image file(s). You can select from `.png`, `.bmp`, `.emf`, `.eps`, `.jpg`, or `.tif`.

6. **Click the Browse button, select the directory and root filename, and click Save.**

   Depending on what you chose to output, the actual output may be multiple files, and they'll all have names derived from the root name you provide. The Save button doesn't write the file(s) — it only inserts your selected name into the Export Output window.

7. **Click OK.**

   The file(s) are written to disk — each in the chosen format, at the chosen location.

Chapter 13

Using Descriptive Statistics

In This Chapter

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

Summaries of individual variables provide the basis for more complex analysis (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, however, given the measurement level of the variables, only some information is meaningful.

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

Looking at Levels of Measurement

The level of measurement of a variable determines the appropriate statistics, and graphs that can be used to describe the data. For example, if we have a variable like marital status, it wouldn't make sense to ask for the mean of this variable; instead, we may ask for the percentages associated with the different categories. In addition, level of measurement also determines the kind of research questions we can answer, so it's a critical step in the research process.

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. 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 can be said to contain the properties of the preceding types and records information at a higher level. The four levels of measurement are as follows:

- **Nominal:** For nominal data, each value represents a category. There is no inherent order to the categories. For example, the variable gender may be coded as 0 (male) and 1 (female), but all these values tell us is that we have two 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, but there is a meaningful order or rank to the categories. However with ordinal data, there is not a measurable distance between categories. For example, if we're measuring the outcome of a foot race, we can determine which contestant came in first, second, third, and so on. However, based on the ranking, we can't tell how much faster each competitor was compared to the others, nor can we 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:** For interval data, a one-unit change in numeric value represents the same change in quantity regardless of where it occurs on the scale. For example, for a variable like 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, they have equal intervals between points on the scale.
- **Ratio:** For ratio data, you have all the properties of interval variables with the addition of a true zero point, representing the complete 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 like number of purchases is a ratio variable because zero indicates no purchases. Ratios can then be calculated (for example, eight purchases represents twice as many purchases as four purchases).

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

- **Categorical:** Nominal and ordinal measurement levels
- **Continuous (or scale):** Interval and ratio measurement levels

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. There are three main measures of central tendency:
  - **Mode:** The category or value that contains the most cases. 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; it is the 50th percentile. If all the cases for a variable are arranged in order according to their value, the median is the value that splits the data into two equally sized groups.
  - **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).
- **Variability:** The amount of spread or dispersion around the measure of central tendency. There are a number of measures of variability:
  - **Maximum:** The highest value for a variable.
  - **Minimum:** The lowest value in the distribution.
  - **Range:** The difference between the maximum and minimum values.
  - **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 measure is expressed in the units of the variable squared. This can cause difficulty in interpretation, so more often, the standard deviation is used. The standard deviation restores the value of variability to the units of measurement of the original variable.

We care about level of measurement because it determines appropriate summary statistics and graphs to describe the data. Table 13-1 summarizes the most common summary statistics and graphs for each of the measurement levels used by SPSS.

Table 13-1 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	Mode, median	Mode, median, mean
Measures of dispersion	None	Min/max/range	Min/max/range, standard deviation/variance
Graph	Pie or bar	Pie or bar	Histogram

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 like 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.**

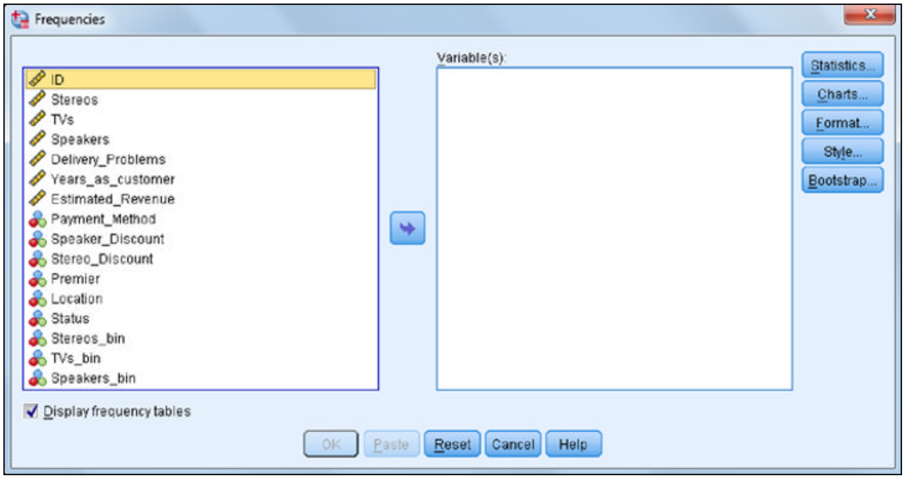
   The file is not in the SPSS installation directory. You have to download it from this book's companion website.

   It contains the customer's purchase history and has 16 variables and 3,338 cases.

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

   The Frequencies dialog box, shown in Figure 13-1, appears.

   In this example, we 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.



☑ Display frequency tables

Figure 13-1: The Frequencies dialog box.

3. **Select the variables `Payment_Method`, `Premier`, and `Status`, and place them in the Variable(s) box, as shown in Figure 13-2.**

   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.

4. **Click the Statistics button.**

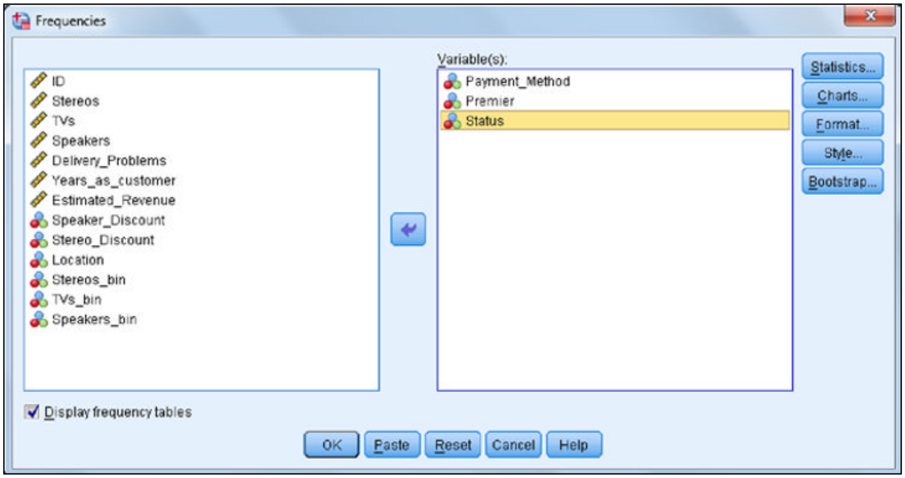
   The Frequencies: Statistics dialog box, shown in Figure 13-3, appears.

   This dialog box 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, only the mode is suitable.

5. **In the Central Tendency section, select the Mode check box, as shown in Figure 13-4.**

6. **Click Continue.**

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



☑ Display frequency tables

Figure 13-2: Place the variables in the Variable(s) box.

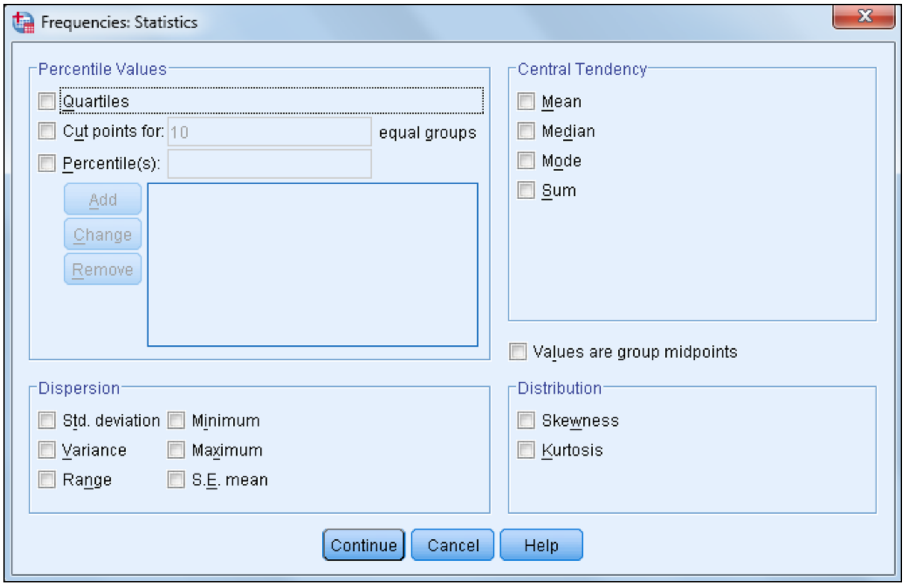
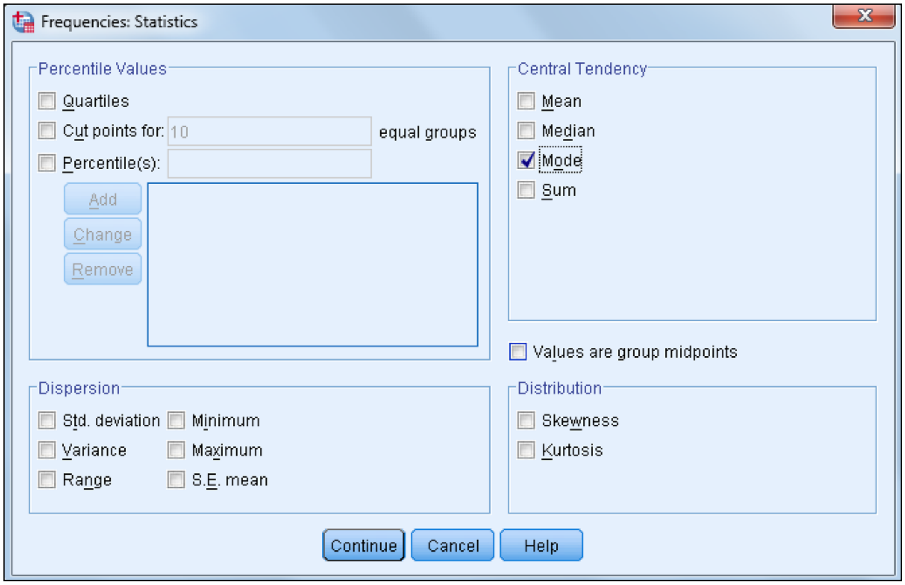


Figure 13-3: The Frequencies: Statistics dialog box.

7. Click the Charts button.

The Frequencies: Charts dialog box, shown in Figure 13-5, appears.

This dialog box 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.



Frequencies: Statistics

Percentile Values

□ Quartiles

□ Cut points for: 10 equal groups

□ Percentile(s):

Central Tendency

□ Mean

□ Median

☑ Mode

□ Sum

□ Values are group midpoints

Dispersion

□ Std. deviation □ Minimum

□ Variance □ Maximum

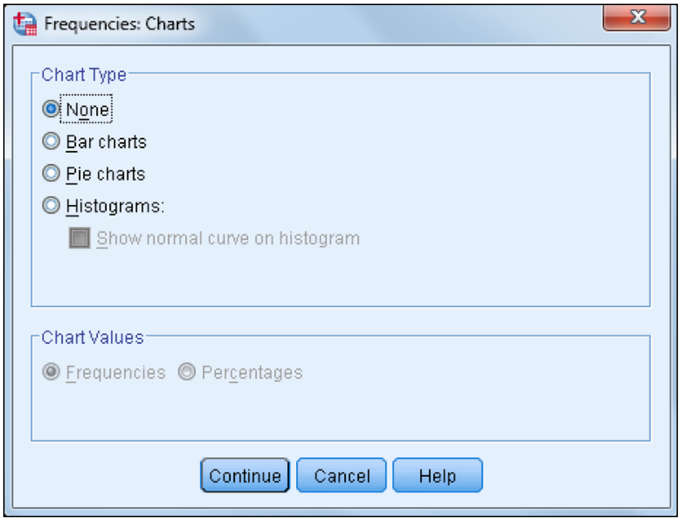
□ Range □ S.E. mean

Distribution

□ Skewness

□ Kurtosis

Figure 13-4: Select the Mode check box.



Frequencies: Charts

Chart Type

● None

○ Bar charts

○ Pie charts

○ Histograms:

□ Show normal curve on histogram

Chart Values

● Frequencies ○ Percentages

Figure 13-5: The Frequencies: Charts dialog box.

8. **In the Chart Type section, click the Bar Charts radio button; in the Chart Values section, click the Percentages radio button (see Figure 13-6).**

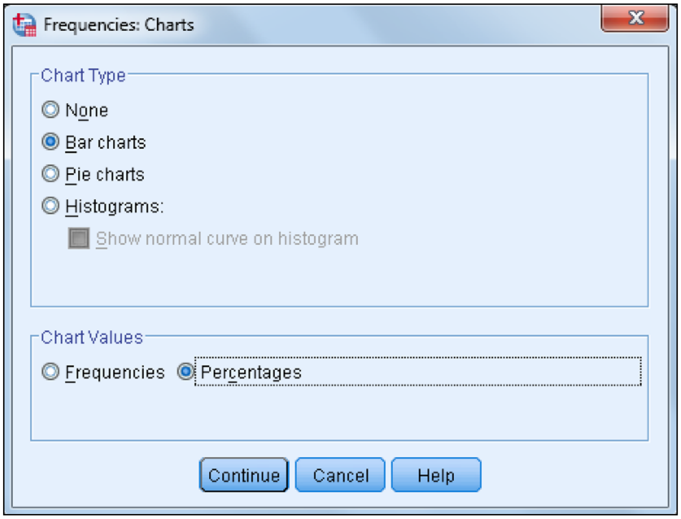


Chart Type

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

Chart Values

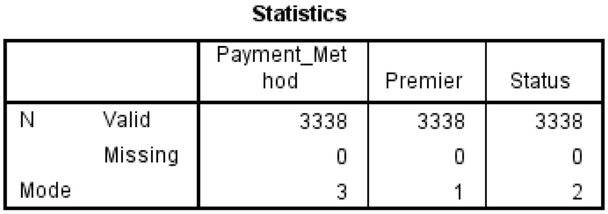
○ Frequencies ● Percentages

Figure 13-6: Click Bar Charts and Percentages.

9. **Click Continue.**
10. **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 13-7) displays the number of valid and missing cases for each variable requested in the Frequencies procedure.



Statistics

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

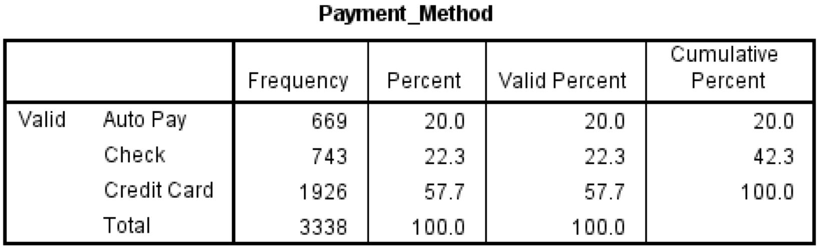
Figure 13-7: The Statistics table.

Be sure to review this table to check the number of missing cases. In this example, we have 3,338 valid cases and we don't have any missing data.

The Statistics table also displays any additional statistics that were requested. In our case, we 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 represents the category of “Credit Card” for Payment_Method, “No” for Premier, and the “Current” group for Status.

We could’ve asked for additional summary statistics like the mean, and the frequencies procedure would’ve produced it. This is why it’s important to understand measurement level and what statistics are relevant.

The Frequency table (shown in Figure 13-8) shows the distribution of the variable Payment_Method (in this case, we focus on the variable Payment_Method because all the other Frequency tables will have similar information). The information in the Frequency table is comprised of counts and percentages. The Frequency column contains counts, or the number of occurrences of each data value. So, for the variable Payment_Method, 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 our current example, those 1,926 customers who paid via credit card account for 57.7% 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.



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 13-8: The Frequency table for the Payment_Method variable.

Bar charts (like the one in Figure 13-9) summarize the distribution that was observed in the Frequency table and allow you to see the distribution. For the variable Payment_Method, more than half of the people fall into the Credit Card category.

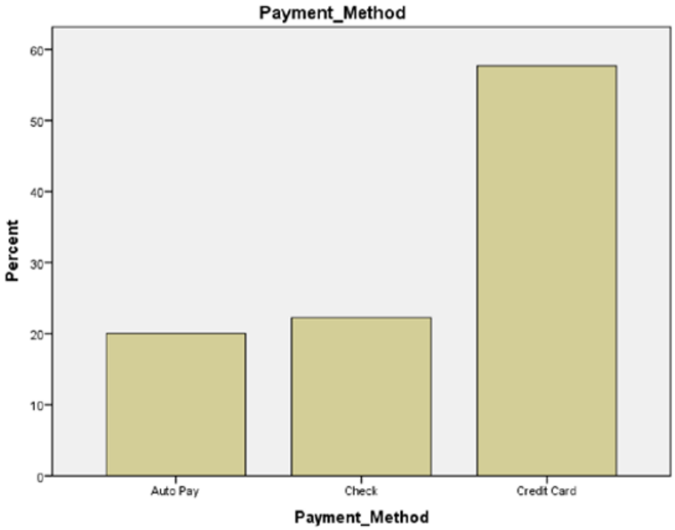


Figure 13-9: A bar chart for the `Payment_Method` variable.

Understanding Frequencies for Continuous Variables

As we have seen, frequency tables show counts and percentages, which are extremely useful when working with categorical variables. However, for continuous variables that have many values, frequency tables become less useful. For example, if we were working with a variable like income, it wouldn't be very useful to know that there was only one person in the dataset who made $22,222 last year. In this case, it's likely that each response would have a different value, so the frequency table would be very large and not particularly 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 the frequencies for continuous variables, follow these steps:

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

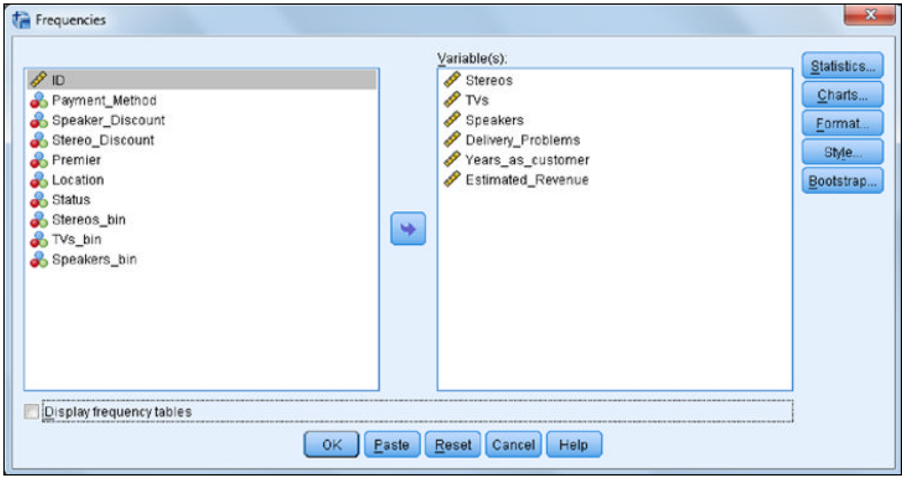
   The file is not in the SPSS installation directory. You have to download it from this book's companion website.

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

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

4. **Deselect the Display Frequency Tables check box, as shown in Figure 13-10.**

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



ID
Payment_Method
Speaker_Discount
Stereo_Discount
Premier
Location
Status
Stereos_bin
TVs_bin
Speakers_bin

Variable(s):
Stereos
TVs
Speakers
Delivery_Problems
Years_as_customer
Estimated_Revenue

□ Display frequency tables

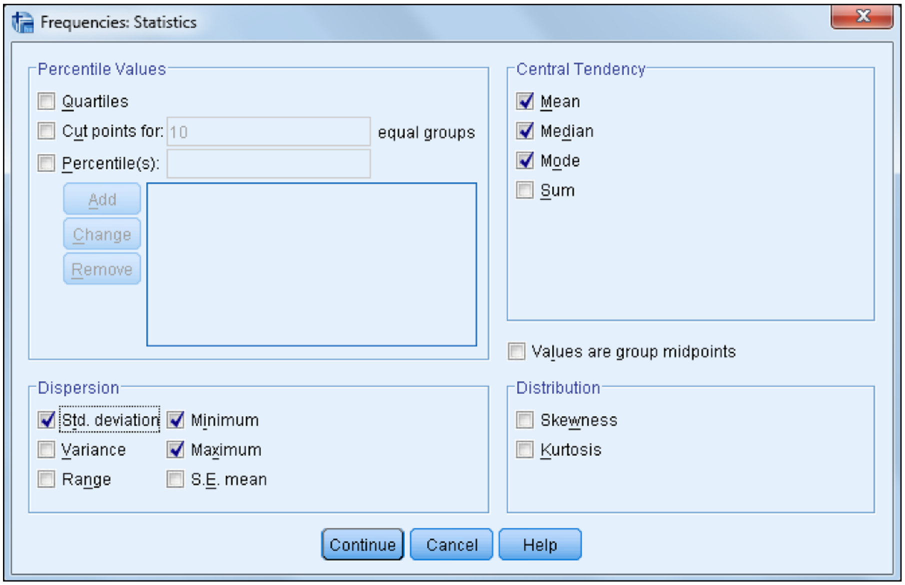
Figure 13-10: The Frequencies dialog box.

5. **Click the Statistics button.**

   The Frequencies: Statistics dialog box appears.

   Several 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 those 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 (see Figure 13-11).**



Frequencies: Statistics

Percentile Values

☐ Quartiles

☐ Cut points for: 10 equal groups

☐ Percentile(s):

Central Tendency

☑ Mean

☑ Median

☑ Mode

☐ Sum

☐ Values are group midpoints

Dispersion

☑ Std. deviation ☑ Minimum

☐ Variance ☑ Maximum

☐ Range ☐ S.E. mean

Distribution

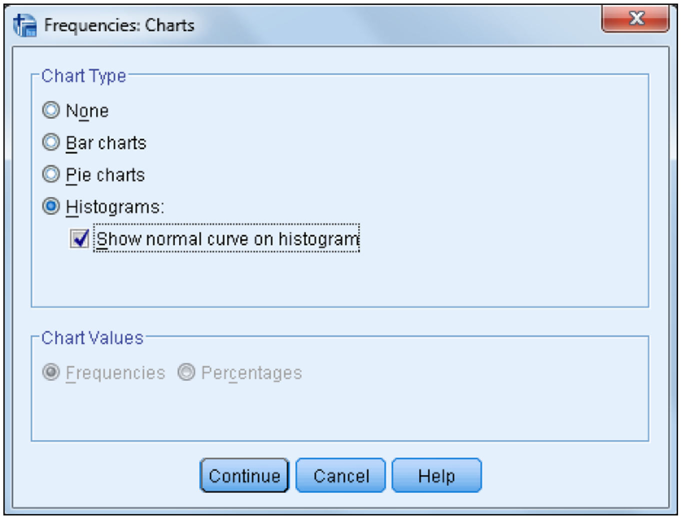
☐ Skewness

☐ Kurtosis

Figure 13-11: The Frequencies: Statistics dialog box.

7. **Click Continue.**
8. **Click the Charts button.**

   The Frequencies: Charts dialog box appears.



Frequencies: Charts

Chart Type

○ None

○ Bar charts

○ Pie charts

● Histograms:

☑ Show normal curve on histogram

Chart Values

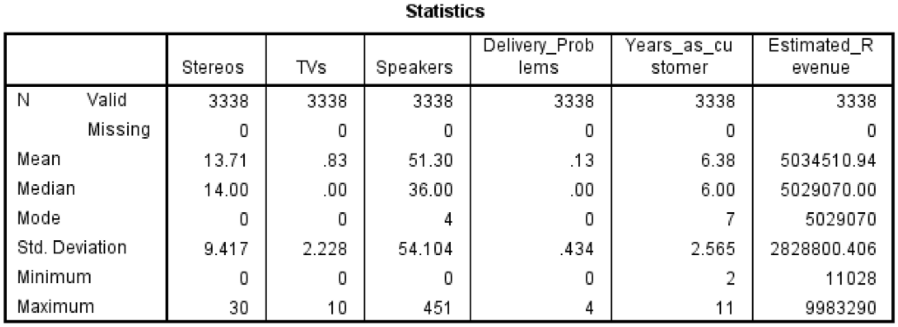
● Frequencies ○ Percentages

Figure 13-12: The Frequencies: Charts dialog box.

9. **Click the Histograms radio button and select the Show Normal Curve on Histogram check box, as shown in Figure 13-12.**
10. **Click Continue.**
11. **Click OK.**

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

The Statistics table (shown in Figure 13-13) shows that we have 3,338 valid cases and we don't have any missing data. The Statistics table contains the requested statistics. For example, for the variable `Speakers`, we can see that the minimum value is 0 and the maximum value is 451. This seems like a very large range of values, so it would be useful to double-check the data to make sure there are no errors. Likewise, in an ideal world, we 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 variable `Speakers`, 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).



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	2828800.406
Minimum		0	0	0	0	2	11028
Maximum		30	10	451	4	11	9983290

Figure 13-13: The Statistics table.

You can visually check the distribution of these variables with a histogram (see Figure 13-14). A histogram has bars, but, unlike the bar chart, they're plotted along an equal interval scale. The height of each bar is the count of values falling within the interval. Notice 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.

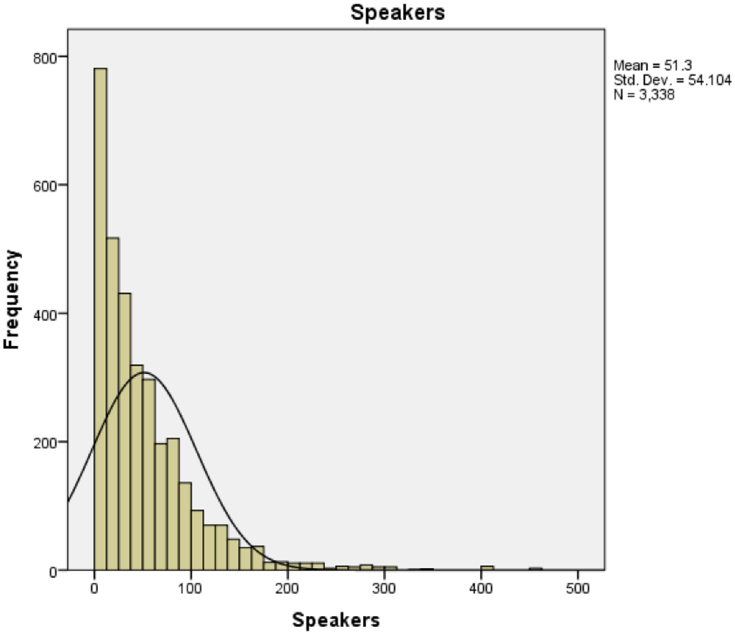


Figure 13-14: A histogram for the `Speakers` variable.

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:

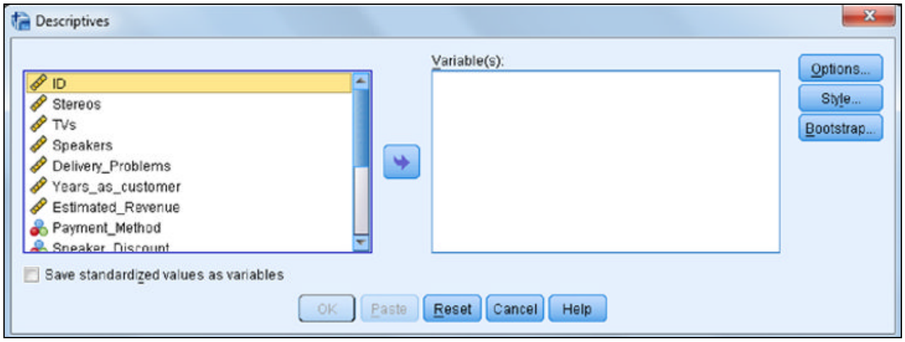
1. **From the main menu, choose File➪Open➪Data and load the `merchandise.sav` data file.**

   The file is not in the SPSS installation directory. You have to download it from this book's companion website.

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

   The Descriptives dialog box, shown in Figure 13-15, appears.

3. **Select the variables `Stereos, TVs, Speakers, Delivery_Problems, Years_as_customer,` and `Estimated_Revenue,` and place them in the Variable(s) box, as shown in Figure 13-16.**

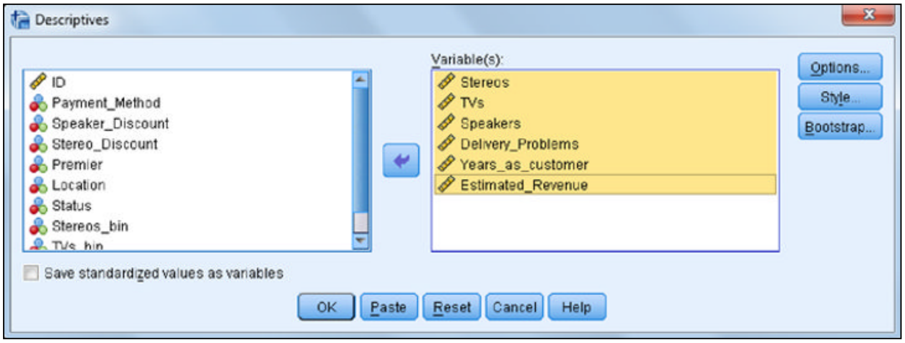


☐ Save standardized values as variables

Figure 13-15: The Descriptives dialog box.

4. **Click OK.**

   SPSS runs the descriptives procedure and calculates the summary statistics.

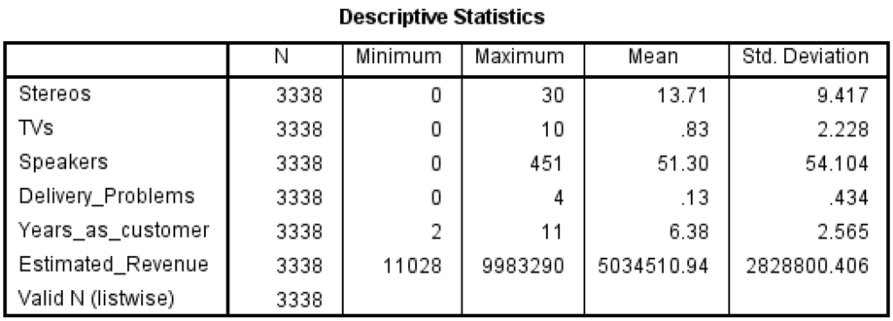


Variable(s):
Stereos
TVs
Speakers
Delivery_Problems
Years_as_customer
Estimated_Revenue

☐ Save standardized values as variables

Figure 13-16: 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 13-17). Likewise, it's always important to investigate the mean and determine if the value makes sense. Sometimes a mean may be lower or higher than expected, which can indicate a problem relating to how the data was coded or maybe even collected. Finally, the last row in the table, Valid N (listwise), gives the number of cases that have a valid value on all the variables appearing in the table. In this example, we have no missing data, so this number isn't particularly useful for this set of variables. However, it would be useful for a set of variables that you intended to use for a *multivariate analysis* (an analysis looking at the relationships between many variables).



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	2828800.406
Valid N (listwise)	3338				

Figure 13-17: The Descriptive Statistics table.