CHAPTER

1

Introduction

Performing a statistical analysis is a little like jumping off the high board into a swimming pool. You may be a little unsure the first few times, but the more you do it, the easier it becomes. Although this book doesn't help with swimming, it does provide you with information that will help you make those initial jumps into data analysis, and we hope the more you do it, the easier it will become. With over 60 combined years (egad!) of consulting and teaching experience behind us, our goal is to help students and researchers jump into the waters of statistical data analysis with confidence.

IBM SPSS by Example is a practical handbook that "cuts to the chase" and explains the when, where, and how of statistical data analysis as it is used for real-world decision making in a wide variety of disciplines. It is designed to assist students and data analysts who have general statistical knowledge to apply the proper statistical procedure to their data and reporting results in a professional manner consistent with commonly accepted practice. Each upcoming chapter discusses the following aspects of performing a statistical analysis and interpreting your experimental data:

- How to make sure you are using an appropriate application of the statistical procedure
- What design considerations you should consider when using a particular statistical procedure
- An explanation of the hypotheses tested by the procedure
- A description of tips and caveats you should know about the procedure
- An example (or two) illustrating the use of the procedure on a data set using step-by step directions on how to perform the analysis in SPSS.
- How to report the analysis results using standard American Psychological Association (APA) and Modern Language Association (MLA) compatible formats (APA, 2013; Gibaldi, 2006)

Before moving on to chapters that discuss specific statistical procedures, the next few sections in this chapter contain general information that pertains to the data analysis process. We cover this information here in part, so it will not have to be repeated individually for later

analyses. We encourage you to review the information in this chapter before moving on to the subsequent chapters.

Getting the Most Out of *IBM SPSS by Example*

The primary purpose of *IBM SPSS by Example* is to provide you with information about how to use and understand the statistical data analysis process. The analysis topics covered in the book are as follows:

- *Chapter 2: Describing and Examining Data*. Explains how to use descriptive statistics to understand and report information about your data.
- *Chapter 3: Creating and Using Graphs*. Explains how to use the SPSS Chart Builder, Graphboard Template Chooser, and Legacy graphs to describe your data.
- *Chapter 4: Comparing One or Two Means Using the t-Test*. Explains the one-sample *t*-test, two-sample *t*-test, paired *t*-test, and appropriate confidence intervals.
- *Chapter 5: Correlation and Regression*. Explains correlation and simple linear regression with a brief discussion of multiple linear regression.
- *Chapter 6: Analysis of Categorical Data*. Explains methods that are applicable to count or categorical data, including contingency table analysis, measures of risk (including relative risk), odds ratios, and goodness of fit.
- *Chapter 7: Analysis of Variance and Covariance*. Explains several methods of comparing means, including one-way analysis of variance (ANOVA), two-way ANOVA, repeated-measures ANOVA, and analysis of covariance (ANCOVA).
- *Chapter 8: Nonparametric Analysis Procedures*. Explains nonparametric statistical procedures, including Spearman's correlation, sign test, the Mann-Whitney *U*, Kruskal-Wallis, and Friedman's test.
- *Chapter 9: Logistic Regression*. Explains logistic regression analyses, including the cases of single or multiple independent variables, variable selection, and evaluation of the model.
- *Chapter 10: Factor Analysis*. Explains how to use Factor Analysis to examine data sets and identify underlying components of the information.

Along with each analysis in these chapters, we include examples along with "step-by-step" instructions describing how to perform the calculations using IBM SPSS. Additional

information that may be helpful to you in analyzing the example data sets and selecting an appropriate analysis for your data is included in the following appendices:

- *Appendix A: A Brief Tutorial for Using SPSS for Windows*. This tutorial gets you started with the essential information needed to work through the examples in this book. We recommend that if your SPSS is rusty, if you have limited experience using SPSS, or if you are new to SPSS, you should go through the examples in this appendix before working the examples in the book. For the more adventurous, it also includes a tutorial for using the SPSS Code Language that describes how you can use the SPSS Syntax Editor to modify existing analyses.
- *Appendix B: Choosing the Right Procedure to Use*. This appendix includes a decision chart that can help you decide which statistical procedure is appropriate to address your research question.

The remainder of this chapter contains material that we believe is important for understanding the examples contained in this book. We know you are in a hurry, maybe faced with a deadline, and anxious to get to your analysis. Take a breath. Relax. Slow down. This chapter provides you with basic refresher information about research methods that can help you successfully navigate your way through the analysis process . . . if you take a few minutes to read the rest of the chapter. Doing so may save you hours of frustration down the road and provide you with the ability to do a quicker and better job with your analysis. The remaining topics covered in the rest of this chapter are as follows:

- A Brief Overview of the Statistical Process
- Understanding Hypothesis Testing, Power, and Sample Size
- Understanding the *p*-Value
- Planning a Successful Analysis
- Guidelines for Creating Data Sets
- Preparing Excel Data for Import
- Guidelines for Reporting Results
- Downloading Sample SPSS Data Files
- Opening Data Files for Examples

A Brief Overview of the Statistical Process

Perhaps you are currently taking a statistics course or you struggled through a statistics course in the past and the concepts you once knew are a bit fuzzy. In this review, we remind

you of the issues that typically motivate the use of statistical data analysis and illustrate the types of analyses that are most commonly used to describe data or make a decision based on observed data. Even if you have studied these concepts before, you might learn something new or gain some insights that hadn't occurred to you previously. In either case, we hope this review is helpful.

Most analyses can be categorized into one of these types:

- Description
- Comparison
- Association/correlation

Using Descriptive Statistics

In today's world, information is being accumulated at a dizzying pace that makes it impossible for anyone to comprehend it all. However, computer software, such as SPSS, enables us to summarize vast quantities of this information into numbers and graphs that can help us understand trends, make decisions, and predict future behavior. The challenge for the data analyst is to interpret this tsunami of information in some logical and practical manner based on accepted statistical practice.

For example, suppose you have been funded by a government agency to evaluate the operation of two charity-sponsored counseling centers. As a part of the analysis, a satisfaction survey is given to 109 clients over a period of 1 month and measured on a scale of 1 to 100. In order to describe the results of the survey, you wouldn't want to present a list of raw results (109 scores). Instead, it would be more informative to report several summaries, such as the following:

Average satisfaction score: 80.3 (on a scale of 0 to 100)

Lowest score: 58.6

Highest score: 94.1

This descriptive information gives you an idea about the average level of satisfaction and something about the variability of scores.

Using Comparative Statistics

Since there are two counseling center locations, your research group might be interested in knowing if there is a difference in level of satisfaction among clients at the two locations. This could be important in deciding which center receives additional funding. Suppose you have the following summary data grouped by location:

Average satisfaction score at the uptown location was 82.4 (based on 54 client scores).

Average satisfaction score at the downtown location was 78.5 (based on 55 client scores).

Assuming that the clients are representative at each location, you have some evidence to make a decision about which center is more effective in terms of satisfaction score. Your data suggest that the uptown location may do a better job as far as the satisfaction score is concerned since the score for uptown is 3.9 points higher than the score for the downtown location. However, what if the average scores were only 1 point apart? Or 10 points apart? What level of difference would it take for you to conclude that the average score for one location was significantly higher than for the other? Could the difference in scores be due to random fluctuation? If you did the survey again during some other time period, is there a reasonable chance that the downtown location would produce a better score? These questions are addressed with a properly designed and executed statistical analysis.

Using Correlational Statistics

To learn more about your survey results, you could examine your data in another way. Ignoring for a moment the location of the center, you may want to compare the relationship between educational level of clients and satisfaction scores. The variables *survey scores* and *years of schooling* are plotted on a scatterplot in Figure 1.1, and a measure of how they are related is summarized in a number called the correlation coefficient, which is found to be $r = 0.37$. From this measure of association, you have evidence that suggests there is a mild positive linear relationship between years of schooling and satisfaction score. There is a tendency for clients with a higher education to have a higher satisfaction score. (Correlation is discussed in more detail in Chapter 5: Correlation and Regression.)

In each of these example analyses, the raw data are summarized into summary statistics or a graph that allows you to discover important information about the data and to provide the basis for making informed decisions. *IBM SPSS by Example* provides you with the information needed to use these and other types of statistical procedures and to interpret the results.

Understanding Hypothesis Testing, Power, and Sample Size

To properly interpret a statistical analysis, you must understand the concept of hypothesis testing. Otherwise, most SPSS output will be so much gibberish. This brief discussion is designed to refresh your memory about these concepts.

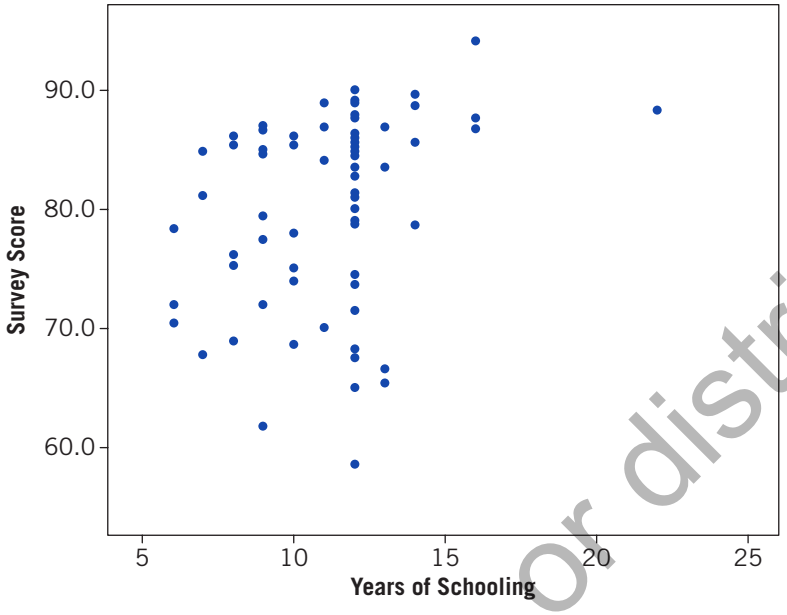

Figure 1.1 Scatterplot of Schooling by Survey Score

Many people have likened hypothesis testing to a criminal jury trial. In the U.S. judicial system, we make an initial assumption that the defendant is innocent (not guilty). Evidence is then presented to show guilt. If there is a preponderance of evidence of the defendant's guilt, you should conclude that the defendant is indeed guilty (you reject innocence). In the same way, a statistical analysis is based on a "null" hypothesis (labeled H_0) that there is "no effect" (e.g., no treatment differences). In research terms, the null hypothesis will typically be a statement such as the following: There is no difference in group means, no linear association between two variables, no difference in distributions, and so on.

An experiment is designed to determine whether evidence refutes the null hypothesis. If your evidence (research results) indicates that what you observed was extreme enough, then you would conclude that you have "significant" evidence to reject the null hypothesis. However, if you do not gather sufficient evidence to reject H_0, this does not prove that the null hypothesis is true, only that we did not have enough evidence to "prove the case." Back to the criminal trial example, failure to prove guilt beyond a reasonable doubt does not prove the "innocence" of the defendant. In fact, we use the terminology "We find the defendant not guilty" as a subtle way to distinguish from a verdict of innocence.

In general, null and alternative hypothesis are of the following form:

- A "null hypothesis" (H_0) is the hypotheses of "no effect" or "no differences" (i.e., the observed differences are only due to chance variation).
- An alternative hypothesis (H_a) states that the null hypothesis is false and that the observed differences, relationships, etc. are real.

In the following chapters, the null and alternative hypotheses related to each statistical test will be presented. They appear in the following form:

H_0: $\mu_1 = \mu_2$ (the population means of the two groups are the same).

H_a: $\mu_1 \neq \mu_2$ (the population means of the two groups are different).

These particular hypotheses are for a two-sample *t*-test as described in Chapter 4: Comparing One or Two Means Using the *t*-Test. In most cases, we will present the hypotheses in both a mathematical form (such as $\mu_1 = \mu_2$) and in words.

The alternative hypothesis is usually what the investigator wants to show or suspects is true. The alternative in the example above is called a two-tailed alternative (also called a two-sided alternative). That is, we would reject H_0 if there is sufficient evidence that the null is not true. For a one-tailed alternative (e.g., H_a: $\mu_1 > \mu_2$), we would reject H_0 only if the evidence against H_0 tends to support H_a. Further discussion of one- and two-tailed alternatives will be given when appropriate for the discussion of various tests in future chapters.

In hypothesis testing, two types of errors can occur, as illustrated in Table 1.1. The top classification is the "truth" that you do not know. The left categories are your decisions. For example, if you reject H_0 when it is false, you've made a correct decision. However, if you reject H_0 when it is true, you've made a "Type I error." Notice that of the four possible outcomes summarized in the table, two are errors.

The Type I error is controlled by your choice of a decision-making criterion, called alpha (α) or the level of significance. It is commonly set small, at 0.05, which means you are willing to risk making a Type I error 5% of the time, or 1 in 20 times.

If H_0 is false and you do not reject H_0, you commit a Type II error. The probability of committing a Type II error is called beta (β). The power of the test is defined to be one minus β. When a test has low power, it increases the chance of making a Type II error (i.e., failing to reject H_0 when it is actually false). Looking at it the other way, the higher the "power," the better your chance of rejecting H_0 when it is false—the better your chance of finding a

difference when it in fact exists. Other factors that we have not discussed here that may be related to power include sample size, effect size (or size of the smallest important difference), whether your test is one or two-tailed, your selection of the alpha-level (α), correlations among samples, and type of distribution (e.g., normal), etc.

An important point is that there are many ways in which a null hypothesis can be "not true." For example, if the null hypothesis is that there is no difference in two population means (measured in inches), then, for example, this hypothesis is "not true" if the actual difference between the two means is 1 in., 5 in., or 50 in. It may be very difficult to develop a test for which we are able to detect a difference in population means of 1 inch. In fact, such a difference may be of no practical importance. On the other hand, it will likely be the case that a true difference of 50 in. may be very easy to detect. That is, if the true difference is 50 in., the power of the test will be large.

Another important point is that for any given level of significance (α), power can be increased by increasing the sample size. Thus, sample size should be a consideration when embarking on an experiment. Many negative (nonsignificant) studies reported in the literature are the result of inadequate sample size resulting in poor power (Friedman, Chalmers, Smith, & Kuebler, 1978). Therefore, the process of selecting a sample size for your analysis should begin early in your study. To follow with this example, the experimenter should determine the level of difference it is desirable to detect and then select a sample size that will detect this difference with an acceptable power (say, at least 0.80). Often, a pilot study will be undertaken to help determine the necessary sample size. SPSS offers a separate program called SamplePower that allows you to calculate a sample size for a given power or range of powers you select. Other commercial programs (PASS, nQuery, and SAS) are also available for these purposes. Or consult your friendly local statistician for help. For more concerning hypothesis testing, see a standard statistical text such as Moore and McCabe (2012). For a good discussion of power and sample size, see Keppel and Wickens (2004). References for effect size are Nakagawa and Cuthill (2007) and Ferguson (2009).

Table 1.1 Hypothesis Test Decisions

	Truth	
Our Decision	**H_0 True**	**H_0 False**
Reject H_0	Type I error (α)	Correct decision ($1 - \beta$ or power)
Do not reject H_0	Correct decision ($1 - \alpha$)	Type II error (β)

Understanding the *p*-Value

The "evidence" used to reject a null hypotheses is summarized in a probability called a *p*-value. The *p*-value is the probability of obtaining results as extreme or more extreme than the ones observed given that the null hypothesis is true. Thus, the smaller the *p*-value, the more evidence you have to reject the null hypothesis.

When your rejection criterion, α, is set at 0.05, then if your *p*-value for that test is 0.05 or less, you reject H_0. All of the examples illustrating statistical tests in *IBM SPSS by Example* use the criterion that a *p*-value less than 0.05 indicates that the null hypothesis should be rejected.

However, don't base your entire decision-making criteria on the *p*-value. For example, suppose two sample means for systolic blood pressure (SBP) differ by 1 point and are found to be statistically significantly different (i.e., $p < 0.05$). This could occur if the sample sizes are large, but such a finding may have no practical or therapeutic importance, even though the results are statistically significant. On the other hand, an observed difference in mean SBP of 20mm Hg based on small sample sizes may not be statistically significant (i.e., $p > 0.05$). However, such a finding may be of sufficient practical importance that this (nonsignificant) result may indicate the need for further investigation with larger sample sizes to increase the power to the extent that you would have a good chance of detecting a difference of 20mm Hg if it really exists. The point here is that the *p*-value is a valuable decision-making tool, but it should not be the only criterion you use to judge the results of your research. Another approach to interpreting results is to make use of *effect size* reporting which provides wording you can use to express the magnitude of a difference.

A word of warning: If you perform multiple statistical tests within the same analysis, you should adjust your α level for individual tests to protect your overall Type I error rate. For example, if 10 independent statistical tests are reported for the same analysis (such as in a table comparing baseline values between two groups), each conducted at the 0.05 significance level, there is about a 40% chance that one or more significant differences would be found even if there are no actual differences. That should be unacceptable to you—and is usually unacceptable to journal reviewers. The proper response to this problem is to adjust *p*-values in multiple tests using a standard technique such as the Bonferroni correction. (SPSS offers several other techniques for *p*-value adjustments. For simplicity, we're only mentioning the Bonferroni technique here.) To perform this simple adjustment, divide your rejection criterion value (α) by the number of tests performed. For example, if you are testing at the $\alpha = 0.05$ level and 10 tests are performed, then your rejection criterion for each test should be $0.05/10 = 0.005$ in order to maintain your 0.05 overall Type I error rate (Miller, 1981). To report these results in your paper, use wording such as "$p < 0.005$ was considered statistically significant for baseline comparisons according to a Bonferroni correction. . . ."

Planning a Successful Analysis

Statistical data analysis begins with planning. Entire university courses are devoted to properly designing experiments. An improperly designed experiment can make data analysis a nightmare. Therefore, it is to the researcher's advantage to spend some up-front time considering how an experiment will be analyzed before collecting the data. Although this book cannot cover all the aspects of good experimental planning, a few important considerations are provided here. And although there are times when plans need to be adjusted once you know more about your experiment, having a solid up-front plan will provide you with the best process to deal with any subsequent issues.

SIDEBAR

All of the examples illustrating statistical tests in *IBM SPSS by Example* use the criterion that a *p*-value less than 0.05 indicates that the null hypothesis should be rejected, unless a multiple-test correction is made.

Formulate a Testable Research Question (Hypothesis)

You should formulate a testable research question (hypothesis) before collecting your data and formulating your research question in a way that is statistically testable. For example, you might test the null hypothesis that there is no difference in satisfaction scores from the two counseling centers in the previous example. You "test" this assumption by gathering data and determining if there is enough information to cast sufficient doubt on your null hypothesis. If there is such evidence, then you may reject the null hypothesis in favor of the alternative (one location has a better satisfaction score than the other, or they are different).

Collect Data Appropriate to Testing Your Hypotheses

Consider the types of variables you will need to answer your research question:

An outcome variable. (Sometimes also called the dependent or response variable.) The outcome variable measures the characteristic that you want to test or describe in some way. It could be some measure such as death, sales amounts, growth rate, test score, time to recovery, and so on.

Predictor variable(s). (Sometimes called independent or explanatory variables, or factors.) The predictor variables are often manipulated by the experimenter (e.g., level or dosage, color of package, type of treatment), although they may also be observed (such as cigarette smoking, blood pressure, gender, amount of rainfall).

For correlational studies. If you are performing a correlational study (examining the association between variables), you may not have a specific outcome variable. Keep in mind, however, that a correlational study by itself cannot be used to conclude cause and effect.

Scales of measurement. The method you use to measure an observation affects the type of analysis that may be performed. Table 1.2 describes three measurement types used by SPSS. As you design your study, keep in mind these general ways of measuring data.

As various statistical analyses are discussed in this text, reference will be made to the measurement types appropriate for the analysis.

SIDEBAR
SPSS makes a distinction between **measurement types** described in Table 1.2 and **data types**, which refer to the way actual values are displayed in the data set. Data types include numeric, comma, dot, scientific notation, date, dollar, custom currency, string, and restricted number. These data types are described more thoroughly in Appendix A: A Brief Tutorial for Using SPSS for Windows.

Decide on the Type of Analysis Appropriate to Test Your Hypothesis

Do you need a descriptive, comparative, or association/correlation analysis? See Appendix B: Choosing the Right Procedure to Use for help in deciding which type of data analysis to use for testing your hypotheses. Wilkinson and the Task Force on Statistical Inference (1999) state,

> The enormous variety of modern quantitative methods leaves researchers with the nontrivial task of matching analysis and design to the research question. Although complex designs and state-of-the art methods are sometimes necessary to address research questions effectively, simpler classical approaches often can provide elegant and sufficient answers to important questions. Do not choose an analytic method to impress your readers or to deflect criticism. If the assumptions and strength of a simpler method are reasonable for your data and research problem, use it. (p. 598)

Table 1.2 SPSS Measurement Types

SPSS Measurement Type	Interpretation	Other Notes
Scale	Scale data has an order and a metric so that a comparison of distance between numbers is appropriate. Examples are measurements such as height, age, income in dollars, etc.	SPSS scale variables include both interval data (data without a fixed origin) and ratio data (data with a fixed origin). We will sometimes refer to scale variables as quantitative.
Categorical/ Nominal	Nominal categorical variables do not have any implied order (e.g., gender, race, eye color, etc.).	Categorical results are generally reported as counts—that is, how many of each category is observed. We will also refer to categorical variables as qualitative.
Categorical/ Ordinal	Ordinal categorical variables have an implied order (e.g., level of difficulty, easy, hard; or box size, small, medium, large; or cancer grade, 1, 2, 3, 4).	Although cancer grades are numbers, they cannot be assumed to be an equal distance apart, and so cannot be considered to conform to a metric (see scale).

A practical interpretation of this is that you need to be able to understand and defend whatever analysis technique you use. In general, select the simplest statistical procedure that adequately answers your research question. How do you know what technique is the simplest and best? Frankly, only experience can provide you with a good answer.

Properly Interpret and Report Your Results

As a part of the discussion of each analysis method in *IBM SPSS for Example*, suggestions for interpreting your results and reporting them in a professional manner are presented.

> The above items are important considerations for your data analyses. The discussion here is not comprehensive and cannot substitute for the expertise of a professional statistician. If you do not understand the relevance of these issues to your own analysis, we recommend that you consult a professional statistician.

Guidelines for Creating Data Sets

Information that consists of thousands (or millions) of raw numbers and codes provides little information useful for decision making. Like a builder who transforms raw materials into a functional skyscraper, statistical data analysis transforms raw data into meaningful and useful information. However, before you can begin to perform your data analysis, you must get that raw data into the software program.

Before entering data for analysis, there are several data issues you should address. The following discussion describes how to prepare a data set for use in any statistical software program. For specific requirements in SPSS, see Appendix A: A Brief Tutorial for Using IBM SPSS for Windows. Also, for a general discussion of this topic, see Elliott, Hynan, Reisch, and Smith (2006).

1. Decide What Variables You Need and Document Them

Your research question determines which variables are needed for your analysis. Researchers should document their variables in a "data dictionary" that contains the important information defining the variables. (Some texts refer to this as a data codebook.) For an example of a data dictionary, see Table 1.3. This table is created during your planning stage. It can be created using a spreadsheet or a word processor. Creating this simple "dictionary" before you collect your data not only forces you to consider which variables you will need in your data set, their types, and how they will be named but it also provides documentation that can be a valuable tool to you and others in performing and interpreting your analyses later on.

Variables may contain values (data types) that are either represented by alphabetic characters called strings (such as M and F or A, B, and C) or numbers (such as 0 and 1) whose meaning may not be completely clear. For example, if you coded a *female* variable as 1 and 0 (1 = female gender, and 0 = not female) and *race* as AA, C, H, and O, you will want to define those codes in your data dictionary, as illustrated for the *female* and *race* variables in Table 1.3.

Note that when you create a categorical variable, you may need to also include an "other" designation when the list does not include an exhaustive collection of possibilities. For example, the *race* variable includes an "O" for other or unknown.

Missing value codes are used to indicate that the information is not available for that variable. It is recommended that you use a proactive code rather than leaving missing data as blanks. See the section "Define Missing Value Codes" below for more information.

Table 1.3 Sample Data Dictionary

Variable Name	Label	Data Type (Width)	Value Codes	Missing Code
id	Identification number	String (4)	0001 to 9999	Not allowed
age	Age January 1, 2015	Numeric (3.0)	None	–9
female	Female Gender	Numeric (1.0)	1 = Female 0 = Not Female	9
tdate	Test date	Date (10) (mm/dd/yyyy)	None	11/11/1111
score	Initial test score	Numeric (6.2)	None	–9
race	Indicated Race	String (2)	AA = African American C = Caucasian H = Hispanic	O = Other or Unknown

One reason to code gender as *female* (1 = yes, 0 = no) rather than (Gender) M and F or (Gender) 1 and 2 is that a 0, 1 variable is easier to interpret in some analyses such as regression, and since 1 is typically set as Yes and 0 as No, the meaning of the variable is easy to interpret. (You could also call the variable *male* with 1 = Yes, is male, and 0 = No, is not male, but you do not need both variables.) You might also use this technique with other categorical variables. Suppose you record race as Caucasian, Hispanic, African-American, and Other. You could create the variable Caucasian as 0, 1; Hispanic as 0, 1; and African-American as 0, 1. The number of variables needed to define all groups is the number of categories minus 1. This allows you to make comparisons such as Hispanic versus all others in an analysis, which is sometimes preferred over using a (nominal) variable containing four categories. On the other hand, for a contingency table analysis, you might want to keep the four categories.

2. Design Your Data Set With One Subject (or Observation) Per Line

The vast majority of data analyses require your data set to contain one subject (or entity) per row. A properly designed data set should look something like Table 1.4.

Notice how this data set is designed. Each row contains data from a single subject. Each column contains the data from a single variable. You may be tempted to have multiple rows per subject or to design your data set with subjects as columns, but if you enter your data in that manner, you are only asking for problems later on in most cases. If your data are already in a data set where the subjects are in columns and your variables are in rows, see the transpose example in Appendix A: A Brief Tutorial for Using IBM SPSS for Windows for a way to realign your data file.

3. Each Variable Must Have a Properly Designated Name

Variable names are often short designations such as ID for subject identification number, SSBP (supine systolic blood pressure), and so on. Each statistical package has a set of restrictions for naming variables. The guidelines given here will help you design your data dictionary with variable names that are acceptable to most statistical programs:

- Variable names should begin with a letter but may also include numbers.
- Keep variable names short. Some programs require variable names of eight or fewer characters, although many allow names up to 64 characters in length.
- Do not use blanks or special characters (e.g., !, ?, ‘, and *).
- Variable names must be unique; no duplicate names are allowed.
- Case usually does not matter. Use any mixture of uppercase and lowercase characters when naming or referring to your variables. (Case does not matter in SPSS but it may matter in other statistical software, e.g., the R language.)

Table 1.4 Table Showing the First Three Records in a Typical Data Set

ID	Age	Female	Tdate	Score
1001	45	0	07/10/2015	60
1002	34	1	06/12/2015	55
1003	65	0	12/02/2015	62

4. Select Descriptive Labels for Each Variable

Creating a variable label allows you to associate a descriptive label with each variable name. Variable labels are important because they help you more clearly understand and interpret statistical output, particularly if the variable names are ambiguous, similar, or difficult to decipher. Typical names and labels might be the following:

age: Age on January 1, 2015

sbp: Systolic blood pressure

s1 to *s50*: Answers to a satisfaction survey

female: Female or non-female (male)

swq1: Sales for the southwest region during the first quarter

5. Select a Data Type for Each Variable

Each variable designates a particular type of information. The most commonly used variable types are numeric (a quantitative value) and string (also called character or text and often used for categorical-type data). (Numeric variables may also be indicated with the data types comma, dot, scientific notation, date, dollar, or custom currency.) A good rule of thumb is to designate as numeric only those variables that could be used in a calculation or that are factor or grouping codes for categorical variables. For example, a Social Security number, an ID number, and a telephone number are not really "numbers" that are used in calculations, and they should be designated as string values. This prevents the program from accidentally using that "number" in a calculation. However, it is common to designate dichotomous or grouping variables using numeric codes such as 0 and 1 or 1, 2, and 3, but care must be taken if you use these numbers in calculations. Also, never use codes such as "NA," "Missing," "> 100," or "10 to 20" as entries in numeric fields (which may occur if you first enter your data into a spreadsheet such as Excel and then import the data into your statistics program). For a list of specific data types in SPSS, see the section in Appendix A titled "Working With Data in SPSS."

6. Additional Tips for Categorical (String) Variables

Keep Case Consistent. For coded variables that are of the string (character or text) type, it is always good advice to maintain consistent case in data values. For example, use all uppercase ("Y" and "N") or all lowercase ("y" and "n") for a string type variable that is coded to represent yes or no. Even when case does not matter for variable names, it does matter for the data contents of the variables. The computer recognizes uppercase *Y* as a different character than lowercase *y*. Therefore, if you haphazardly use *Y, y, N,* and *n* as data entries, your program may recognize the data as having four categories instead of two.

Avoid Long Data Codes. Avoid long (and easy to misspell) string variables such as Influenza or Timer Clock Malfunction. Use shortened codes such as FLU and TCM instead. The Label field (see item number 4) can be used for a more complete description of the variable if needed.

Consider Binary Coding. If your data are binary (having only two levels such as diabetic or non-diabetic), creating a numeric variable that uses the values 0 and 1 may save time later since some analyses (such as regression) require numeric data.

7. Define Missing Values Codes

Sometimes data are lost or never collected. For example, a test tube is broken, a subject refuses to answer, or a patient fails to show up for an appointment. This type of data should be coded using a missing value code. Always select a missing value code that is an "impossible value" for the particular variable. For example, a –9 (negative 9) is an appropriate missing value code for *age, weight,* or *height* since that value would never be observed for those variables. Specifically, avoid using a blank or a 0 as a missing value code since that may cause confusion as to whether the data value was ever recorded and may cause an incorrect number to be used in a calculation. For a date variable, you can use a legitimate but highly unlikely date such as 11/11/1111 or 1/1/1800 as a missing value code (assuming your data do not include observations from the 12th or 19th century!). Once you specify a missing value code in your statistics program, the program will take that missing value into account when performing an analysis.

8. Consider the Need for a Grouping Variable

A grouping variable is a code that tells the statistical program how to separate records into groups—such as control group and experimental group. Therefore, if your data set contains information on two or more groups, you should include a variable that specifies the group membership of each observation. A grouping specification could be a single character (A, B, C), numeric (1, 2, 3), or names (CONTROL, TRT1, TRT2). For example, suppose you will be comparing the mean heights of 24-month-old males who were fed regularly with breast milk and those who were fed on formula. You could choose numeric grouping codes to be 1 and 0, where 1 means breast-fed and 0 means formula-fed. Or you could use string grouping codes such as B and F or BREAST and FORMULA or any other designation that makes sense to you. For example, Table 1.5 contains a grouping variable (named *group*) as well as two other variables, *subject* and *height*.

From this example, you can see how the program can tell that the height 30.4 belongs to Subject 1001 in Group B, the height 35.9 belongs to a subject in Group F, and so on.

Table 1.5 Sample Grouped Data

Subject	group	height
1001	B	30.4
1002	F	35.9
1003	B	30.2
1004	B	38.0
1005	F	34.3
etc. . . .		

Preparing Excel Data for Import

A number of researchers choose to first enter data using the Microsoft Excel program and then subsequently import that data set into a statistical program. This section describes how you should prepare your data in Excel (or any other spreadsheet or database program) for importation into a statistics program such as SPSS. Using the guidelines in the previous section, here are several additional items you should keep in mind. (The procedure for importing an Excel spreadsheet into SPSS is illustrated in Appendix A: A Brief Tutorial for Using IBM SPSS for Windows.)

1. Row 1 of your Excel spreadsheet should contain only variable names. Do not extend names to row 2.
2. Each subsequent row (line) in the Excel spreadsheet should contain data for a single subject or observed entity (in almost all cases).
3. Avoid blank rows—it will complicate your import and analysis.
4. If you have missing data in your data set, define a missing value code and place that code in any cell that contains missing data.
5. Always use date variables with four-digit year formats in Excel. That is, enter the date in Excel using the format 01/01/2015 and not 01/01/15. Otherwise, the old Y2K gotcha can still be a problem for date calculations because the date 1/1/15 could either represent the year 1915 or 2015.
6. Use your data dictionary (previously discussed), making sure to include all of the variables you will need. Use the specifications in the data dictionary such as codes, formats, and data ranges to determine how you will enter your data into Excel.

7. If you have the time or resources, enter your data twice (preferably using two different data entry people) and compare the two files. See Elliott et al. (2006) for an example of how to do a simple double-entry comparison in Excel.

8. Avoid putting any extraneous text into your spreadsheet. Instead, put explanatory information in other sheets in the same spreadsheet file. Extraneous data in your primary spreadsheet can make importing the data more difficult.

Guidelines for Reporting Results

All the statistics in the world will not get your point across unless you properly report your results. Most journals and publications have guidelines that you must follow when submitting your results. Along with each example in *IBM SPSS by Example*, we illustrate how you might report your findings using statements that are compatible with generally accepted formats. Since certain guidelines are commonly adopted when reporting statistical results, we present these general rules:

- Computer programs tend to report statistics (such as means, medians, standard deviations, etc.) to more digits than are necessary or meaningful. A generally accepted practice is to report statistics to one decimal place more than the resolution of the original measurements. For example, if age is measured as integer, report the average *age* using one decimal place. Occasionally, if precision is important, you may report more decimals. APA guidelines state that two or three significant digits (e.g., digits that convey information and are not merely placeholders) are usually sufficient for reporting any statistic. (However, you should use all decimal places reported in the computer output when using these results in further calculations.)

- For very large numbers, you may want to limit the number of significant digits depending on the nature of the measure. For example, if you are reporting the average salary of corporate presidents, you might report a mean of $723,000 and a standard deviation of $59,000 rather than $723,471.20 and $59,356.10.

- Whenever a number is less than 0, place a zero before the decimal. For example, use 0.003 instead of .003.

- When reporting percentages, include the counts as well. For example, "There were 19% males (12 of 64) represented in the sample." Note also that the percentage was rounded. In general, give percentages as whole numbers if the sample size is less than 100 and to one decimal place if the sample size is larger than 100 (Lang & Secic, 1997, p. 41).

- When using the APA format for reporting statistical results, use the appropriate abbreviations for common statistical measures. Examples are the following:

  Mean: $M = 1.34$

  Standard deviation: $SD = 3.21$

  Sample size: $N = 203$

  p-value: $p = 0.03$ or $p < 0.001$

  t-statistic with degrees of freedom: $t(13) = 2.12$

  Chi-square results: $\chi^2(2, N = 97) = 7.6, p = 0.02$

  F-test: $F(2, 21) = 3.33, p = 0.04$

- Information on creating and reporting results using graphs is covered in Chapter 3: Creating and Using Graphs.

Downloading Sample SPSS Data Files

This *IBM SPSS by Example* uses a number of data files that are used to illustrate the procedures described in this book. These data files are available for you to download from the Internet. To download these files onto your local hard drive, point your browser to the following site (enter in all lowercase):

http://www.alanelliott.com/spss2

Follow the instructions on this Web page to download and install the files onto your computer.

Opening Data Files for Examples

Once you have downloaded the sample data onto your hard drive, you can open these data files in the IBM SPSS program using the following steps:

1. Begin IBM SPSS.
2. From the main menu select **File/Open/Data**. . . .
3. In the "Look In" option on the Open Data dialog box, drill down to the C:\SPSSDATA folder on your computer (or wherever you stored your data).

SIDEBAR
Examples in the subsequent chapters assume that the data files have been downloaded and stored in a directory (folder) on your computer and that you know the name of the directory where the files are stored. The examples in this book use the convention that data files are stored in the C:\SPSSDATA folder.

SIDEBAR

To designate a series of menu selections using the SPSS drop down dialogs, we used the convention that a slash (/) means the next menu item selection. For example, **File/Open/Data** . . . indicates to select the File menu option from the main SPSS data screen. From the File menu options, select Open, and from the Open options, select Data. . . .

4. Select the file to open (such as EXAMPLE.SAV) and click OK. (Or simply double-click on the file name.) The data will be opened into the SPSS data grid.

5. You are now ready to use that data in an analysis.

In the examples referenced in subsequent chapters, you should use the above steps to open designated data files to perform the analyses under discussion.

SUMMARY

This chapter includes a description of the goals of this book, a brief review of statistical concepts, guidelines for creating a data set, entering data into Excel, presenting results, and instructions on how to download example data. The next chapter plunges you into the analysis process, beginning with a look at how to describe your data.

REFERENCES

American Psychological Association (APA). (2013). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Elliott, A. C., Hynan, L. S., Reisch, J. S., & Smith, J. P. (2006). Preparing data for analysis using Microsoft Excel. *Journal of Investigative Medicine*, *54*(6), 334–341.

Ferguson, Christopher J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40.5*, 532.

Friedman, J. A., Chalmers, T. C., Smith, H., & Kuebler, R. R. (1978). The importance of beta, the Type II error and sample size in the design and interpretation of the randomized control trial. *NEJM*, *299*, 690–696.

Gibaldi, J. (2006). *MLA handbook for writers of research papers* (7th ed.). New York, NY: Modern Language Association of America.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Mahwah, NJ: Pearson Prentice Hall.

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

Miller, R. G. (1981). *Simultaneous statistical inference* (2nd ed.). New York, NY: Springer Verlag.

Moore, D., & McCabe, G. (2012). *Introduction to the practice of statistics* (7th ed.). New York, NY: Freeman.

Nakagawa, Shinichi, & Cuthill, Innes C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews*, *82.4*, 591–605.

Wilkinson, L., & the Task Force on Statistical Inference, APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604.