Chapter 1

Introducing IBM SPSS Statistics

A statistic is a number. A raw statistic is a measurement of some sort. It is fundamentally a count of something — occurrences, speed, amount, or whatever. IBM SPSS Statistics is a piece of software that takes in raw data and combines them into new statistics that can be used as predictors.

"There are three kinds of lies: lies, damn lies, and statistics." That statement is often attributed to Mark Twain, but that's not quite right. Mark Twain did say it, but he attributed it to someone else. He indirectly attributed it to Disraeli, but his attribution was vague, and the original statement, if it exists, can't be located. Speaking statistically, the odds are we'll never know who said it first.

Garbage In, Garbage Out

Statistical analysis is like a sewer. What you get out of it depends on what you put into it.

Eighty-two percent of all statistics are made up on the spot to try to prove a point.

If you're not careful, you can conclude just about anything from your data and your calculations. SPSS performs calculations for you, but the raw data, and which calculations are performed, are up to you.

Let me show you a simple example of using raw data to produce an obviously wrong conclusion. Suppose you want to demonstrate, by sampling, that every odd number is prime. (A prime number can be evenly divided only by 1 and itself.) The first thing to do is gather a collection of data points, as shown in Table 1-1.

Table 1-1 Odd Numbers and Whether They Are Prime

Number	*Prime?*	*Comment*
1	Yes	It fits the definition exactly
3	Yes	It is certainly both odd and prime
5	Yes	It fits the pattern of primes
7	Yes	So far, so good
9	No	Must be a bad data point, so throw it out
11	Yes	Now we're back on track
13	Yes	Looking good

Lots of things are already wrong with the data in Table 1-1. For one, the sample is too small. For another, the sampling cannot be considered random. All too often it happens that data points are omitted if they don't fit a preconceived conclusion. The result of the data in this table can be used as "proof" of a "fact" that is dead wrong.

This book is not about the accuracy, correctness, or completeness of the input data. Your data is up to you. This book shows you how to take the numbers you already have, put them into SPSS, crunch them, and display the results in a way that makes sense. Gathering valid data and figuring out which crunch to use is up to you.

Where Did SPSS Come From?

SPSS is probably older than you are. In 2009 it became 40 years old, and the average age of an American is 35.3.

At Stanford University in the late 1960s, Norman H. Nie, C. Hadlai (Tex) Hull, and Dale H. Bent developed the original software system named Statistical Package for the Social Sciences (SPSS). They needed to analyze a large volume of social science data, so they wrote software to do it. The software package caught on with other folks at universities and, consistent with the open-source tradition of the day, the software spread through universities around the country.

The three men produced a manual in the 1970s, and the software's popularity took off. A version of it existed for each of the different kinds of mainframe computers existing at the time. Its popularity spread from universities into other areas of government, and it began to leak out into private enterprise.

In the 1980s, a version of the software was moved to the personal computer. In 2008, the name was briefly changed to Predictive Analysis Software (PASW). In 2009, SPSS Inc. was acquired by IBM Corporation and the name of the product was returned to the more familiar SPSS. The official name of the software today is IBM SPSS Statistics.

Maybe it has been continuously successful because the software does such a good job of making predictions, and the SPSS people could always figure out what they should do next.

The practical application of the software has always been to attempt to predict the future. Predictive models are used on business data to identify both risks and opportunities. Relationships among many factors are analyzed to guide decision-makers in selecting from among a number of possible actions.

The software is available in several forms — single user, multiuser, client-server, student version, and so on. The software also has a number of special purpose add-ons available. You can find out about them all at the following Web site: `www.spss.com`

The Four Ways to Talk to SPSS

More than one way exists for you to command SPSS to do your bidding. You can use any of four approaches to perform any of the SPSS functions, but the one you should choose depends not only on which interface you prefer, but also (to an extent) on the task you want performed. The available interfaces are as follows:

- **GUI (graphic user interface):** SPSS has a windowing interface; you can issue commands by using the mouse to make menu selections that cause dialog boxes to appear. This is a fill-in-the-blanks approach to statistical analysis that guides you through the process of making choices and selecting values. The advantage of the GUI approach is that, at each step, SPSS makes sure you enter everything necessary before you can proceed to the next step. This is the preferred interface for those just starting out — and if you don't go into depth with SPSS, this may be the only interface you ever use.
- **Syntax:** This is the internal language used to command actions from SPSS. It is the command syntax of SPSS, hence its name. It's often referred to as *the command language*. You can use the Syntax command language to enter instructions into SPSS and have it do anything it's capable of doing. In fact, when you select from menus and dialog boxes to command SPSS, you're actually generating Syntax commands internally that do your bidding. That is, the GUI is nothing more than the *front*

end of a Syntax command-writing utility. Writing (and saving) command-language programs is a good way to create processes that you expect to repeat. You can even grab a copy of the Syntax commands generated from the menu and save them to be repeated later.

- **Python:** This is a general-purpose language that has a collection of SPSS modules written for it; you can use it to write programs that work inside SPSS. You can also run Python with the Syntax language to command SPSS to perform statistical functions. One advantage of using Python is that it's a modern language, complete with the power and convenience that come with such languages, including the capability of constructing a more readable program. In addition, because Python is a general-purpose language, you can read and write data in other applications and in files.
- **Scripts:** The items that SPSS calls *scripts* are actually programs written in BASIC. This language is simple and many people are familiar with it. Also, a BASIC program can be written as an *autoscript* — a script that executes automatically whenever SPSS produces certain output. Both BASIC and Python are scripting languages, but where the SPSS documentation talks about a script, it is referring to a BASIC program.

What You Can and Cannot Do with SPSS

The full-blown SPSS package comes in many parts. The *Base system* is the center around which the rest of SPSS revolves. If you have SPSS, you have a Base system.

You may also have one or more add-ons. With only one exception — the Python programming language, which requires some additional software available for free on the SPSS distribution CD — everything described in this book is included in the Base system, so you will be able to do anything you read about. Chapter 20 describes other modules you can add to your Base system.

SPSS works with numbers. Only. If you cannot express your information as a number, you can't run it through SPSS. You will see names and descriptions seemingly being processed by SPSS, but that's because each name has been assigned a number. (Sneaky.) That's why survey questions are written like this: "How *much* do you enjoy eating rhubarb? Select your answer: Very much, sort of, don't care, not really, I hate the stuff." A number is assigned to each of the possible answers, and these numbers are fed through the statistical process. SPSS uses the numbers, not the words, so be careful about keeping all your words and numbers straight.

You must keep accurate records describing your data, how you got the data, and what it means. SPSS can do all the calculations for you, but only you can decipher what it means. In *The Hitchhiker's Guide to the Galaxy,* a computer the size of a planet crunched on a problem for generations and finally came out with the answer, 42. But the people tending the machine had no idea what the answer meant because they didn't remember the question. They hadn't kept track of their input. You must keep careful track of your data or you may later discover, for example, that what you've interpreted to be a simple increase is actually an increase in your rate of decrease. Oops.

SPSS lets you enter the data and tag it to help keep it organized, but you already have the data written down someplace and fully annotated. Don't you?

How SPSS Works

The developers of SPSS have made every effort to make the software easy to use. It prevents you from making mistakes or even forgetting something. That's not to say it's impossible to do something wrong, but the SPSS software works hard to keep you from running into the ditch. To foul things up, you almost have to work at figuring out a way of doing something wrong.

You always begin by defining a set of *variables,* then you enter data for the variables to create a number of *cases.* For example, if you're doing an analysis of automobiles, each car in your study would be a case. The variables that define the cases could be things such as the year of manufacture, horsepower, and cubic inches of displacement. Each car in the study is defined as a single case, and each case is defined as a set of values assigned to the collection of variables. Every case has a value for each variable. (Well, you *can* have a missing value, but that's a special situation described later.)

Each variable is a specific type. That is, each variable is defined as containing a certain kind of number. For example, a *scale* variable is a numeric measurement, such as weight or miles per gallon. A *categorical* variable contains values that define a category; for example, a variable named `gender` could be a categorical variable defined to contain only values 1 for female and 2 for male. Things that make sense for one type of variable don't necessarily make sense for another. For example, it makes sense to calculate the average miles per gallon, but not the average gender.

After your data is entered into SPSS — your cases are all defined by values stored in the variables — you can easily run an analysis. You've already finished the hard part. Running an analysis on the data is simple compared to entering the data. To run an analysis, you select the one you want to run

from the menu, select appropriate variables, and click the OK button. SPSS reads through all your cases, performs the analysis, and presents you with the output as tables or graphs.

You can instruct SPSS to draw graphs and charts directly from your data the same way you instruct it to do an analysis. You select the desired graph from the menu, assign variables to it, and click OK.

When you're preparing SPSS to run an analysis or draw a graph, the OK button is unavailable until you've made all the choices necessary to produce output. Not only does SPSS require that you select a sufficient number of variables to produce output, it also requires that you choose the right kinds of variables. If a categorical variable is required for a certain slot, SPSS will not allow you to choose any other kind. Whether the output makes sense is up to you and your data, but SPSS makes certain that the choices you make can be used to produce some kind of result.

All output from SPSS goes to the same place — a dialog box named SPSS Viewer. It opens to display the results of whatever you've done. After you have produced output, if you perform some action that produces more output, the new output is displayed in the same dialog box. And almost anything you do produces output.

Where SPSS Works

More than one version of IBM SPSS Statistics 18 exists, for execution under different operating systems.

IBM SPSS Statistics 18 for Windows can be run on Windows XP (32-bit) or on Windows Vista (32-bit or 64-bit). You can run IBM SPSS Statistics 18 for Mac on Macintosh 10.5*x* (Leopard) or on Macintosh 10.6*x* (Snow Leopard), both 32- and 64-bit. IBM SPSS Statistics 18 for Linux has been tested only on Red Hat Enterprise Linux 5 and Debian 4.0, but it should run on any sufficiently updated Linux system.

All the Strange Words

Statistics seems to have been born in the land of strange words. Lots of them. If you come across a term that you don't understand, such as *dichotomy, variable,* or *kurtosis,* you're not stopped: You can look it up in the glossary at the back of this book.

It's not only new words that can trip you up. You will find common words used in a special way. For example, the word *case* has a special meaning. And a *break variable* has a special purpose when organizing tabular data.

All Those Files

Input data and statistics are stored in files. Different kinds of files. Some files contain numbers and definitions of numbers. Some files contain graphics. Some files contain both.

The examples in this book require the use of files that contain data configured to demonstrate capabilities of PASW. Some of the files are already on your computer, and others can be found on the Internet. Most are in the same directory you used to install PASW. That is, the action of installing PASW also installs a number of data files ready to be loaded into PASW and used for analysis. A few of the files used in the examples can be found in the compressed file `PASW.zip` found on this book's companion Web site (it's listed in the Introduction).

Where to Get Help When You Need It

You're not alone. Some immediate help comes directly from the PASW software package, and other help can be found on the Internet. If you find yourself stumped on some point, you can look in several places, as follows:

- **Topics:** Choosing Help➪Topics from the main window of the PASW application is your gateway to immediate help. The help is somewhat terse, but often it provides exactly what you need. The information is in one large help document, presented one page at a time. Choose Contents to select a heading from an extensive table of contents, choose Index to search for a heading by entering its name, or choose Search to enter a string search inside the body of the help text.

  In the help directory, the titles in all uppercase are descriptions of Syntax language commands.

- **Tutorial:** Choose Help➪Tutorial to open a dialog box with the outline of a tutorial that guides you through many parts of PASW. You can start at the beginning and view each lesson in turn, or you can select your subject and view just that.

- **Case Studies:** Choose Help➪Case Studies to open a dialog box containing examples in a format similar to that of the Tutorial selection. You can select titles from its outline and view descriptions and examples of specific instances of using PASW. You can also find descriptions of the different types of calculations. If some particular analysis type is eluding your comprehension, this is a good place to look.
- **Statistics Coach:** Choose Help➪Statistics Coach if you have a good idea of what you want to do but need some specific information on how to go about doing it.
- **Command Syntax Reference:** Choose Help➪Command Syntax Reference to display more than 2000 pages of references to the Syntax language in your PDF viewer. The regular help topics, mentioned previously, provide a brief overview of each topic, but this document is much more detailed.
- **Algorithms:** Choose Help➪Algorithms to get detailed information on how processes work internally. This is where you can dive far down into the internals. If you want to take a look at the math and how it's applied, this is where you look.

Your Most Valuable Possession

The most valuable possession you have in dealing with statistics is not your computer. It's not your PASW software. It's not even this book, or any other book you may be using to learn statistical procedures. You can lose any one of those, but any one of them can be replaced.

Your most valuable possession is your data. Sure, you can always go and get more data, but you can't go and get the *same* data. The world doesn't hold still long enough. Be sure to make backup copies of your data.

Back up your data to memory that does not live in the same building with the computer you're using. You can swap backups with a friend, or if you have access to a remote Web site, you can stuff files in a blind directory.

This message about backing up your data comes to you from someone who has been stung. And I don't want to talk about it again. Ever.

You Can Dive as Deep as You Want to Go

PASW makes no effort to keep anything secret. It's designed to be as easy to use as possible, so you really don't have to know all that much to make it work. However, if you want to understand how things are working internally, you can find out if you dig. And you don't have to dig very far. Choosing Help is the first step to finding out anything you want to know about what's going on inside.

Let's say you're working on your numbers and want to use some specific algorithm to do your calculations. PASW has been at this longer than you have, so the algorithm you want to use is almost certainly built in. If you're not sure exactly what PASW is doing to calculate some of the numbers, you can go to the Help menu and read through the supplied documentation to find out how the calculations are being performed. But, before you start looking, make sure you really want to know, because the equations and how they are applied are explained in excruciating detail.

The purpose of this book is to give the shallow divers enough information to be able to swim and to show the deeper divers how to begin. I don't explain all the details because there are too many. There's simply not enough room in a book this size to explain PASW in depth.