Statistics for Linguists

Statistics for Linguists:

A Step-by-Step Guide for Novices

By

David Eddington

Cambridge
Scholars
Publishing

Statistics for Linguists: A Step-by-Step Guide for Novices

By David Eddington

This book first published 2015

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Copyright © 2015 by David Eddington

Reprints throughout Courtesy of International Business Machines Corporation, © International Business Machines Corporation. SPSS® Inc. was acquired by IBM® in October, 2009.

All rights for this book reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the copyright owner.

ISBN (10): 1-4438-7638-0
ISBN (13): 978-1-4438-7638-4

FOR SILVIA

CONTENTS

ACKNOWLEDGEMENTS

So many people have had a hand in the production of this book. I am indebted to the College of Humanities and the Department of Linguistics and English Language at Brigham Young University for their financial support. Mel Thorne, Jennifer McDaniel, and the staff at the Faculty Editing Service are responsible for the production of a much more polished manuscript than I could have managed on my own. Karen Grace-Martin of the Analysis Factor also deserves a great deal of credit for the numerous hours she spent consulting me on statistical matters. The following researchers were gracious enough to make their data available: Benjamin Bergen, Travis Bradley, Ann Marie Delforge, Vineeta Chand, Ewa Dąbrowska, Jean-Marc Dewaele, Dan Dewey, Jennifer Bown, Wendy Baker, Rob Martinsen, Carrie Gold, Dennis Eggett, Martin Hilpert, Mary Johnson, John Grinstead, Frank Keller, Mirella Lapata, Azadeh Nemati, Steve Parker, Daniel Tight, Michael Taylor, Bodo Winter, and Sven Grawunder. I thank my wife, Silvia Gray, for putting up with my obsession with all things statistical while preparing this text.

Reprints throughout Courtesy of International Business Machines Corporation, © International Business Machines Corporation. SPSS® Inc. was acquired by IBM® in October, 2009.

IBM, the IBM logo, ibm.com, and SPSS are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at “IBM Copyright and trademark information” at www.ibm.com/legal/copytrade.shtml.

INTRODUCTION

Over the course of the last few decades, linguistics has seen a gradual shift away from armchair theorizing toward more data-oriented analysis. Subfields such as corpus linguistics, psycholinguistics, language acquisition, and sociolinguistics—all of which rely on quantitative data—have experienced a great deal of growth. Acoustic phonetic research, which used to require equipment costing the equivalent of the entire yearly salary of four assistant professors, can now be carried out on any personal computer thanks to programs such as PRAAT. Corpora designed for linguistic data-mining, such as the British National Corpus and Corpus of Contemporary American English contain hundreds of millions of words, while the Google Book corpus contains a number of corpora containing billions of words. With this proliferation of data, language researchers are obliged to find means to describe and summarize them, as well as to use them to test linguistic hypotheses.

While quantitative analyses are ever more frequent in the linguistic literature, many linguistics programs have not kept pace. Far too many linguists lack familiarity and training in statistics simply because it was not offered by or required in their programs. Of course, a plethora of introductory texts on statistics exist, and colleges and universities offer courses on the topic. Nevertheless, linguists often find that existing materials focus on solving problems in social and physical sciences. Many texts use topics such as birthrate trends in Asia or moisture content in soil samples to illustrate statistical principles, which often leaves linguists scratching their heads trying to understand how those procedures could apply to formant frequencies, reaction times, or pronoun deletion. What's more, language data lend themselves to statistical methods that are often not dealt with in introductory texts. My experience with statistics labs in a number of different universities is that those who staff the labs find it hard to wrap statistics around the kinds of data linguists use. For this reason, it is heartening to see a number of recent books on statistics written specifically for the field of linguistics (Baayen 2008, Cantos Gómez 2013, Gries 2013, Johnson 2008, Larson-Hall 2010, Rasinger 2008).

At this point, I would like the reader to imagine a person named Kim who has grown tired of fast food and wants to learn to cook. Kim shows up for class at a prestigious school of culinary arts. The first day is spent discussing the inner workings of industrial mixers and ovens. The next day Kim learns about different kinds of yeasts and their reproduction rates in environments that vary in their sugar content and temperature. The third day the class revolves around the chemical process of fermentation.

After a few weeks of this, Kim finally gets up the courage to admit to the instructor, "I just want to learn the steps to take to prepare a nice meal. Why do I need to know about how the ingredients are fabricated and what kitchen tools look like on the inside?" "Ah," replies the instructor, "this information makes you more knowledgeable and better able to appreciate the complexities behind the ingredients you will use in preparing *haute cuisine* and the machinery used in their preparation. You will be more empowered to understand their chemistry and mechanics." "But I just want to learn how to make a quiche," protests Kim, "I don't want to repair kitchen appliances or produce raw ingredients."

Ultimately, Kim leaves the class in despair and picks up a pizza on the way home In too many cases, I have observed statisticians getting so wrapped up in the mechanics of the math that goes on under the hood, so overzealous about teaching every intricacy and nuance of statistical theory, they overlook the fact that many researchers just want to learn to correctly apply statistical analysis to their data—not learn the math behind it. For many, an introductory book on statistics that contains pages and pages of formulas like this,[1]

$$l = 2c \cdot \ln\left[\left(\frac{S^2 - h^2}{4c^2}\right)^{\frac{1}{2}} + \left(\frac{S^2 - h^2}{4c^2} + 1\right)^{\frac{1}{2}}\right] \quad (4)$$

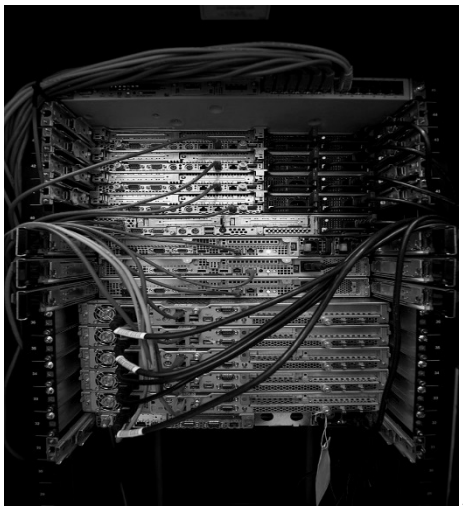
or instructions that look like this,

```
twosam <- function(y1, y2) {
    n1 <- length(y1); n2 <- length(y2)
    yb1 <- mean(y1);  yb2 <- mean(y2)
    s1 <- var(y1);  s2 <- var(y2)
    s <- ((n1-1)*s1 + (n2-1)*s2)/(n1+n2-2)
    tst <- (yb1 - yb2)/sqrt(s*(1/n1 + 1/n2))
    tst
    }
```

feels the same as most people do when they see this,[2]

[1] Image retrieved from http://commons.wikimedia.org/wiki/File:Formula10.png.

[2] Image retrieved from http://commons.wikimedia.org/wiki/File:Servers_in_a_Rack.jpg.



All of this can be overwhelming, especially when housed in a hefty 600-page book. Many of the statistically uninitiated will not get past the initial flip through and will simply toss the book aside in despair.

The present book is designed to address this problem. It is written for linguists who would like to have a practical, hands-on understanding of statistics so they can carry out quantitative studies of their own and so they can better understand the linguistic literature that includes such analysis. It is meant to be a truly basic introduction—not just in title, but in essence. The focus is on applying statistical recipes without cracking the machinery open to explore the mechanisms that underlie the mathematical engine. By adopting this approach, I hope to draw researchers into the quantitative world by avoiding what often prove to be stumbling blocks to the statistically uninitiated: mathematical formulas and total reliance on line command interfaces.

By using the previous cookbook analogy, by no means do I imply that I condone sloppy methodology. Too many basic introductory texts let crucial concepts, such as statistical assumptions, slide by with only a mention of how most statistics are robust in the face of violations (meaning that they still give valid results when the assumptions aren't met). Others merely intimate that there are other non-parametric tests readers can use, rather than showing readers how to actually use them. Therefore, the goal of this book is to provide recipes that are complete enough that, when followed, will result in sound results.

In order to achieve the aims of this book, I have designed it with a number of things in mind. First, the examples and data sets are mostly linguistic in nature, which should not only make them more accessible to the reader, but suggest ways in which other linguistic questions may be answered using the same statistical method. Second, the sheer number of different statistical analyses is somewhat overwhelming; for this reason, I have included the analyses that, in my opinion, are the most often used or the most appropriate to use with linguistic data. When there is more than one way to analyze a particular kind of data, I've mostly limited myself to explaining only one of them. Third, I have intentionally chosen a very informal writing style where I refer to the reader in the second person *you* and do not hide the identity of the author behind the passive voice, so I use the pronouns *I*, *we,* and *me* liberally. Fourth, for simplicity's sake, I have chosen to demonstrate how to carry out the different statistical analyses using software that has a graphical user interface (GUI), principally SPSS® version 20.[3] The differences between this version of SPSS and other newer versions (16 and higher, 2007 and later) are not large enough that the instructions and graphics in this book differ much from other recent versions. I understand that this excludes the program R (R Development Core Team 2011). I am well aware that in comparison to SPSS, R is more powerful, produces better graphics, and is free. However, its command line interface makes the learning curve much steeper in comparison to programs with graphical user interfaces,[4] and it is also quite intimidating for many computer users who are accustomed to point and click programs. Given the introductory nature of the present book, SPSS is a much gentler way of introducing people to statistics.

Of course, a primer of this sort is not meant to be a comprehensive answer-all; many topics are left uncovered. Instead, this primer serves as a gateway to the field. I trust that once a person grasps the basics of statistical analysis contained in this book, they will then be in an ideal position to perform more complex analyses, and begin to dig deeper, if so inclined.

[3] SPSS Inc. was acquired by IBM in October, 2009.

[4] There are a number of graphical interfaces for R (e.g., Deducer, R Commander, JGR, RKWard, Rattle) and choosing among them is problematic in and of itself. In contrast, SPSS has one well-documented GUI. In any event, R enthusiasts generally avoid using GUIs when writing texts.

CHAPTER ONE

GETTING TO KNOW SPSS

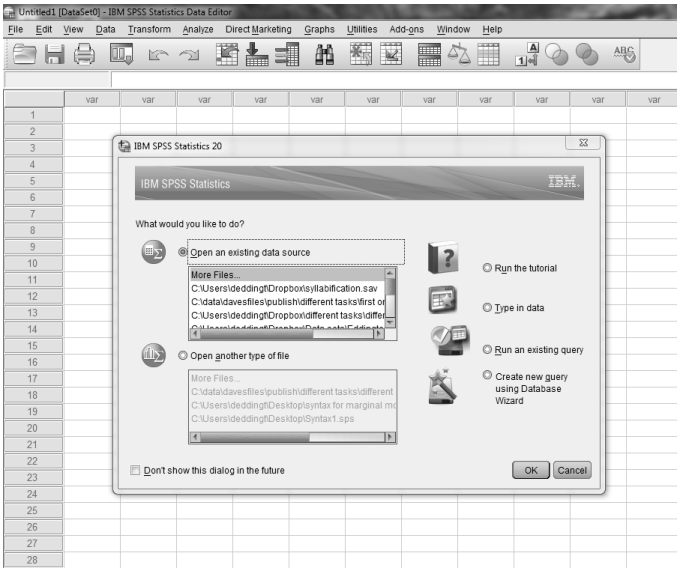
The best way to work through this book is with SPSS opened to the data that we are discussing on the screen. This allows you to replicate what you see on the pages in SPSS. Don't get freaked out if the windows on your computer look just a bit different from the screen shots in the book, or if they give a slightly different numeric output. These differences may be due to the fact that you have a different version of SPSS or run a different operating system on your computer.

To follow along, you will also need the data sets we'll use. They can be downloaded from http://linguistics.byu.edu/faculty/eddingtond/Data_sets. Once you have the data sets saved in a handy place on your computer, you are ready to begin.

1.1 Opening an Existing File in SPSS Software

Open the SPSS program by clicking on the icon (or doing whatever you do to start a new program). When you first run SPSS, it either opens two windows (Figure 1-1) or just the spreadsheet window (depending on the version). At this point, you can begin to enter data manually by checking *Type in data* on the right-hand side of the screen, which allows you to enter your data like a spreadsheet. However, hopefully you will already have your data in a file of some sort and won't need to do this.

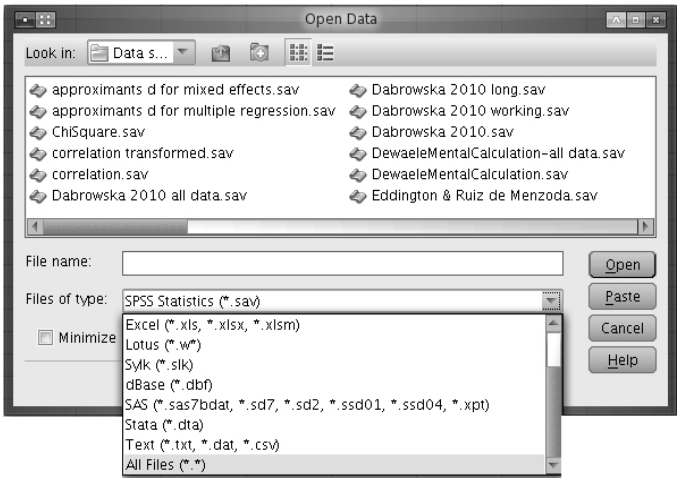
Figure 1-1. Opening an existing file in SPSS software.



Let's start by opening one of the data files that you have downloaded. (You have downloaded the data sets, right?) Do this by checking *Open an existing data source* (again, see Figure 1-1). Any files you have already opened in SPSS will appear in the box directly underneath this option (and you could click on the file name then *OK,* if that were the case). However, your file is probably not named in the box, so click on *More Files*, which brings up the *Open Data* dialog box (Figure 1-2). Use this window and the *Look in* drop-down menu to navigate to where you stored the files that accompany this text.

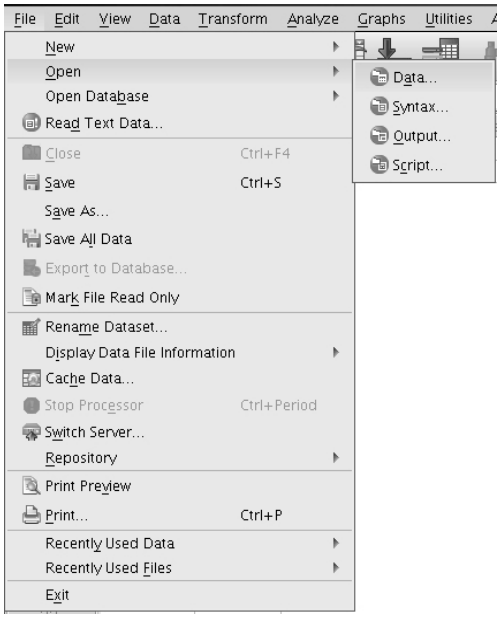
SPSS has a very narcissistic habit of only highlighting SPSS files that have the *.sav* extension. If the file you are looking for is of a different type, it won't show up until you pull down the *Files of type* drop-down menu (Figure 2). At the very bottom of that menu, you can choose to display *All Files.* As you can see in the drop-down menu, SPSS can import data from many different file formats.

Figure 1-2. *Files of Type* drop-down menu in the *Open Data* dialog box.



If you are already in SPSS and need to open a file, or if your version doesn't automatically open up the dialog box in Figure 1-1, you can access the *Open Data* dialog box by clicking on *File* in the top left-hand corner of the screen, choosing *Open*, and then choosing *Data* (Figure 1-3).

Figure 1-3. Opening a file from within SPSS.



To shorten things a bit, and to avoid saying "then choosing" ad nauseam, from now on I'll describe series of commands like this:

Click on *File* > *Open* > *Data*

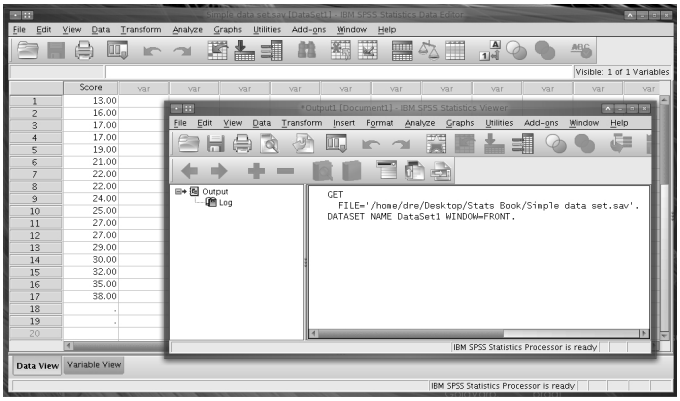
1.2 The *Statistics Viewer* and *Data Editor* Windows

When you open the data file both of SPSS's windows will open: the *IBM SPSS* window and the *IBM SPSS* window (see Figure 1-4; both windows will be open, although one may be on top of the other).

The *Statistics Viewer* window keeps track of all the commands you have given SPSS, and it is where you can see the results of your analyses. The *Data Editor* window works like a spreadsheet. This is where you put your data.

In the file, there are 17 scores. Notice that things like scores, age, gender, time, participant number, and other things like that go in columns. Rows are for cases, which in linguistics are usually participants or test items or responses in a study. So in this case, the 17 rows represent the test scores of 17 imaginary people (or "ideal speaker-hearers," if you prefer).

Figure 1-4. *Data Editor* and *Statistics Viewer* windows.

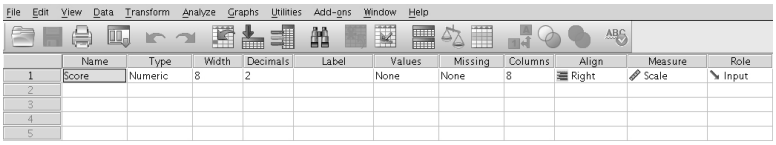


1.3 Specifying Information about the Data in the Window

The raw data are the heart of any analysis, but SPSS is not very bright on its own—it needs you to tell it more about what those numbers and letters mean before it can correctly deal with them. You can specify this information in the window. On the bottom left-hand corner there are two tabs: *Data View* and *Variable View*.

Click on the *Variable View* tab to see that window (Figure 1-5). The switch between windows can be confusing since the information that appears in columns in the *Data View* window appears in rows in the *Variable View* window.

Figure 1-5. The *Variable View* window.

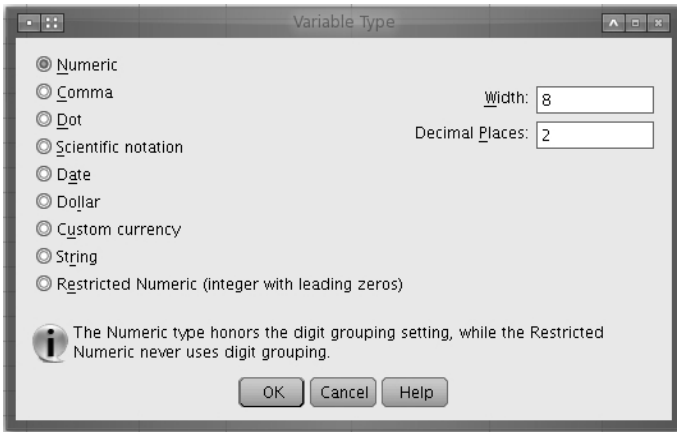


The name I gave the numbers is the bland-sounding *score*, which appears in the column. You can give your variables any name you want, except that you can't use spaces or characters like /, +, =, -, &, or you'll get an error message. The name you put here is the name that shows up in the column header in the *Data View* window.

HINT
The columns in the *Data View* window appear as rows in the *Variable View* window and vice versa.

In the column, you need to specify what kind of data appear in that column. When you click in this column, a drop-down button appears that, when clicked, brings up the *Variable Type* dialog box in Figure 1-6. Of all the possibilities in that box, we linguists generally only need to use two. If the data are numeric, choose *Numeric*; if they are letters or words, choose *String*. Here's a heads up: in a lot of cases (but not all), anything that you are going to use in an analysis, whether it is numeric or not, needs to be made numeric here. (Sorry—I didn't invent the program, I just work around it.) So, if you have three different languages, under the heading *Languages* you can call them 1, 2, 3 here. If you have Male, Female, call them 1, 2. You can sort out what the numbers mean in the column. Save for things like participant name, test word, test condition, or test sentence.

Figure 1-6. The *Variable Type* dialog box.



Sometimes when you have missing data you want to give it a special value, which is what the column is for. Remember that in the column, you specify how many characters are allowable. In the *Columns* column, you say how many of those characters will actually show up in the spreadsheet. *Align* lets you change whether the values in a column are left-aligned, right-aligned, or centered.

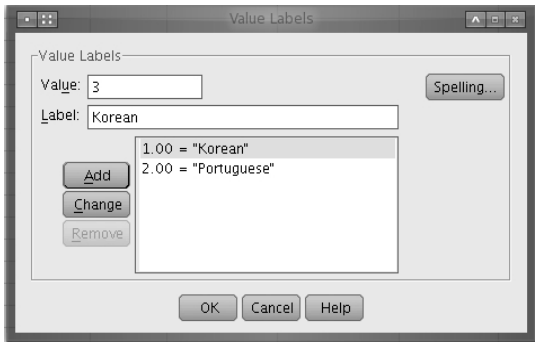
The *Measure* and columns aren't crucial to performing an analysis. They are places to help you keep track of what kind of data you have in each column and what they are used for. In *Measure* you can specify what kinds of data are in the column, which we'll discuss this in greater detail in Chapter 2. For now, just remember that is what I'll refer to as "continuous." Older versions of SPSS don't have a *Role* column at all. In newer versions, you can keep track of what role the variable you are defining plays in the analysis. For example, *Input* is an independent variable and *Target* is a dependent variable. We'll get to this later.

HINT
Sometimes it's helpful to have the same variables coded twice, once by name (e.g., *Gender* = male, female) and once by numbers (e.g., *GenderNumeric* = 0, 1). Some analyses allow string variables and others require only numerically coded data.

Also, in this dialog box, you can choose how many decimal places to display (*Decimal Places*) as well as the maximum number of characters that can fit in the column (*Width*). If you forget to specify this information here, notice that this same information can be entered in the *Width* and *Decimal* columns in the *Variable View* window.

The column is a place where you can describe in detail what the name you put in the *Name* column means. For example, if the value in the column is *NumYears,* which means "number of years living in target language country," put that long descriptor here since it is way too big for the *Name* column and has a bunch of spaces in it anyway that make it illegal in the *Name* column.

Figure 1-7. The *Value Labels* dialog box.

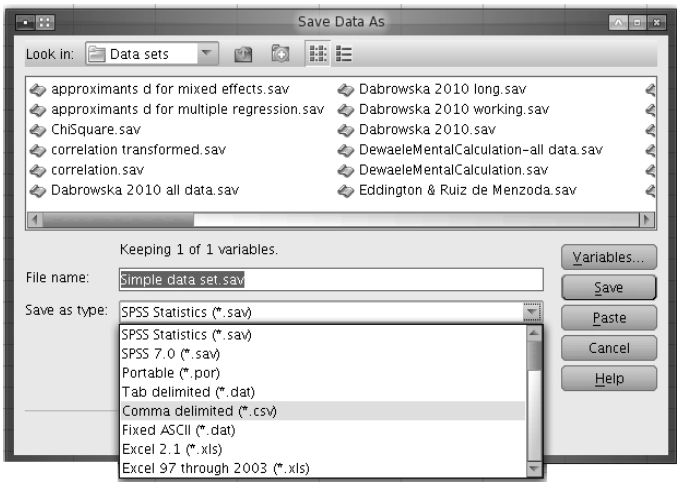


The *Values* column is where you tell SPSS exactly what the numbers in a variable like *Languages* actually mean. Clicking on a cell in the column opens the *Value Labels* dialog box (Figure 1-7), which is already partially filled out. Specify the number you want to define in the *Value* box, give it a name in the *Label* box, and press *Add.* SPSS has even provided a handy spelling check button here for your continuing statistical enjoyment.

1.4 Saving a File in SPSS

Click on *File* > *Save as* to open the *Save Data As* dialog box (Figure 1-8). Write the name you want to give the file in the *File name* box and select the format in the *Save as type* drop-down menu. There are many different formats that can be chosen. SPSS uses its own *.sav* format.

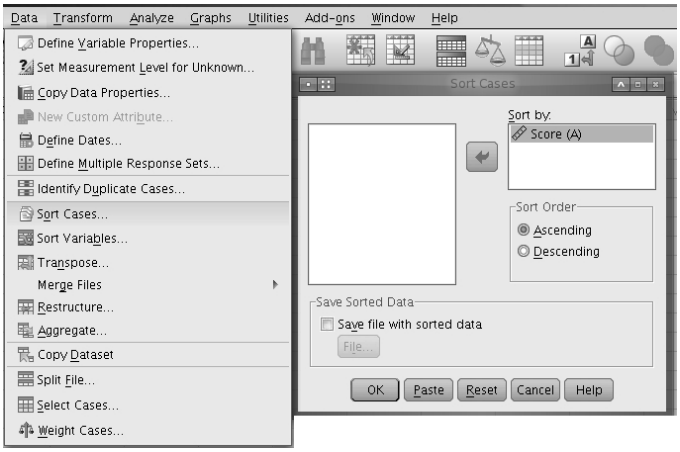
Figure 1-8. *Save Data As* dialog box.



1.5 Sorting the Data

Click on *Data* > *Sort Cases* and click on *Scores* in the box on the left. Move the term into the *Sort by* box by clicking on the arrow button between the boxes (Figure 1-9). Now choose whether you want to sort it in ascending or descending order, then click *OK*.

Figure 1-9. *Sort Cases* dialog box.



This is a fairly simple sort with only one column involved. It is possible to do more complex sorts. Remember that the first column of data that you specify takes priority over subsequent ones. So if you had *test score* and *gender* as columns and you put *gender* first in the *Sort by* field, you would get the first two columns in Table 1-1. If you put *test score* in before *gender*, the result would be the last two columns.

Table 1-1. Sorting gender and test score by different priorities.

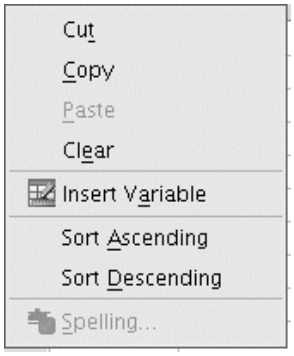
Gender First	
Gender	*Score*
F	89
F	79
F	76
F	61
M	90
M	87
M	87
M	42

Score First	
Gender	*Score*
M	90
F	89
M	87
M	87
F	79
F	76
F	61
M	42

1.6 Adding, Deleting, Copying, Pasting, and Cutting Rows and Columns

Go to the *Data View* tab. If you needed to add a column to the left of the *Score* column, you right-click on the title of the column (*Score*) and choose *Insert Variable* from the dialog box that appears (Figure 1-9). From this handy dialog box, you can also clear, cut, paste, or sort the contents of the column. To add a row, right-click on the row number on the left-hand side of the screen and choose *Insert cases*. The option to cut, clear, or copy the row also appears.

Figure 1-9. Insert, cut, copy, paste, clear, sort variable dialog box.



Of course, what I've shown you here is far from an exhaustive list of commands, but it should get you started. I certainly don't want to bore you only a few pages into the book, so I'll introduce other commands as they relate to the concepts in each chapter.

CHAPTER TWO

DESCRIPTIVE AND INFERENTIAL STATISTICS

2.1 Types of Data

The purpose of statistics is to organize, describe, summarize, and interpret data. It allows us to find trends, discover relationships between variables, and determine whether there are any causal relationships between them. Most importantly, statistics help us explore whether results can be extrapolated beyond the cases we consider, which allows us to test hypotheses about linguistic issues. However, to correctly order the data to be used in a statistical procedure as well as to choose which procedure to carry out, it is crucial to correctly classify the data, since not all data are of the same type.

2.1.1 Categorical data

In linguistics, much of the data we are interested in is CATEGORICAL.[1] This means that it can fit into named, unordered categories. Some examples of categorical data are

- Gender: male or female
- Country of origin: Korea, Canada, Brazil, or France
- Education: high school graduate or not
- Ethnicity: Hispanic, Caucasian, Asian, or Polynesian
- Childhood language background: monolingual, bilingual, or multilingual
- Prodrop: subject pronoun used with verb or subject pronoun not used with verb
- Language abilities of participant: native or nonnative
- Teaching method: total physical response, audiolingual, or grammar translation
- Word used for "large sandwich": *hoagie*, *subway, grinder,* or *po' boy*
- Pronunciation of /t/: [t^h], [ʔ], or [ɾ]

Notice that the number of things that are categorical may be counted, but you may not treat the category as if it were a number. For example, if you imposed a number system on the words people use for "large sandwich," such that 1 = *hoagie*, 2 = *subway*, 3 = *grinder*, and 4 = *po' boy*, the numbers make it seem like *subway* is somehow closer to *grinder* and further from *po' boy*. This, of course, makes no sense. It is impossible to plug categorical categories into mathematical formulas. These variables are merely named, not numbered. There are times, however, when you will need to identify categorical items with numerals, then specify what those numbers mean (see Chapter 1). However, the variables are always marked as categorical when they are used in an analysis.

2.1.2 Ordinal data

ORDINAL DATA are best exemplified by the order in which runners cross the finish line. The first-place runner may finish only a fraction of a second before of the second-place runner, yet the time between the second- and third-place runners could be several seconds. For ordinal data, the distance between each number is not relevant—only the order in which the numbers occur. Some examples of ordinal data in linguistics are

- The order in which children acquire certain morphemes.
- The way a test participant orders a series of five recordings of nonnatives from *most fluent* to *least fluent*.

Notice that you cannot subdivide ordinal data. It makes no sense to say that the runner from Ghana came in third and a half in the race.

Some measures involve choosing a point along a labeled scale. For example, in response to the statement, "The person in this recording has a strong foreign accent," you can choose one of the following: *strongly agree*, *agree*, *neither agree nor disagree*, *disagree*, *strongly disagree.* It is clear that *strongly agree* indicates a judgment of more foreign accent than *agree*, but does choosing *strongly agree* really mean twice as much accent as *agree*? In this case, the intervals between each label are very subjective and hard to quantify numerically, so it may be better to treat them

[1] Some other names for categorical variables are NOMINAL or FACTOR variables.

as ordinal rather than continuous (see next section), although some claim that it's okay to treat them as continuous (Carifio & Perla 2007), so I'll leave it up to your discretion. Because ordinal data have unequal (or unclear) intervals between each point, they are treated differently in statistical analyses from other numeric data that do have equal intervals.

2.1.3 Continuous data

CONTINUOUS DATA[2] are numeric data in which the distances between numbers are equal. For instance, a person who has lived in a foreign country for two years has lived twice as long there as a person who has only lived there one year. Examples of continuous data are

- Age
- Number of years of formal schooling
- Months spent living in a foreign country
- Time required to recognize a word during an experiment
- Frequency of a formant
- Duration of consonant closure
- Hours spent sending text messages

In contrast to ordinal data, subdividing continuous data results in interpretable outcomes: it is possible for a person to be 21.3 years old and to spend 5.4 hours a day writing text messages. Believe me; I know such a person.

2.2 Variables

Variables are simply characteristics that change from situation to situation, object to object, or person to person. A person's biographic information is a series of variables. Some variables are continuous: age, number of years of schooling past high school, number of years living at the present address, number of children. Others are categorical: gender, ethnicity, county of residence, marital status. Some variables are ordinal: ranking among high school class, birth order among siblings, order of preference among potential dates in your little black book.

2.2.1 Independent and dependent variables

In quantitative research, we are interested in the relationship between certain variables. Perhaps the best way to frame a research question is this: What is the influence of X on Y? X is called the INDEPENDENT VARIABLE.[3] It is what we think makes a difference in Y, or causes Y, or is related to Y, or changes Y somehow. The relationship of the independent variable to the DEPENDENT VARIABLE is what you really want to research. Y is the dependent variable. It is the thing that may be changed, related to, or influenced by X. This is what you measure to determine if the independent variable has any effect on it.

If a research project cannot be stated in terms of how X affects Y, it is difficult to determine how to apply a quantitative analysis to it. For some, the most difficult step in doing research is framing one's ideas in terms of variables. Below are a few examples of research questions and the variables they entail.

Idea 1: People seem to use *myself* as the nonreflexive object of a preposition rather than as a reflexive a lot more nowadays (e.g., *as for myself)*.
Quantified question 1: What is the effect of time (1950s, 1960s, etc.) on the use of *myself* as the nonreflexive object of a preposition?
Variables in 1: Time is a continuous independent variable, and number of uses of *myself* as the object of a preposition is a continuous dependent variable.

Idea 2: It seems that women always outnumber men in foreign language classes.
Quantified question 2: What is the effect of gender on enrollment in foreign language classes?
Variables in 2: Gender is the categorical independent variable, and number of students enrolled is the continuous dependent variable.

Idea 3: I wonder if daily consumption of greasy American-style fast food is likely to shorten my life—yes.

[2] For those who know that there are actually two kinds of continuous data, I have collapsed ratio and interval data into this category, since it is not relevant to what kind of statistical analysis is used. Continuous variables are also sometimes called COVARIATES.
[3] It is also known as the PREDICTOR VARIABLE, although some people make subtle differences between the two.

2.2.2 Control variables

One difficulty in carrying out research is that linguistic behavior is human behavior, which is always influenced by a myriad of variables, some of which we may be aware of and others that lurk silently in the dark, waiting to pounce on us and undermine our—mostly—carefully constructed experiment. If we know of a variable that may affect the results, we would want to control for its influence. Such variables are called CONTROL VARIABLES. Suppose you are interested in how textual genre influences the use of the alternative past tenses of *sneak* (*snuck* and *sneaked*). You also know or suspect that there may be differences between British and American uses, but that is not the focus of your study. You could control for the influence of variety of English simply by limiting your search to only British or American corpora. Another way to control for the effect of any such variables that you are aware of is by including them in the statistical analysis.

2.2.3 Confounding variables

Confounding variables are the most pesky, because we often do not know they exist, or we only find out about their existence after the data have been gathered and analyzed. Although a good study tries to control for as many variables as possible so that the only one that affects the results of the study is the independent variable, sometimes confounding variables creep in and sabotage the study. Imagine a study designed to determine whether words with many morphemes, such as *industrialize* and *postmodernist*, take longer to recognize than monomorphemic words, such as *cabbage* and *alligator*. The results suggest this is true because polymorphemic words take longer to recognize. However, as psycholinguists know, this study contains a number of confounding variables. Word length and word frequency are also at work. Shorter words and more frequent words are recognized more quickly. Since these variables are not controlled for, they confound the results of this study and make it impossible to determine the exact effect the number of morphemes has on response time.

2.3 Descriptive Statistics

If you were handed a list of numbers, merely looking them over wouldn't give you much of a sense of the story they tell. This is where descriptive statistics come in. They provide concise summaries of the numbers taken as a whole, rather than individually. We'll use the data in Table 2-1 to illustrate descriptive statistics. Later on I'll describe how to use SPSS to obtain these statistics. If you really want to see where they come from, skip forward to Table 2-3.

2.3.1 Central tendency: Mean

A spreadsheet containing raw data is something that only true geeks can easily wrap their heads around. Descriptive statistics were invented to help the rest of us, since they summarize the data into easily digestible numbers. For example, the first thing a student generally wants to know upon receiving his or her test score is what the average of the class is. This is because a test score by itself is not very meaningful. It needs to be considered in the context of the population it comes from, that is, the scores everyone else received. The average test score, more precisely referred to as the MEAN rather than the average, is a measure of central tendency. It lets you know what middle number the scores cluster around. Consider the scores in Table 2-1. If your score is one of the two 27s, and the mean on the test is 24.35, then you did better than the average student. However, if your score is 5, then look for another major. The mean is calculated by summing up all of the scores and dividing by the total number of scores.

Table 2-1. Distribution of 17 scores.

Test Scores	
13	
16	
17	
17	
19	
21	
22	
22	
24	← median (24)
25	← mean (24.35)
27	
27	
29	
30	
32	
35	
38	

2.3.2 Central tendency: Median

Now, consider the test scores again. Another measure of central tendency is the MEDIAN, which is the middle score (or the grassy area between traffic lanes for some people). If you order all of the scores from lowest to highest, the median is the middle one.

In this case the median is 24, which is not far from the mean of 24.35. If the mean and median are about the same, why are they both needed? To answer this, look back at the list of scores and assume that there are two additional scores: 62 and 69. It should be obvious that these scores fall extremely far from the rest of the scores. They are outliers. When you include these two scores and recalculate the mean, it moves from 24.35 to 28.65, which is now quite a bit off-center. The median, on the other hand, only increases from 24 to 25. In other words, when there are outliers in the data, they affect the mean and make it a less accurate measure of central tendency. That is why the median is better when there are extreme scores. Medians are also used to report ordinal data. For example, imagine that you graduated tenth in your class—is that good? Well, if it was a small graduating class with a median of 10, not really. You are right in the middle of the class. However, if the median is 40, you are starting to look stellar.

2.3.3 Central tendency: Mode

Another measure of central tendency, that is much less often used, is the MODE, which is simply the most often occurring score. The above data actually has three modes, 17, 22, and 27, which are bolded in the data set in Table 2-1. Means and medians make sense for continuous data, but what gives us a sense of the middle if the data are categorical? In this case the mode is the measure of central tendency to use. Table 2-2 shows the number of students from different countries in an ESL class.

Table 2-2. Number of students in a class by country of origin.

Country	Number of students
Korea	6
Spain	3
Jordan	1
Mexico	4
China	4
Japan	2
Uruguay	1

For categorical data, the mode as a number makes little sense; you couldn't say that the mode country is 3.75. Instead it is reported as a category, so the mode for these data is Korea because that is the country that more students come from.

2.3.4 Dispersion

Knowing where the middle of the data falls is important but by itself is actually not very useful. Another crucial piece of data is how spread out the scores are, known as the score's DISPERSION. If you received a score of 27 on a test whose mean was 24.35, you may feel smug because you scored above average, but what if the highest score on the test were 50 and the lowest, 5? That look on your face melts away as you realize that the scores are more spread out or disperse. On the other hand, if the highest score were 30 and the lowest 17, your 27 is a phenomenal score because the scores are less disperse and cluster closer to the mean, which allows you to retain your air of self-importance. A simple measure of dispersion is the RANGE, which is simply the highest minus the lowest score. In the first case, the range is 45 (50 − 5), whereas the range is only 13 (30 − 17) in the second.

2.4 Using SPSS to Calculate Descriptive Statistics

We've been talking about the data in Table 2-1 and the statistics that describe them. Let's go over the steps of how to get SPSS to calculate them. I warn you that this produces a boatload of tables and graphs. I'll explain each one in time. As far as the tables are concerned, they contain an overwhelming amount of data. Be patient, and I will get to them and point out what to focus on. I include them all here so that you can see all of the output of the descriptive statistics that SPSS produces when you follow these instructions. I'll refer you back to some of these tables in later chapters.

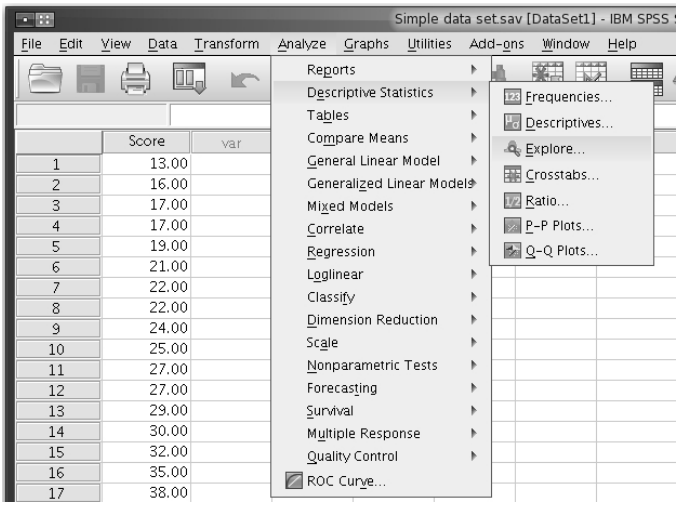
1. Download the file *Simple data set.sav* from the text website. Open SPSS > Check *Open existing data source* > *More files* > *OK* and navigate to where you downloaded the file > *OK*.

Remember that SPSS will automatically highlight only the *.sav*-type files unless you pull down the *Files of type* drop-down menu. At the very bottom of that menu you can choose to display *All Files.*

At this point, complete the following steps to generate the descriptive statistics and graphs used in this chapter.

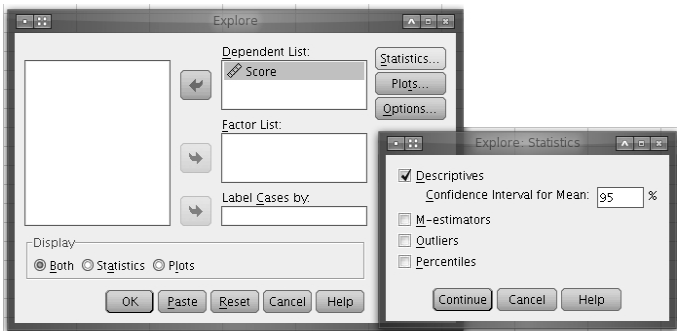
2. Click on *Analyze* > *Descriptive Statistics* > *Explore* (Figure 2-1).

Figure 2-1. Descriptive statistics using *Explore*.



3. Click on the variable name > *Score*, and move it to the *Dependent List* by clicking on the arrow between the boxes. Click on *Statistics* > check the box next to *Descriptives* > *Continue* (Figure 2-2).

Figure 2-2. *Explore* dialog box.



4. Click on *Plots* > check the boxes next to *Histogram* and *Normality plots with tests* > *Continue* > *OK* (Figure 2-3).

All of these tables and figures will appear in your *Statistics Viewer* window, and you may need to scroll down to see them. This generates the histogram in Figure 2-6, the boxplot in Figure 2-7, and the Q-Q plot in Figure 2-12. The descriptive statistics that are produced are given in Tables 2-3 and 2-4. These tables and figures are referred to later on in this chapter.

Figure 2-3. *Explore Plots* dialog box.

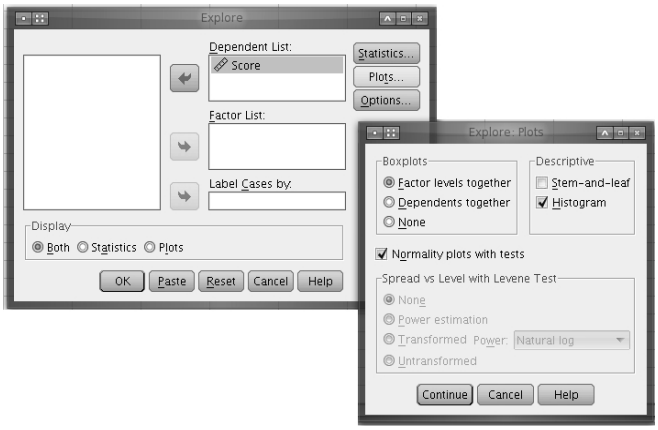


Table 2-3. Descriptive statistics for data in Table 2-1.

Descriptives				
			Statistic	**Std. Error**
Score	Mean		24.3529	1.69762
	95% Confidence Interval for Mean	**Lower Bound**	20.7541	
		Upper Bound	27.9517	
	5% Trimmed Mean		24.2255	
	Median		24.0000	
	Variance		48.993	
	Std. Deviation		6.99947	
	Minimum		13.00	
	Maximum		38.00	
	Range		25.00	
	Interquartile Range		11.50	
	Skew		.301	.550
	Kurtosis		−.589	1.063

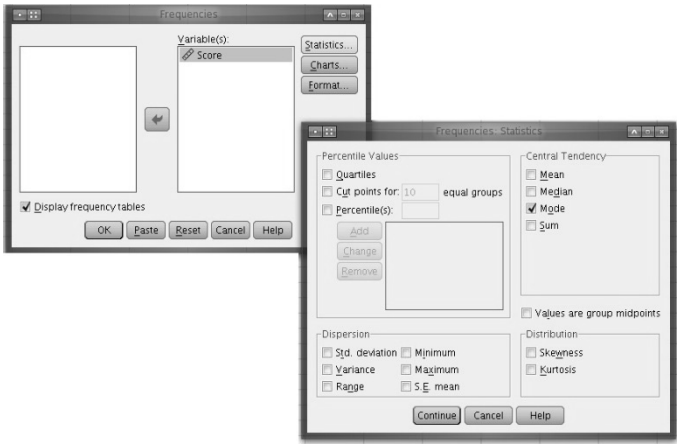
Table 2-4. Tests of Normality for data in Table 2-1.

Tests of Normality						
	Kolmogorov-Smirnov[a]			**Shapiro-Wilk**		
	Statistic	df	Sig.	Statistic	df	Sig.
Score	.102	17	.200*	.978	17	.939

*. This is a lower bound of the true significance.
a. Lilliefors Significance Correction

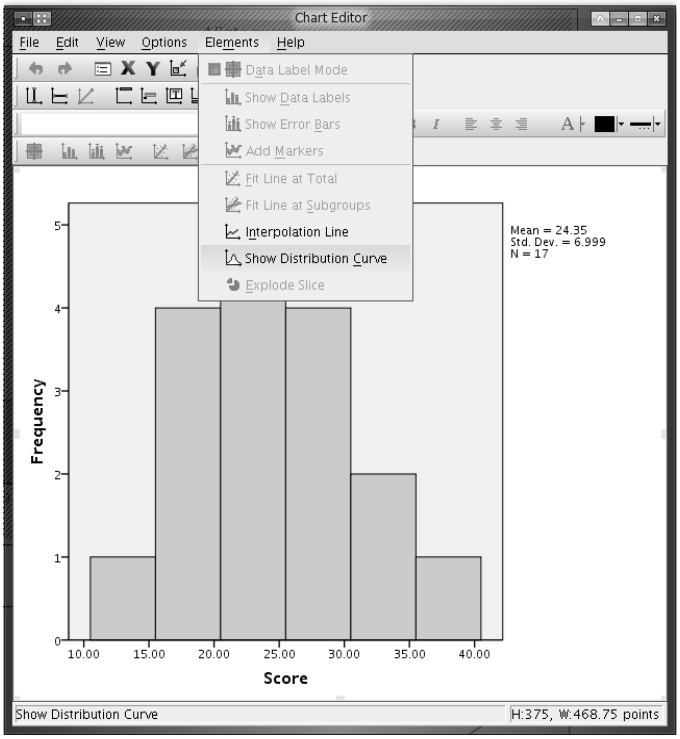
To get the mode, which is not included in Table 2-3 click on *Analyze* > *Descriptive Statistics* > *Frequencies*. Move *Score* into the *Variables* box > click on the *Statistics* button > check *Mode* box > *Continue* > *OK* (Figure 2-4).

Figure 2-4. Obtaining the mode.



In order to superimpose the normal distribution curve on the histogram in Figure 2-6, double-click on the histogram itself, which appears in the *Statistics Viewer* window. (The histogram looks pretty much like what you see in Figure 2-5. You will have to scroll down to see it.) Then in the *Chart Editor* choose *Elements* > *Show Distribution Curve*. Close the *Properties* and *Chart Editor* windows.

Figure 2-5. Superimposing the distribution curve.

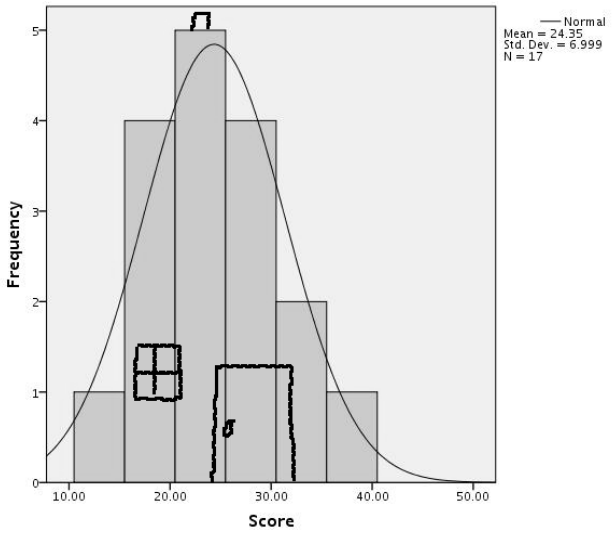


2.5 Visualizing the Data

2.5.1 Histogram

One way to visualize dispersion is with a HISTOGRAM. Figure 2-6 is a histogram of the scores we've been discussing. The bars indicate how many scores there are with a particular value or set of close values. For example, the leftmost bar contains the single score of 13. The second bar includes the four scores 16, 17, 17, and 19, and so on. The range for these data is 38 − 13 = 25. It should be clear that the central part of the distribution of these scores is about 24, where the mean and median indicate that it is. I will explain the curve in the histogram later.

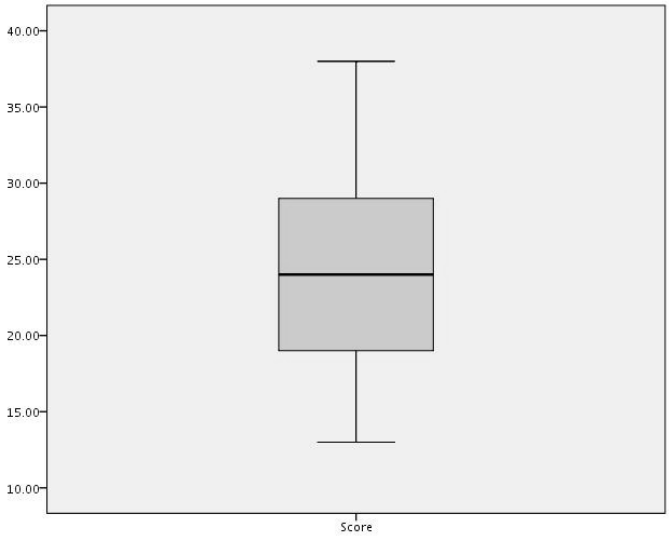
Figure 2-6. Histogram of data in Table 1-1.



2.5.2 Boxplot

Another way to visualize the data is with a BOXPLOT, as in Figure 2-7. One way of conceptualizing the relationship between the histogram in Figure 2-6 and the boxplot in Figure 2-7 is to imagine the histogram as a new age–type building seen from street level. The door and window should give you the idea. The boxplot is more analogous to viewing the same building from a helicopter hovering directly above the building. The dark line in the middle of the box represents the median, which is usually the tallest bar on the histogram. In the histogram, the median is where the chimney is. The box itself incorporates the data that are in the twenty-fifth to seventy-fifth percentiles. This means that only about 25% of the data fall above the box and about 25% below, so the box shows the middle 50% of the scores. The whiskers are the lines that extend above and below the box. On the histogram the ends of the whiskers indicate more or less where the curve comes closest to touching the street. They extend to lines that mark the largest and smallest scores as long as those are not farther away than one and a half times the length of the box. Scores beyond the ends of the whiskers are outliers because of their extreme values. There are no possible outliers in Figure 2-7, but if there were, they would appear as dots.

Figure 2-7. Boxplot of data in Table 2-1.



2.6 Normal Distribution

The curve that is superimposed on the histogram in Figure 2-6 shows what the data would look like if they followed a NORMAL DISTRIBUTION, which is often referred to as a bell curve. The much dreaded but often loved practice of grading on the curve entails forcing the test scores into a bell curve regardless of what their actual distribution is. In a normal distribution, the mean, median, and mode are identical. They all appear in the center of the