CHAPTER 2

ADVANCED DATA SCREENING

Outliers and Missing Values

2.1 INTRODUCTION

Extensive data screening should be conducted prior to all analyses. Univariate and bivariate data screening are still necessary (as described in Volume I [Warner, 2020]). This chapter provides further discussion of outliers and procedures for handling missing data. It is important to formulate decision rules for data screening and handling prior to data collection and to document the process thoroughly.

During data screening, a researcher does several things:

- Correct errors.
- "Get to know" the data (for example, identify distribution shapes).
- Assess whether assumptions required for intended analyses are satisfied.
- Correct violations of assumptions, if possible.
- Identify and remedy problems such as outliers, skewness, and missing values.

The following section suggests ways to keep track of the data-screening process for large numbers of variables.

2.2 VARIABLE NAMES AND FILE MANAGEMENT

2.2.1 Case Identification Numbers

If there are no case identification numbers, create them. Often the original case numbers used to identify individuals during data collection in social sciences are removed to ensure confidentiality. Case numbers that correspond to row numbers in the SPSS file can be created using this command: COMPUTE idnumber = $casenum (where $casenum denotes row number in the SPSS file). The variable idnumber can be used to label individual cases in graphs and identify which cases have outliers or missing values.

2.2.2 Codes for Missing Values

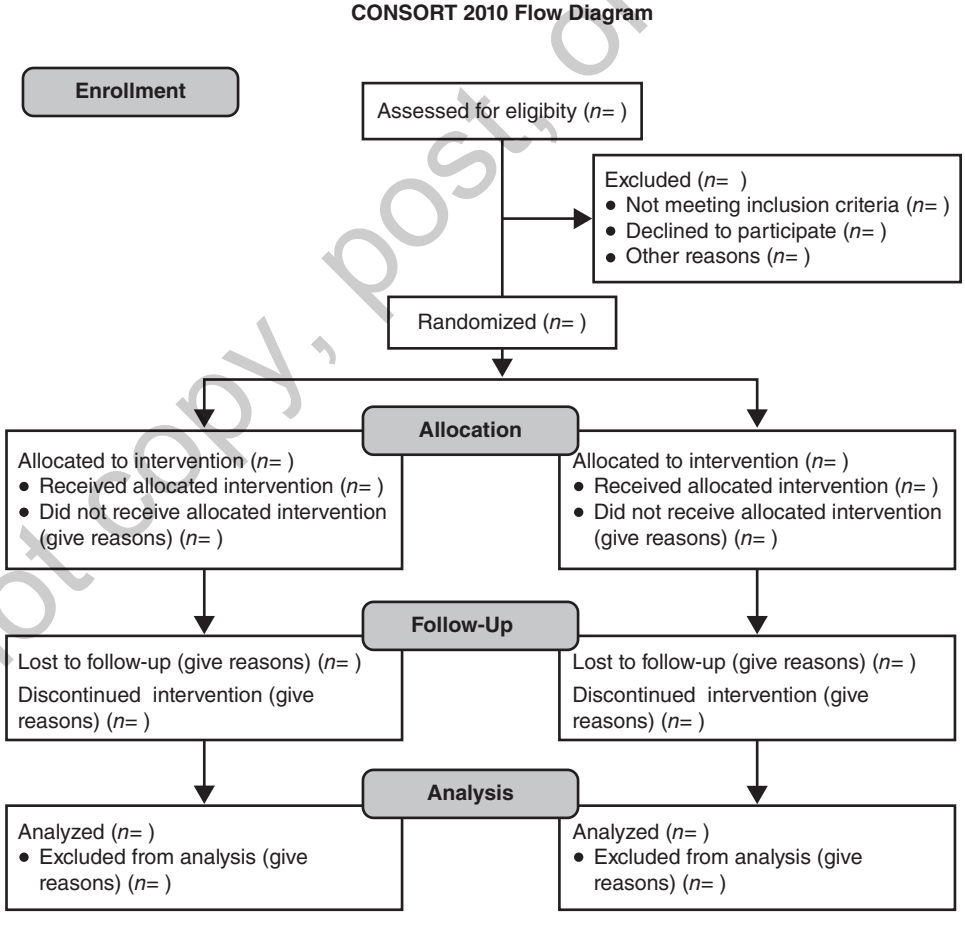
Missing values are usually identified by leaving cells in the SPSS data worksheet blank. In recode or compute statements, a blank cell corresponds to the value $sysmis. In some kinds of research, it is useful to document different reasons for missing values (Acock, 2005). For

example, a survey response can be missing because a participant refuses to answer or cannot remember the information; a physiological measure may be missed because of equipment malfunction. Different numerical codes can be used for each type of missing value. Be sure to use number codes for missing that cannot occur as valid score values. For example, number of tickets for traffic violations could be coded 888 for "refused to answer" and 999 for "could not remember." Archival data files sometimes use multiple codes for missing. Missing values are specified and labeled in the SPSS Data Editor Variable View tab.

2.2.3 Keeping Track of Files

It is common for data analysts to go through a multiple-step process in data screening; this is particularly likely when longitudinal data are collected. A flowchart may be needed to keep track of scores that are modified and cases that are lost due to attrition. The CONSORT (Consolidated Standards of Reporting Trials) protocol describes a way to do this (Boutron, 2017). Figure 2.1 shows a template for a CONSORT flow diagram.

Figure 2.1 Flowchart: CONSORT Protocol to Track Participant Attrition and Data Handling



Source: http://www.consort-statement.org/consort-statement/flow-diagram.

It is important to retain the original data file and save modified data files at every step during this process. If you change your mind about some decisions, or discover errors, you may need to go back to earlier versions of files. Keep a detailed log that documents what was done to data at each step. Use of file names that include the date and time of file creation and/or words that remind you what was done at each step can be helpful when you need to locate the most recent version or backtrack to earlier versions. Naming a file "final" is almost never a good idea. (Files are time-stamped by computers, but these time stamps are not always adequate information.)

2.2.4 Use Different Variable Names to Keep Track of Modifications

If a variable will be transformed or recoded before use in later analysis, it is helpful to use an initial variable name indicating that this change has not yet been done. For example, an initial score for reaction time could be named raw_rt. The log-transformed version of the variable could be called log_rt. As another example, some self-report measures include reverse-worded questions. For example, most items in a depression scale might be worded such that higher degree of agreement indicates more depression (e.g., "I feel sad most of the time," rated on a scale from 1 = *strongly disagree* to 5 = *strongly agree*). Some items might be reverse worded (such that a high score indicates less depression; e.g., "Most days I am happy"). Before scores can be summed to create a total depression score, scoring for reverse-worded questions must be changed to make scores consistent (e.g., a score of 5 always indicates higher depression). The initial name for a reverse-worded question could be rev_depression1. The "rev" prefix would indicate that this item was worded in a direction opposite from other items. After recoding to change the direction of scoring, the new variable name could be depression1 (without the "rev" prefix). Then the total scale score could be computed by summing depression1, depression2, and so on. Avoid using the same names for variables before and after transformations or recodes, because this can lead to confusion.

2.2.5 Save SPSS Syntax

The Paste button in SPSS dialog boxes can be used to save SPSS commands generated by your menu selections into a syntax file. Save all SPSS syntax used to recode, transform, compute new variables, or make other modifications during the data preparation process. The syntax file documents what was done, and if errors are discovered, syntax can be edited to make corrections and all analyses can be done again. This can save considerable time.

Data screening is needed so that when the analyses of primary interest are conducted, the best possible information is available. If problems such as outliers and missing values are not corrected during data screening, the results of final analyses are likely to be biased.

2.3 SOURCES OF BIAS

Bias can be defined as over- or underestimation of statistics such as values of M, t ratios, and p values. Bias means that the sample statistic over- or underestimates the corresponding population parameters (e.g., M is systematically larger or smaller than μ). Bias can occur when assumptions for analyses are violated and when outliers or missing data are present.

Most of the statistics in this book (except for logistic regression) are special cases of the general linear model (GLM). Most GLM analyses were developed on the basis of the following assumptions. Some assumptions are explicit (i.e., assumed in derivations of statistics). There are also implicit assumptions and rules for the use of significance tests in practice (e.g., don't run hundreds of tests and report only those with $p < .05$; selected p values will greatly underestimate the risk for Type I decision error). Problems such as outliers and

missing data often arise in real-world data. The actual practice of statistics is much messier than the ideal world imagined by mathematical statisticians. Here is a list of concerns that should be addressed in data screening. Some of these things are relatively easy to identify and correct, while others are more difficult.

- **Scores within samples must be independent of one another.** Whether this assumption is satisfied depends primarily on how data were collected (Volume I [Warner, 2020], Chapter 2). Scores in samples are not independent if participants can influence one another's behavior through processes such as persuasion, cooperation, imitation, or competition. See Kenny and Judd (1986, 1996) for discussion. When this assumption is violated, estimates of *SD* or SS_{within} are often too small; that makes estimates of *t* or *F* too large and results in inflated risk for Type I error. Violations of this assumption are a serious problem.
- **All relationships among variables are linear.** This is an extremely important assumption that we can check in samples by visual examination of scatterplots and by tests of nonlinearity. Nonlinear terms (such as X^2, in addition to *X* as a predictor of *Y*) can be added to linear regression models, but sometimes nonlinearity points to the need for other courses of action, such as nonlinear data transformation or analyses outside the GLM family.
- **Missing values** can lead to bias in the composition of samples and corresponding bias in estimation of statistics. Often, cases with missing values differ in some way from the cases with complete data. Suppose men are more likely not to answer a question about depression than women, or that students with low grades are more likely to skip questions about academic performance. If these cases are dropped, the sample becomes biased (the sample will underrepresent men and/or low-performing students). Later sections in this chapter discuss methods for evaluation of amount and pattern of missing values and replacement of missing values with estimated or imputed scores. Whether problems with missing values can be remedied depends on the reasons for missingness, as discussed in that section.
- **Residuals or prediction errors are independent of one another, are normally distributed, and have mean of 0 and equal variance for all values of predictor variables.** For regression analysis and related techniques such as time-series analysis, these assumptions can be evaluated using plots and descriptive statistics for residuals. Data analysts should beware the temptation to drop cases just because they have large residuals (Tabachnick & Fidell, 2018). This can amount to trimming the data to fit the model. Users of regression are more likely to focus on residuals than users of analysis of variance (ANOVA).
- **Some sample distribution shapes make *M* a poor description of central tendency.** For example, a bimodal distribution of ratings on a 1-to-7 scale, with a mode at the lowest and highest scores (as we would see for highly polarized ratings), is not well described by a sample mean (see Volume I [Warner, 2020], Chapter 5). We need to do something else with these data. If sample size is large enough, we may be able to treat each *X* score (e.g., $X = 1, X = 2, \ldots, X = 7$) as a separate group. With large samples (on the order of thousands) it may be better to treat some quantitative variables as categorical.
- **Some distribution types require different kinds of analysis.** For example, when a *Y* dependent variable is a count of behaviors such as occasions of drug use, the histogram for the distribution of *Y* may have a mode at 0. Analyses outside the GLM family, such as zero-inflated negative binomial regression, may be needed for this

kind of dependent variable (see Appendix 2A). The remedy for this kind of problem is to choose an appropriate analysis.

- **Skewness of sample distribution shape.** Skewness can be evaluated by visual examination of histograms. SPSS provides a skewness index and its standard error; statistical significance of skewness can be assessed by examining z = skewness/SE_{skewness}, using the standard normal distribution to evaluate z. However, visual examination is often adequate and may provide insight into reasons for skewness that the skewness index by itself cannot provide. Skewness is not always a major problem. Sometimes sample skewness can be eliminated or reduced by removal or modification of outliers. If skewness is severe and not due to just a few outliers, transformations such as log may be useful ways to reduce skewness (discussed in a later section).

- Derivations of many statistical significance tests assume that **scores in samples are randomly selected from normally distributed populations**. This raises two issues. On one hand, some data analysts worry about the normality of their population distributions. I worry more about the use of convenience samples that were not selected from any well-defined population. The use of convenience samples can limit generalizability of results. On the other hand, Monte Carlo simulations that evaluate violations of this normally distributed population assumption for artificially generated populations of data and simple analyses such as the independent-samples t test often find that violations of this assumption do not seriously bias p values, provided that samples are not too small (Sawilowsky & Blair, 1992). There are significance tests, such as Levene F, to test differences between sample variances for t and F tests. However, tests that are adjusted to correct for violations of this assumption, such as the "equal variances not assumed" or Welch's t, are generally thought to be overly conservative. The issue here is that we often don't know anything about population distribution shape. For some simple analyses, such as independent-samples t and between-S ANOVA, violations of assumption of normal distribution in the population may not cause serious problems.

- **For more advanced statistical methods, violations of normality assumptions may be much more serious.** These problems can be avoided through the use of robust estimation methods that do not require normality assumptions (Field & Wilcox, 2017; Maronna, Martin, Yohai, & Salibián-Barrera, 2019).

- **Violation of assumption that all variables are measured without error** (that all measures are perfectly reliable and perfectly valid). This is almost never true in real data. Advanced techniques such as structural equation modeling include measurement models that take measurement error into account (to some extent).

- **Model must be properly specified.** A properly specified model in regression includes all the predictors of Y that should be included, includes terms such as interactions if these are needed, and does not include "garbage" variables that should not be statistically controlled. We can never be sure that we have a correctly specified model. Kenny (1979) noted that when we add or drop variables from a regression, we can have "bouncing betas" (regression slope estimates can change dramatically). The value of each beta coefficient depends on context (i.e., which other variables are included in the model). Significance tests for b coefficients vary depending on the set of variables that are controlled when assessing each predictor. Another way to say this is that we cannot obtain unbiased estimates of effects unless we control for the "right" set of variables. Decisions about which variables to control are limited by the variables that are available in the data set. Unfortunately, it is

common practice for data analysts to add and/or drop control variables until they find that the predictor variable of interest becomes statistically significant.

Some of these assumptions (such as normally distributed scores in the population from which the sample was selected) cannot be checked. Some potential problems can be evaluated through screening of sample data.

2.4 SCREENING SAMPLE DATA

From a practical and applied perspective, what are the most important things to check for during preliminary data screening? First, remember that rules for identification and handling of problems with data, such as outliers, skewness, and missing values, should be established before you collect data. If you experiment with different rules for outlier detection and handling, run numerous analyses, and report selected results, the risk for committing Type I decision error increases, often substantially. Doing whatever it takes to obtain statistically significant values of *p* is called *p*-hacking (Wicherts et al., 2016), and this can lead to misleading results (Simmons, Nelson, & Simonsohn, 2011). Committing to decisions about data handling prior to data collection can reduce the temptation to engage in *p*-hacking.

2.4.1 Data Screening Needed in All Situations

- Individual scores should always be evaluated to make sure that all score values are plausible and accurate and that the ranges of scores in the sample (for important variables) corresponds at least approximately to the ranges of scores in the hypothetical population of interest. If a study includes persons with depression scores that range only from 0 to 10, we cannot generalize or extrapolate findings to persons with depression scores above 10. A frequency table provides information about range of scores.
- Missing values. Begin by evaluating how many missing data there are; frequency tables tell us how many missing values there are for each variable. If there are very few missing values (e.g., less than 5% of observations), missing data may not be a great concern, and it may be acceptable to let SPSS use default methods such as listwise deletion for handling missing data. If there are larger amounts of missing data, this raises concerns whether data are missing systematically. A later section in this chapter discusses the missing data problem further.
- Evaluate distribution skewness. When a distribution is asymmetrical, it often has a longer and thinner tail at one end than the other. It is possible that an appearance of skewness arises because of a few outliers. If this is the case, I recommend that you handle this as an outlier problem. A later section discusses possible ways to handle skewness if it is not due to just a few outliers. Some variables (such as income) predictably have very strong positive skewness. Sometimes nonlinear data transformations are used to reduce skewness.

2.4.2 Data Screening for Comparison of Group Means

- Make sure all groups have adequate *n*'s. If we have at least $n = 30$ cases per group, and use two-tailed tests, violations of the population normality assumption and of assumptions about equal population variances do not seriously bias *p* value estimates (Sawilowsky & Blair, 1992). Some authorities suggest that even smaller values of *n* may be adequate. I believe that below some point (perhaps *n* of 20 per group), there is just not sufficient information to describe groups or to evaluate whether

the group is similar to populations of interest. However, this is not an ironclad rule. In some kinds of research (such as neuroscience), it is reasonable for researchers to assume little variation among cases with respect to important characteristics such as brain structure and function, and recruiting and paying for cases can be expensive because of time-consuming procedures. For example, in behavioral neuroscience animal research, each case may require extensive training, then surgery, then extensive testing or evaluation or costly laboratory analysis of specimen materials. Procedures such as magnetic resonance imaging are very costly. Sometimes smaller *n*'s are all we can get.

- Check for outliers within groups. Outliers within groups affect estimates of both *M* and *SD*, and these in turn will affect estimates of *t* and *p*. The effect of outliers may be either to inflate or deflate the *t* ratio. Boxplots are a common way to identify outliers within groups.
- Examine distribution shapes in groups to evaluate whether *M* is a reasonable description of central tendency. Some distribution shapes (such as a bimodal distribution with modes at the extreme high and low ends of the distribution and distributions with large modes at 0) can make *M* a poor way to describe central tendency (Volume I [Warner, 2020], Chapter 5). If these distribution shapes are seen in sample data, the data analyst should consider whether comparison of means is a good way to evaluate outcomes.

2.4.3 Data Screening for Correlation and Regression

- Check that relations between all variables are linear if you plan to use linear correlation and linear regression methods. In addition, predictor variables should not be too highly correlated with one another. Visual examination of a scatterplot may be sufficient; regression can also be used to evaluate nonlinearity (discussed in Section 2.8).
- Check for outliers. Bivariate and multivariate outliers can inflate or deflate correlations among variables. Bivariate outliers can be detected by visual examination of an *X*, *Y* scatterplot. For more than two variables, you need to look for multivariate outliers (described later in this chapter).
- Evaluate whether *X* and *Y* have similar distribution shapes. It may be more important that *X* and *Y* have similar distribution shapes than that their sample distribution shapes are normal. When distribution shapes differ, the maximum obtainable value for *r* will have a limited range (not the full range from –1 to +1). This in turn will influence estimates for analyses that use *r* as a building block. Visual examination of histograms may be sufficient. See Appendix 10D in Volume I (Warner, 2020).
- Evaluate plots of residuals from regression to verify that they are (a) normally distributed and (b) not related to values of *Y* or Y' (these are assumptions for regression). If you have only one predictor, screening raw scores on variables may lead to the same conclusions as screening residuals. Tabachnick and Fidell (2018) pointed out that when a researcher runs the final analysis of primary interest and then examines residuals, it can be tempting to remove or modify cases specifically because they cause poor fit in the final analysis. In other words, the data analyst may be tempted to trim the data (post hoc) to fit the model.
- Sample distributions that differ drastically from normal may alert you to the need for different kinds of analyses outside the GLM family (an example is provided in Appendix 2A).

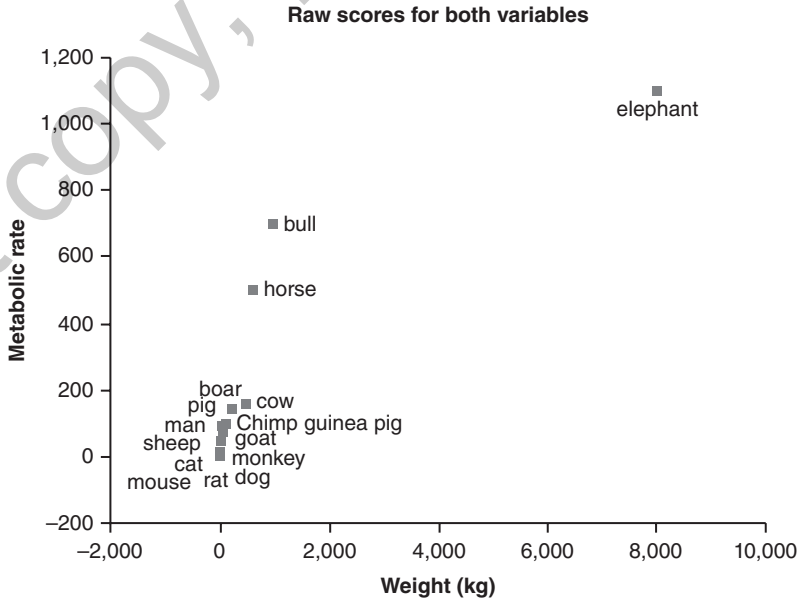
2.5 POSSIBLE REMEDY FOR SKEWNESS: NONLINEAR DATA TRANSFORMATIONS

Nonlinear transformations of X (such as $1/X$, X^2, X^c for any value of c, base 10 or natural log of X, arcsine of X, and others) can change the shape of distributions (Tabachnick & Fidell, 2018). Although log transformations can potentially reduce positive or negative skewness in an otherwise normal distribution, they are not always appropriate or effective. In many situations, if distribution shape can be made reasonably normal by modifying or removing outliers, it may be preferable to do that. Log transformations make sense when at least one of the following conditions are met:

- The underlying distribution is exponential.
- It is conventional to use log transformations with this variable; readers and reviewers are familiar with it.
- Scores on the variable differ across orders of magnitude. Scores differ across orders of magnitude when the highest value is vastly larger than the smallest value. Consider the following example: The weight of an elephant can be tens of thousands of times greater than the weight of a mouse. Typical values for body weight for different species, given in kilograms, appear on the X axis in Figure 2.2.

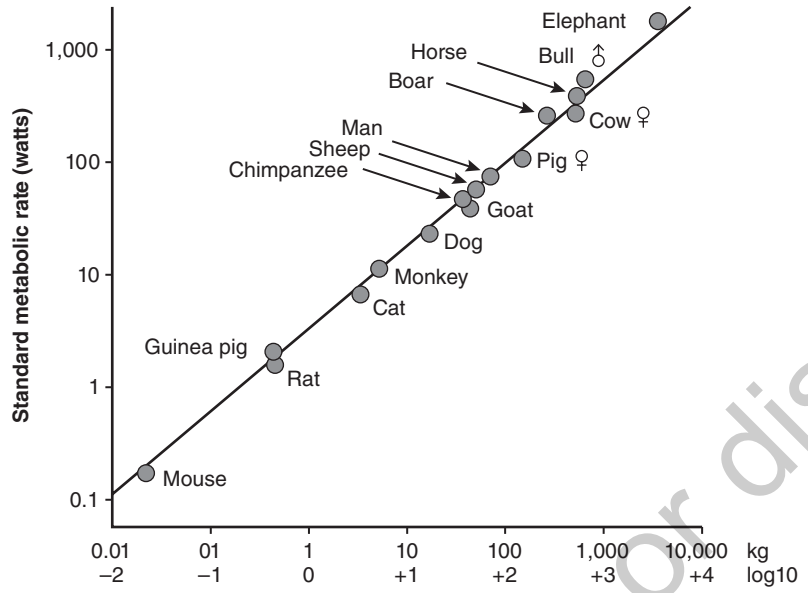
Because of outliers (weight and metabolic rates for elephants), scores for body weight (and metabolism) of smaller species are crowded together in the lower left-hand corner of the graph, making it difficult to distinguish differences among most species. When the base 10 log is taken for both variables, as shown in Figure 2.3, scores for species are spread out more evenly on the X and Y axes. Differences among them are now represented in log units (orders of magnitude). In addition, the relation between log of weight and log of metabolic rate becomes linear (of course, this will not happen for all log-transformed variables).

Figure 2.2 Scatterplot of Metabolic Rate by Body Weight (Raw Scores)



Source: Reprinted with permission from Dr. Tatsuo Motokawa.

Figure 2.3 Scatterplot of Log Metabolic Rate by Log Body Weight



Source: Reprinted with permission from Dr. Tatsuo Motokawa.

Other transformations commonly used in some areas of psychology involve power functions, that is, replacing X with X^2, or X^c (where c is some power of X; the exponent c is not necessarily an integer value). Power transformations are used in psychophysical studies (e.g., to examine how perceived heaviness of objects is related to physical mass).

When individual scores for cases are proportions, percentages, or correlations, other nonlinear transformations may be needed. Data transformations such as arcsine (for proportions) or Fisher r to Z are used to correct problems with the shapes of sampling distributions that arise when the range of possible score values has fixed end points (–1 to +1 for correlation, 0 to 1.00 for proportion).

If you use nonlinear transformations to reduce skewness, examine a histogram for the transformed scores to see whether the transformation had the desired effect. In my experience, distributions of log-transformed scores often do not look any better than the raw scores. When X does not have a very wide range, the correlation of X with X^2, or X with log X, is often very close to 1. In these situations, the transformation does not have much effect on distribution shape.

2.6 IDENTIFICATION OF OUTLIERS

2.6.1 Univariate Outliers

Outliers can be a problem because many widely used statistics, such as the sample mean M, are not robust against the effect of outliers. In turn, other statistics that use M in computations (such as SD and r and t) can also be influenced by outliers. Outliers can bias estimates of parameters, effect sizes, standard errors, confidence intervals, and test statistics such as t and F ratios and their corresponding p values (Field, 2018).

It is often possible to anticipate which variables are likely to have outliers. If scores are ratings on 1-to-5 or 1-to-7 scales, extreme outliers cannot occur. However, many variables

(such as income) have no fixed upper limit; in these situations, outliers are common. When you know ahead of time that some of your variables are likely to generate outliers, it's important to make decisions ahead of time. What rules will you use to identify scores as outliers, and what methods will you use to handle outliers? Outliers are sometimes obtained because of equipment malfunction or other forms of measurement error.

If groups will be compared, outlier evaluation should be done separately within each group (e.g., a separate boxplot within each group).

To review briefly:

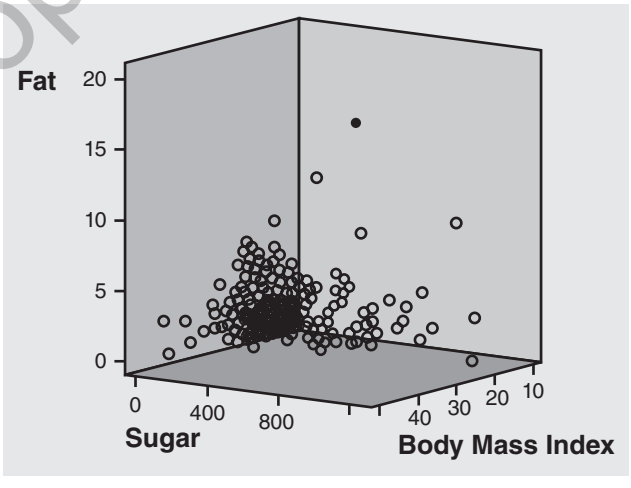
- In boxplots, scores that lie outside the "whiskers" can be considered potential outliers (an open circle represents an outlier; an asterisk represents an extreme outlier). Boxplots are particularly appropriate for non-normally distributed data.
- Scores can be identified as outliers if they have z values greater than 3.29 in absolute value for the distribution within each group (Tabachnick & Fidell, 2018).

These are arbitrary rules; they are suggested here because they make sense in a wide range of situations. Aguinas, Gottfredson, and Joo (2013) provided numerous other possible suggestions for outlier identification.

2.6.2 Bivariate and Multivariate Outliers

Bivariate outliers affect estimates of correlations and regression slopes. In bivariate scatterplots it is easy to see whether an individual data point is far away from the cloud that contains most other data points. This distance can be quantified by computing a Mahalanobis distance. Mahalanobis distance can be generalized to situations with larger numbers of variables. A score with a large Mahalanobis distance corresponds to a point that is outside the cloud that contains most of the other data points, as shown in the three-dimensional plot for three variables in Figure 2.4. The most extreme multivariate outlier is shown as a filled circle near the top.

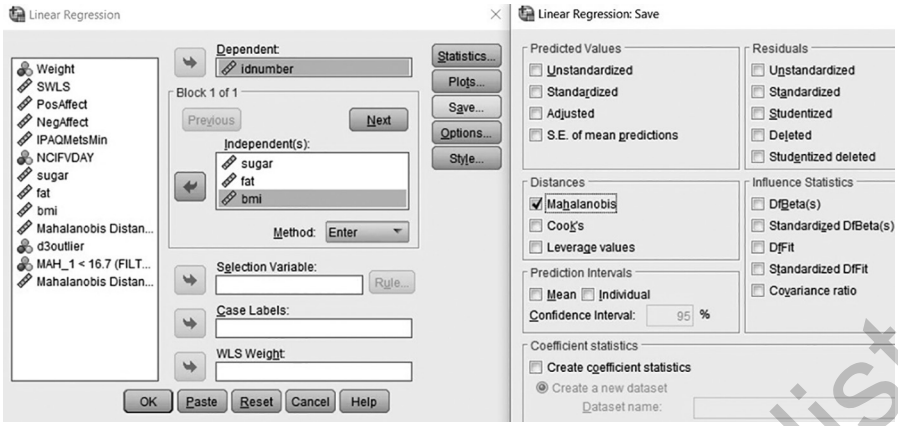
Figure 2.4 Multivariate Outlier for Combination of Three Variables



Source: Data selected and extensively modified from Warner, Frye, Morrell, and Carey (2017).

Note: Fat is the number of fat servings per day, and sugar is the number of sugar calories per day.

Figure 2.5 Linear Regression Dialog Boxes



Mahalanobis distance can be obtained as a diagnostic when running analyses such as multiple regression and discriminant analysis. Tabachnick and Fidell (2018) suggested a method to obtain Mahalanobis distance for a set of variables without "previewing" the final regression analysis of interest. Their suggested method avoids the temptation to remove outliers that reduce goodness of fit for the final model. They suggested using the case identification number as the dependent variable in a linear regression and using the entire set of variables to be examined for multivariate outliers as predictors. (This works because multivariate outliers among predictors are unaffected by subject identification number; Tabachnick & Fidell, 2018).

Data in the file outlierfvi.sav are used to demonstrate how to obtain and interpret Mahalanobis distance for a set of hypothetical data. The initial menu selections are <Analyze> → <Regression> → <Linear>. This opens the Linear regression dialog box on the left-hand side of Figure 2.5. Idnumber is entered as the dependent variable. To examine whether there are multivariate outliers in a set of three variables, all three variables are entered as predictor variables in the Linear Regression dialog box. Click the Save button. This opens the Linear Regression: Save dialog box that appears on the right-hand side of Figure 2.5. Check the box for "Mahalanobis." Click Continue, then OK.

After the regression has been run, SPSS Data View (Figure 2.6) has a new variable named MAH_1. (The tag "_1" at the end of the variable name indicates that this is from the first regression analysis that was run.) This is the Mahalanobis distance score for each individual participant; it tells you the degree to which that person's combination of scores on fat, sugar, and body mass index (BMI) was a multivariate outlier, relative to the cloud that these scores occupied in three-dimensional space (shown previously in Figure 2.4). The file was sorted in descending order by values of MAH_1; the part of the file that appears in Figure 2.6 shows a subset of persons whose scores could be identified as multivariate outliers, because they had large values of Mahalanobis distance (many other cases not shown in Figure 2.6 had smaller values of Mahalanobis distance).

Mahalanobis distance has a χ^2 distribution with *df* equal to the number of predictor variables (Tabachnick & Fidell, 2018). The largest value was MAH_1 = 77.76 (for idnumber = 421). The critical value of chi squared with 3 *df*, using $\alpha = .001$, is 16.27. Using that value of χ^2 as a criterion, MAH_1 would be judged statistically significant for all cases listed in Figure 2.6. If the decision to use Mahalanobis distance as a criterion for the identification of outliers was made prior to data screening, scores for all three variables for the cases with significant values of Mahalanobis distance could be converted to missing values. If this results in fewer than 5% missing values, this small amount of missing data may not bias results. If more than 5% of cases have missing values, some form of imputation (described elsewhere in the chapter) could be used to replace the missing values with reasonable estimates.

Figure 2.6 SPSS Data View With Saved Mahalanobis Distance

idnumber	MAH_1	sugar	fat	bmi
421.00	77.76280	624.00	16.00	24.13
258.00	67.01817	338.00	11.50	21.25
213.00	40.08815	1248.00	.00	22.50
153.00	39.69702	.00	.00	43.16
278.00	39.51425	1040.00	9.00	17.16
134.00	34.02619	.00	4.50	18.47
210.00	33.44055	1248.00	3.00	21.92
151.00	33.32354	.00	.00	44.16
173.00	32.72449	.00	2.50	45.10
97.00	32.03592	966.00	4.50	31.20
147.00	23.09725	873.00	4.00	18.67
317.00	22.35106	1040.00	2.00	23.49
357.00	22.29595	104.00	8.00	21.14
125.00	22.11335	.00	3.50	23.12
425.00	21.08898	884.00	2.00	25.10
152.00	20.55901	244.00	3.50	39.06
112.00	20.04464	.00	.00	23.75
115.00	20.02894	602.00	8.00	21.77
80.00	19.83612	.00	2.00	39.53
145.00	18.91604	104.00	1.50	39.58
12.00	18.72081	936.00	1.00	21.63

Examination of scores for sugar and fat consumption and BMI for the case on the first row in Figure 2.6 indicates that this person had a BMI within normal range (the normal range for BMI is generally defined as 18.5 to 24.9 kg/m^2), even though this person reported consuming 16 servings of fat per day. (The value of 16 servings of fat per day was a univariate outlier.) Although this might be physically possible, this seems unlikely. In actual data screening, 16 servings of fat would have been tagged as a univariate outlier and modified at an earlier stage in data screening.

Several additional cases had statistically significant values for Mahalanobis distance. When there are numerous multivariate outliers, Tabachnick and Fidell (2018) suggested additional examination of this group of cases to see what might distinguish them from nonoutlier cases.

2.7 HANDLING OUTLIERS

2.7.1 Use Different Analyses: Nonparametric or Robust Methods

The most widely used parametric statistics (those covered in Volume I [Warner, 2020], and the present volume) that are part of the GLM are generally not robust against the effect of outliers. One way to handle outliers is to use different analyses. Many nonparametric statistics convert scores to ranks as part of computation; this gets rid of outliers. However, it would be incorrect to assume that use of nonparametric statistics makes everything simple. Statistics such as the Wilcoxon rank sum test do not require scores to be normally distributed, but they assume that the distribution shape is the same across groups, and in practice, data often violate that assumption.

Robust statistical techniques, often implemented using R (Field & Wilcox, 2017; Maronna et al., 2019) do not require the assumptions made for GLM. Robust methods are beyond the scope of this volume. They will likely become more widely used in the future.

2.7.2 Handling Univariate Outliers

Suppose that you have identified scores in your data file as univariate outliers (because they were tagged in a boxplot, because they had $z > 3.29$ in absolute value, or on the basis of other rules). Rules for identification and handling of outliers should be decided before data collection, if possible.

Here are the four most obvious choices for outlier handling; there are many other ways (Aguinas et al., 2013).

- Do nothing. Run the analysis with the outliers included.
- Discard all outliers. Removal of extreme values is often called truncation or trimming.
- Replace all outliers with the next largest score value that is not considered an outlier. The information in boxplots can be used to identify outliers and find the next largest score value that is not an outlier. This is called Winsorizing.
- Run the analysis with the outliers included, and also with the outliers excluded, and report both analyses. (Do not just report the version of the analysis that you liked better.)

No matter which of these guidelines you choose, you must document how many outliers were identified, using what rule, and what was done with these outliers. Try to avoid using different rules for different variables or cases. If you have a different story about each data point you remove, it will sound like *p*-hacking, and in fact, it will probably be *p*-hacking. (That said, there may be precedent or specific reasons for outlier handling that apply to some variables and not others.) Do not experiment with different choices for outlier elimination and modification and then report the version of analysis you like best. That is *p*-hacking; the reported *p* value will greatly underestimate the true risk for Type I decision error.

2.7.3 Handling Bivariate and Multivariate Outliers

Consider bivariate outliers first. If you have scores for height in inches (X) and body weight in pounds (Y), and one case has $X = 73$ and $Y = 110$, the univariate scores are not extreme. The combination, however, would be very unusual. Winsorizing might not get rid of the problem, but you could do other things (exclude the case, or run the analysis both with and without this case), as long as you can justify your choice on the basis of plans you made prior to data collection.

For multivariate outliers, it may be possible to identify which one or two variables make the case an outlier. In the previous example of a multivariate outlier, the extremely high value of fat (in row 1 of the data file that appears in Figure 2.6) seemed inconsistent with the normal BMI score. A decision might be made to replace the high fat score with a lower valid score value that is not an outlier. However, detailed evaluation of multivariate outliers to assess whether one or two variables are responsible may be too time consuming to be practical.

Some multivariate outliers may disappear when univariate outliers have been modified. However, multivariate outliers can arise even when none of the individual variables is a univariate outlier.

2.8 TESTING LINEARITY ASSUMPTION

If an association between variables is not linear, it can be described as nonlinear, curvilinear, or perhaps a polynomial trend. Visual examination of bivariate scatterplots may be sufficient to evaluate possible nonlinearity. It is possible to test whether departure from linearity is statistically significant using regression analysis to predict Y from X^2 and perhaps even X^3 (in addition

to X). If adding X^2 to a regression equation that includes only X as a predictor leads to a significant increase in R^2, then the association can be called significantly nonlinear. The actual increase in R^2 would tell you whether nonlinearity predicts a trivial or large part of the variance in Y. For discussion of regression with two predictors, see Chapter 4.

If Y is a function of:

- Only X, then the X, Y function is a straight line; this represents a linear trend.
- X and X^2, then the X, Y function has one curve; this is a quadratic trend. It may resemble a U or inverted U shape.
- X, X^2, and X^3, then the function has two curves; this is a cubic trend.

Note that the number of curves in the X, Y function equals the highest power of X minus 1.

The bivariate regression model for a simple linear relationship is $Y' = b_0 + b \times X$. This can be expanded to include a quadratic term: $Y' = b_0 + b_1X + b_2X^2$. If the b_2 coefficient associated with the X^2 predictor variable is statistically significant, this indicates a significant departure from linearity. SPSS transform and compute commands are used to compute a new variable, Xsquared, that corresponds to X^2. (Similarly, we could compute $X^3 = X \times X \times X$; however, trends that are higher order than quadratic are not common in psychological data.)

The hypothetical data that correspond to the graphs in Figure 2.7 are in a file named linearitytest.sav (with $N = 13$ cases). Visual examination of the scatterplots in Figure 2.7 suggests a linear association of Y with X (left) and a quadratic association of Y with Q (right).

Let's first ask whether Y has a significantly nonlinear association with X for the scatterplot on the left of Figure 2.7. To do this, first compute the squared version of X. This can be done as follows using an SPSS compute statement. (If you are not familiar with compute statements, see Volume I [Warner, 2020], or an SPSS guide, or perform a Google search for this topic.)

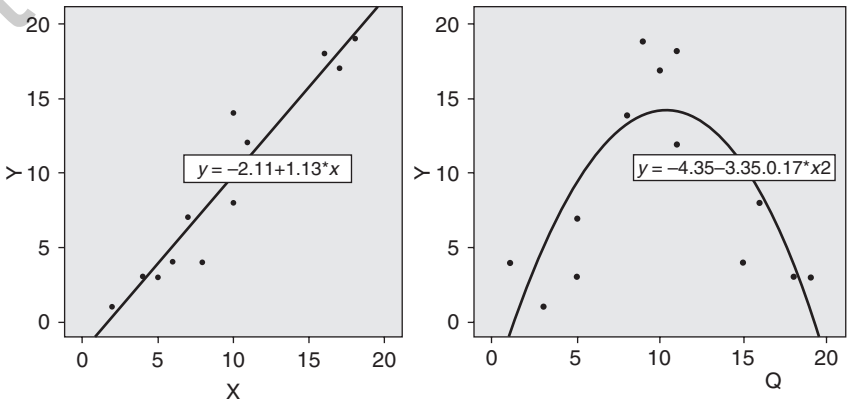
```
COMPUTE Xsquared = X * X
```

(A better way to compute X^2 is $(X - M_X) \times (X - M_X)$, where M_X is the mean of X.[1])

Then run SPSS linear regression using X and the new variable that corresponds to the squared value of X (named Xsquared) as predictors.

```
REGRESSION
/MISSING LISTWISE
```

Figure 2.7 Linear Versus Quadratic Trend Data



```
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT Y
/METHOD=ENTER X Xsquared
```

Partial results for this regression appear in Figure 2.8.

The X^2 term represents quadratic trend. If the *b* coefficient for this variable is statistically significant, the assumption of linearity is violated. In this situation, for X^2, $b = -.001$, $\beta = -.025$, $t(10) = -.048$, $p = .963$, two tailed (the *df* error term for the *t* test appears in a part of regression output that is not included here). The assumption of linearity is not significantly violated for *X* as a predictor of *Y*.

The same procedure can be used to ask whether there is a violation of the linearity assumption when *Q* is used to predict *Y*. First, compute a new variable named Qsquared; then run the regression using Q and Qsquared as predictors.

For Q^2, $b = -.173$, $\beta = -3.234$, $t(10) = -4.20$, and $p = .002$, two tailed (values from Figure 2.9). (While β coefficients usually range between –1 and +1, they can be far outside that range when squared terms or products between variables are used as predictors.) The linearity assumption was significantly violated for *Q* as a predictor of *Y*.

What can be done when the linearity assumption is violated? Sometimes a data transformation such as log will make the relation between a pair of variables more nearly linear; possibly the log of *Q* would have a linear association with the log of *Y*. However, this works in only a few situations. Another option is to incorporate the identified nonlinearity into later analyses, for example, include X^2 as a predictor in later regression analyses, so that the nonlinearity detected during data screening is taken into account.

Figure 2.8 Regression Coefficients for Quadratic Regression (Prediction of Y Scores From Scores on X and X^2)

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-2.200	2.276		-.967	.357
	X	1.214	.522	.980	2.325	.042
	Xsquared	-.001	.025	-.020	-.048	.963

a. Dependent Variable: Y

Figure 2.9 Regression Coefficients for Prediction of Y From Q and Q^2

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-4.350	3.970		-1.096	.299
	Q	3.552	.873	3.134	4.070	.002
	Qsquared	-.173	.041	-3.234	-4.199	.002

a. Dependent Variable: Y

2.9 EVALUATION OF OTHER ASSUMPTIONS SPECIFIC TO ANALYSES

Many analyses require additional evaluation of assumptions in addition to these preliminary assessments. In the past, you have seen that tests of homogeneity of variance were applied for independent-samples *t* tests and between-*S* ANOVA. Often, as pointed out by Field (2018), assumptions for different analyses are quite similar. For example, homogeneity of variance assumptions can be evaluated for the independent-samples *t* test, ANOVA, and regression. For advanced analyses such as multivariate analysis of variance, additional assumptions need to be evaluated. Additional screening requirements for new analyses are discussed when these analyses are introduced.

2.10 DESCRIBING AMOUNT OF MISSING DATA

2.10.1 Why Missing Values Create Problems

There are several reasons why missing data are problematic. Obviously, if your sample is small, missing values make the amount of information even smaller. There is a more subtle problem. Often, missing responses don't occur randomly. For example, people who are overweight may be more likely to skip questions about body weight. SPSS listwise deletion, the default method of handling missing data, just throws out the persons who did not answer this question. If you focus just on the subset of people who did answer a question, you may be looking at a different kind of sample (probably biased) than the original set of people recruited for the study.

Blank cells are often used to represent missing responses in SPSS data files. (Some archival data files use specific numerical values such as 99 or 77 to represent missing responses.)

SPSS does not treat these blanks as 0 when computing statistics such as means; it omits the cases with missing scores from computation. For many procedures there are two SPSS methods for handling missing values. Consider this situation: A researcher asks for correlations among all variables in this list: $X_1, X_2, \ldots, X_k$. If **listwise deletion** is chosen, then only the cases with valid scores for all of the X variables on the list are used when these correlations are calculated. If **pairwise deletion** is chosen, then each correlation (e.g., $r_{12}, r_{13,} r_{23}$) is computed using all the persons who have valid scores for that pair of variables. When listwise deletion is used, all correlations are based on the same N of cases. When pairwise deletion is used, if there are missing values, the N's for different correlations will vary, and some of the N's may be larger than the N reported using listwise deletion.

If the amount of missing data is less than 5%, use of listwise deletion may not cause serious problems (Graham, 2009). When the amount of missing data is larger, listwise deletion can yield a biased sample. For example, if students with low grades are dropped from a sample used in the analysis because they refused to answer some questions about grades, the remaining sample will mostly include students with higher grades. The sample will be biased and will not represent responses from students with lower grades.

I'll add another caution here. If you pay no attention to missing values, and you do a series of analyses with different variables, the total N will vary. For example, in your table of descriptive statistics, you may have 100 cases when you report M and SD for many variables. In a subsequent regression analysis, you may have only 85 cases. In an ANOVA, you might have only 50 cases. Readers are likely to wonder why N keeps changing. In addition, results can't be compared across these analyses because they are not based on data for the same set of cases. It is better to deal with the problem of missing values at the beginning and then work with the same set of cases in all subsequent analyses.

Missing value analysis involves two steps. First, we need to evaluate the amount and pattern of missing data. Then, missing values may be replaced with plausible scores prior to other analyses.

To illustrate procedures used with missing data, I used a subset of data obtained in a study by Warner and Vroman (2011). A subset of 240 cases and six variables with complete data

Figure 2.10 Number of Missing Values for Each of Six Variables (in Data File missingwb.sav)

Statistics

		Depression	Satisfactionw Life	NegativeAffect	Neuroticism	Sex	Socialdesirability
N	Valid	218	150	226	220	240	240
	Missing	22	90	14	20	0	0

was selected and saved in a file named nonmissingwb.sav. To create a corresponding file with specific patterns of missingness, I changed selected scores in this file to system missing and saved these data in the file named missingwb.sav.

2.10.2 Assessing Amount of Missingness Using SPSS Base

Initial assessments of amount of missing data do not require the SPSS Missing Values add-on module. Amount of missing data can be summarized three ways: for each variable, for the entire data set, and for each case or participant. To make an initial assessment, the SPSS frequencies was used; results appear in Figure 2.10.

For each variable: Four variables had some missing values; two variables did not have missing data (in other words, 4/6 = 66.7% of variables had at least one missing value). What number of cases (or percentage of values) were missing on each variable? This is also obtained from the frequencies procedure output in Figure 2.10. For example, out of 240 cases, depression had missing values on 22 cases (22/240 = 9.2%).

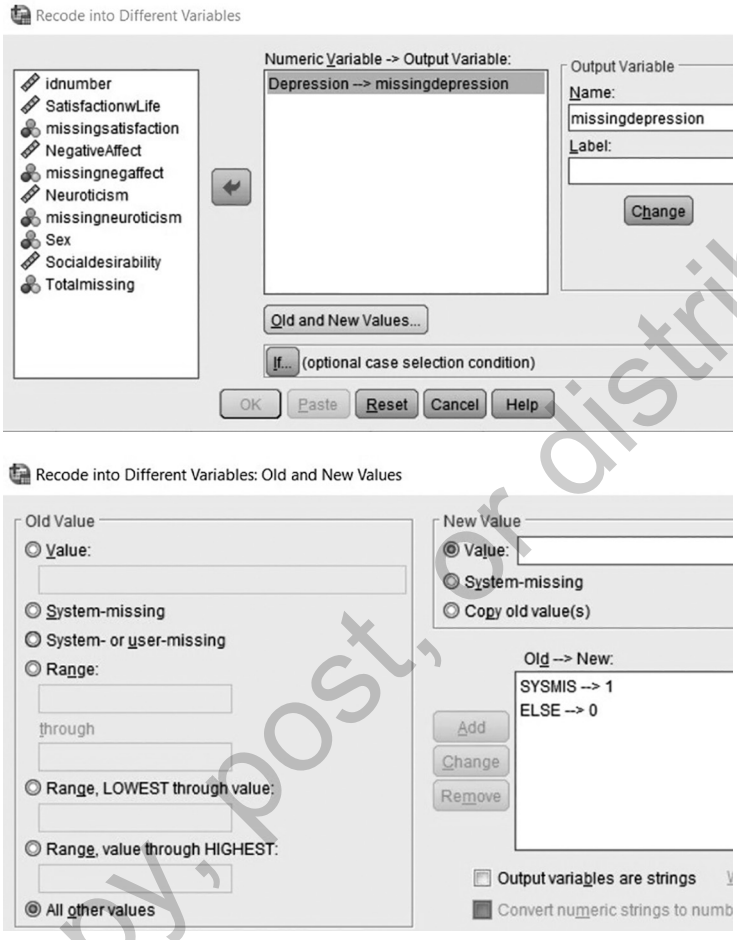
For the entire data set: Out of all possible values in the data set, what percentage were missing? The number of possible scores = number of variables × number of cases = 6 × 240 = 1,440. The number of missing values is obtained by summing the values in the "Missing" row in Figure 2.10: 22 + 90 + 14 + 20 + 0 + 0 = 146. Thus 146 of 1,440 scores are missing, for an overall missing data percentage of approximately 10%.

For each participant or case: Additional information is needed to evaluate the number of missing values for each case. To obtain this, create a dummy variable to represent missingness of scores on each variable (as suggested by Tabachnick & Fidell, 2018). The variable missingdepression corresponds to this yes/no question: Does the participant have a missing score on depression? Responses are coded 0 = no, 1 = yes. Dummy variables for missingness were created using the <Transform> → <Recode into Different Variables> procedure, as shown in Figure 2.11.

In the dialog box on top in Figure 2.11, specify the name of the existing (numerical) variable, in this example, depression. Create a name for the output variable in the right-hand side box (in this example, the output variable is named missingdepression). Click Change to move this new output variable name into the window under "Numeric Variable -> Output Variable." Then click Old and New Values. This leads to the second dialog box in Figure 2.11.

To define the first value of the dummy variable (a code of 1 if there is a missing value for depression), click the radio button to select the system missing value for depression as the old value; then enter the code for the new or output variable (1) into the "New Value" box on the right. Each participant who has a system missing value for depression is assigned a score of 1 on the new variable, missingdepression. Click Add to move this specification into the "Old --> New" box. To define the second value, select the radio button on the left for "All other values," and input 0 for "New Value" on the right; click Add. A participant with any other value, other than system missing, on depression is given a score of 0 on the new variable named missingdepression. Click Continue to return to the main dialog box, then click OK. The SPSS syntax that corresponds to these menu selections is:

Figure 2.11 Recode into Different Variables Dialog Box




```
RECODE Depression (SYSMIS=1) (ELSE=0) INTO missingdepression
EXECUTE
```

The same operations can be used to create missingness variables for other variables (NegativeAffect, SatisfactionwLife, and Neuroticism). To find out how many variables had missing values for each participant, sum these new variables:

```
COMPUTE Totalmissing = missingdepression + missingsatisfaction + missingnegaffect
+ missingneuroticism
```

Then obtain a frequencies table for the new variable Totalmissing (see Figure 2.12). Only one person was missing values on all three variables. Most cases or participants were missing values on no variables (n = 116) or only one variable (n = 103).

2.10.3 Decisions Based on Amount of Missing Data

Amount of Missing Data in Entire Data Set

Graham (2009) stated that it may be reasonable to ignore the problem of missing values if the overall amount of missing data is below 5%. When there are very few missing data, the

Figure 2.12 Numbers of Participants or Cases Missing 0, 1, 2, and 3 Scores Across All Variables

Totalmissing

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	116	48.3	48.3	48.3
	1	103	42.9	42.9	91.3
	2	20	8.3	8.3	99.6
	3	1	.4	.4	100.0
	Total	240	100.0	100.0	

use of listwise deletion may be acceptable. In listwise deletion, cases that are missing values for any of the variables in the analysis are completely excluded. For example, if you run correlations among X_1, X_2, X_3, and X_4 using listwise deletion, a case is excluded if it is missing a value on any one of these variables. Pairwise deletion means that a case is omitted only for correlations that require a score that the case is missing; for example, if a case is missing a score on X_1, then that case is excluded for computation of r_{12}, r_{13}, and r_{14}, but retained for r_{23}, r_{34}, and r_{24}. Listwise and pairwise deletion are regarded as unacceptable for large amounts of missing data. Even with less than 5% missing, Graham still recommended using missing values imputation (discussed in upcoming sections) instead of listwise deletion.

Amount of Missing Data for Each Variable

Tabachnick and Fidell (2018) suggested that if a variable is not crucial to the analysis, that variable might be entirely dropped if it has a high proportion of missing values. Suppose that prior to data analysis, the analyst decided to discard variables with more than 33% missing values. Satisfaction with life was missing 38% of its values; it might be dropped using this preestablished rule. If a variable has numerous missing values, this may have been information that was not obtainable for many cases. (If the missing value were planned missing, the variable would not be dropped. For example, if only smokers are asked additional questions about amount of smoking, these variables would not be dropped simply because nonsmokers did not answer the questions.) It is not acceptable to drop variables after final analyses; dropping variables that influence outcomes such as p values at a late stage in the analysis can be a form of p-hacking. Any decision to drop a variable must be well justified.

Amount of Missing Data for Each Case

Analysts might also consider dropping cases with high percentages of missing values (as suggested by Tabachnick & Fidell, 2018). Completely dropping cases is equivalent to listwise deletion, and experts on missing values agree that listwise deletion is generally poor practice. However, it's worth considering the possibility that some participants may have provided really poor data. Some possible examples of extremely low quality survey data include the following: no answers for many questions, ridiculous or impossible responses (height 10 ft or 3 m), a series of identical ratings given for a long list of questions that assess different things (e.g., a string of scores such as 5, 5, 5, 5, 5, 5, 5 . . .), and inconsistent responses across questions (e.g., person responds "I have never smoked" to one question and then responds "I smoke an average of 10 cigarettes per day" to another question). These problems can arise because of poorly worded questions, or they may be due to lack of participant attention and effort or deliberate refusal to cooperate. If a decision is made to omit entire cases on the basis of data quality, be careful how this decision is presented, and make it clear that case deletions were thoughtful decisions, not (mindless, automatic) listwise deletion. Ideally, specific criteria for case deletion would

be specified prior to data collection. However, participants can come up with types of poor data that are difficult to anticipate. In research other than surveys, analogous problems may arise.

The dummy variables used here to evaluate participant- or case-level missing data can also be used to evaluate patterns in missingness, as discussed in Section 2.13.

2.10.4 Assessment of Amount of Missingness Using SPSS Missing Values Add-On

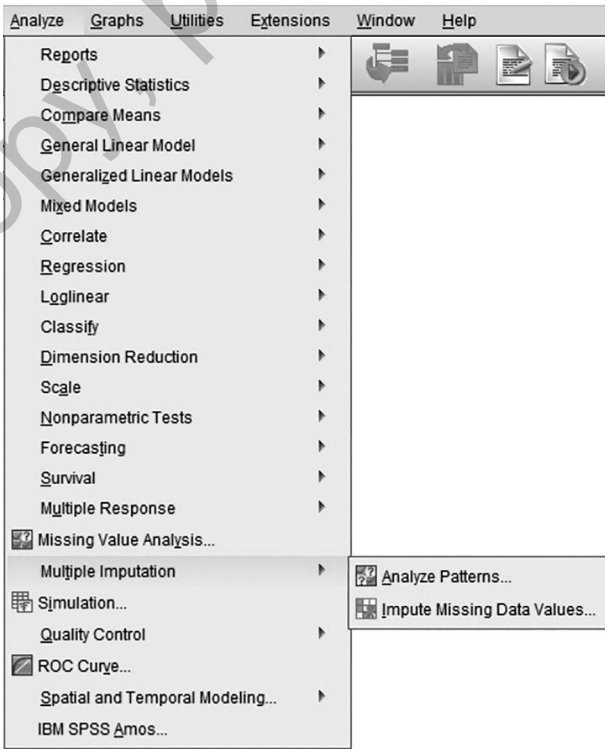
The SPSS Missing Values add-on module can be used to obtain similar information about amount of missing data in a different format (without the requirement to set up dummy variables for missingness.)

The SPSS Missing Values add-on module provides two different procedures for analysis and **imputation of missing values**. Unfortunately, the menu options for these (at least up until SPSS Version 26) are confusing. (You can locate SPSS manuals by searching for "SPSS Missing Values manual" and locating the manual for the version number you are using.)

When you purchase a license for the Missing Values add-on, two new choices appear in the pull-down menu under <Analyze>. The first choice can be obtained by selecting these menu options <Analyze> → <Missing Value Analysis>. I have not used this procedure in this chapter, and I do not recommend it. The procedure that corresponds to these menu selections has an important limitation; it does not provide multiple imputation (only single imputation). Multiple imputation is strongly preferred by experts.

For all subsequent missing value analysis, I used these menu selections: <Analyze> → <Multiple Imputation>, as shown in Figure 2.13. The pull-down menu that appears when

Figure 2.13 Drop-Down Menu Selections to Open SPSS Missing Values Add-On Module



you click <Multiple Imputation> offers two choices: <Analyze Patterns> and <Impute Missing Data Values>. The procedures demonstrated in this chapter are run using these two procedures. First, descriptive information about the amount and pattern of missing data is obtained using the menu selections <Analyze> → <Multiple Imputation> → <Analyze Patterns>. Then the menu selections <Analyze> → <Multiple Imputation> → <Impute Missing Data Values> are used to generate multiple imputation of missing score values.

To obtain information about the amount of missing data, make these menu selections: <Analyze> → <Multiple Imputation> → <Analyze Patterns>, as shown in Figure 2.13. (The <Multiple Imputation> command appears in the <Analyze> menu only if you or your organization has purchased a separate license; it is not available in SPSS Base.)

In the Analyze Patterns dialog box (Figure 2.14), checkboxes can be used to select the kinds of information requested. I suggest that you include all variables in the "Analyze Across Variables" pane, not only the ones that you know have missing values. ("Analyze patterns" is a bit of a misnomer here; the information provided by this procedure is mainly for the amount of missing values rather than patterns of missingness.)

Only one part of the output is shown here (Figure 2.15). Figure 2.15 tells us that four of six of the variables (66.67%) had at least one missing value. One hundred sixteen of 240 of cases or participants (48.33%) had at least one missing value. Of the 2,400 values in the entire data set, 146 or 10.14% were missing. These graphics present information already obtained from SPSS Base. The Missing Values add-on module also generates graphics to show the co-occurrence of pairs or sets of missing variables (e.g., how many cases were missing scores on both depression and sex?). However, more useful ways to assess patterns of missingness are discussed in Sections 2.12 and 2.13.

Figure 2.14 Dialog Box for Analyze Patterns Procedure

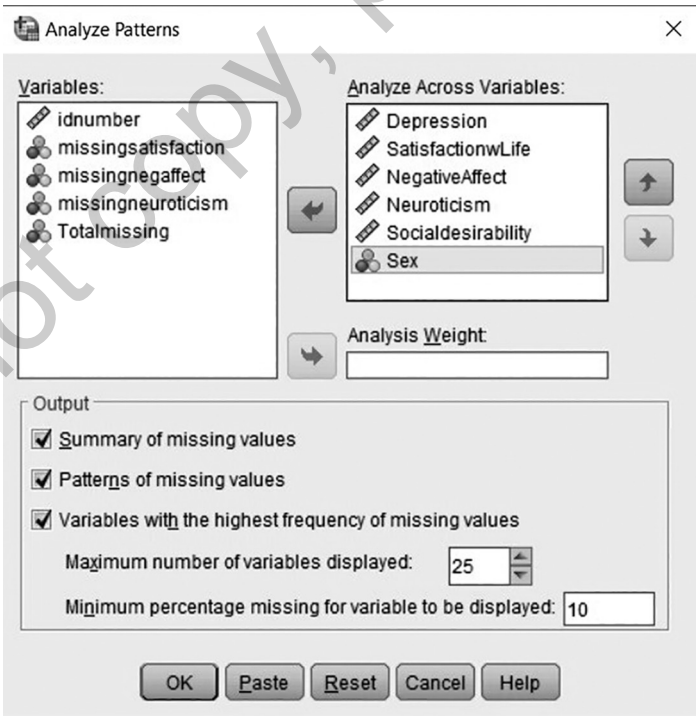
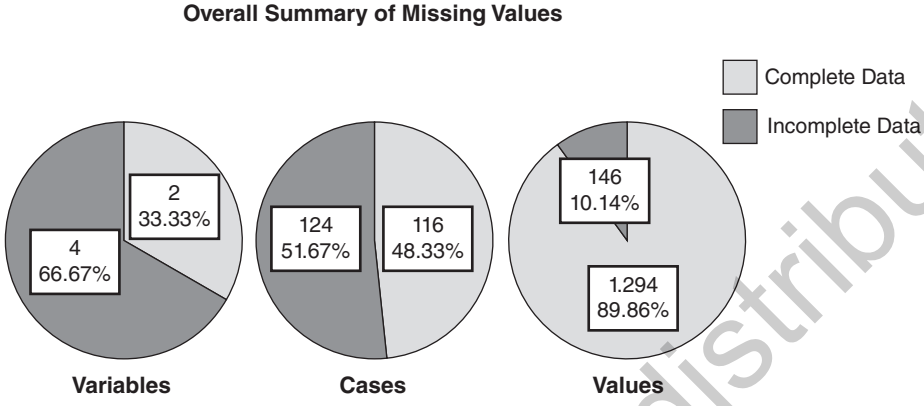


Figure 2.15 Selected Output From Missing Values Analyze Patterns Procedure



2.11 HOW MISSING DATA ARISE

Data can be missing for many reasons. Four common reasons are described; however, this does not exhaust the possibilities.

Refusal to participate: A researcher may initially contact 1,000 people to ask for survey participation. If only 333 agree to participate, no data are available for two thirds of the intended sample. Refusal to participate is unlikely to be random and can introduce substantial bias. There is nothing that can be done to replace this kind of missing data (the researcher could ask another 2,000 people to participate and obtain 666 more people). People who volunteer, or consent, to participate in research differ systematically from those who refuse (Rosenthal & Rosnow, 1975). It is essential to report numbers of person who refused to participate. It would also be useful to know why they refused. Refusal to participate leads to bias that cannot be corrected through later procedures such as imputation of missing values; imputation cannot replace this kind of lost data. The likelihood that the sample is not representative of the entire population that was contacted should be addressed in the discussion section when considering potential limitations of generalizability of results.

Attrition in longitudinal studies creates another kind of missingness. Imagine a longitudinal study in which participants are assigned (perhaps randomly) to different treatment conditions. Assessments may be made before treatment and at one or more times after the treatment or intervention. There is usually attrition. Participants may drop out of the treatment program, move and leave no contact information, die, or become unwilling or unable to continue. Some participants may miss one follow-up assessment and return for a later assessment. Samples after treatment or intervention can be smaller than the pretreatment sample, and they may also differ from the pretreatment sample in systematic ways.

Missing data may be planned: A survey might contain a funnel question, such as "Have you ever smoked?" People who say "yes" are directed to additional questions about smoking. People who say "no" would skip the additional smoking questions. Missing values would almost certainly not be imputed for these skipped questions. To shorten the time demands of a long survey, participants may be given only random subsets of the questions (and thus not have data for other questions, but in a planned and random manner). Development of better methods for handling planned missing data has encouraged the development of planned missing studies (Graham, 2009).

Missing values may have been used to replace outliers in previous data screening: One possible way to handle outliers (particularly when they are unbelievable or implausible) is to convert them to system missing values.

In an ideal situation, missing values would occur randomly, in ways that would not introduce bias in later data analysis. In actual data, missing values often occur in nonrandom patterns.

2.12 PATTERNS IN MISSING DATA

2.12.1 Type A and Type B Missingness

Patterns of missingness are usually described as one of these three types: missing completely at random, missing at random, and missing not at random (Rubin, 1976). To explain how these kinds of missingness differ, here is a distinction not found elsewhere in the missing values literature: I will refer to Type A and Type B missingness.

Consider **Type A missingness**. Suppose we have a Y variable (such as depression) that has missing values, and we also have data for other variables X_1, X_2, X_3, and so on (such as sex, neuroticism, and social desirability response bias). It is possible that missingness on Y is related to scores on one or more of the X variables; for example, men and people high in social desirability may be more likely to refuse to answer the depression questions than women and persons with low social desirability response bias. I will call this Type A missingness. The next few sections show that this kind of missingness can easily be detected and that state-of-the-art methods of replacement for missing values, such as **multiple imputation (MI)**, can correct for bias due to this type of missingness.

Now consider **Type B missingness**. It is conceivable that the likelihood of missing scores on Y (depression) depends on people's levels of depression. That is, people who would have had high scores on depression may be likely not to answer questions about depression. I will call this Type B missingness. Type B missingness is more difficult to identify than Type A missingness. (Sometimes it is impossible to identify Type B missingness.) Also, potential bias due to Type B missingness is more problematic and may not be correctable.

2.12.2 MCAR, MAR, and MNAR Missingness

The three patterns of missingness that appear widely in research on missing values were described by Rubin (1976). These are **missing completely at random (MCAR)**, **missing at random (MAR)**, and **missing not at random (MNAR)**. Each of these patterns can be defined by the presence or absence of Type A and Type B missingness.

First consider MCAR missingness, as described by Schafer and Graham (2002): Assume that "variables X ($X_1, \ldots X_p$) are known for all participants but Y is missing for some. If participants are independently sampled from the population . . . MCAR means that the probability that Y is missing for a participant does not depend on his or her own values of X or Y." Using the terms I suggest, MCAR does not have either Type A or Type B missingness.

The name MAR (missing at random) is somewhat confusing, because this pattern is not completely random. Schafer and Graham (2002) stated, "MAR means that the probability that Y is missing may depend on X but not Y . . . under MAR, there could be a relationship between missingness and Y induced by their mutual relationships to X, but there must be no residual relationship between them once X is taken into account." Using my terms, MAR may show Type A missingness (however, MAR must not show Type B missingness after corrections have been made for any Type A missingness).

The third and most troubling possible pattern is MNAR. Schafer and Graham (2002) stated that "MNAR means that the probability of missingness depends on Y. . . . Under MNAR, some residual dependence between missingness and Y remains after accounting for X." Using terms I suggest, MNAR has Type B missingness (and it may or may not also have Type A missingness).

MAR and MCAR patterns of missingness are called ignorable. This does not mean that we don't have to do anything about missing data if the pattern of missingness is judged to be MAR or MCAR. "Ignorable" means that, after state-of-the-art methods for replacement of missing values are used, results of analyses (such as *p* values) should not be biased.

MNAR (and Type B missingness, its distinguishing feature) are nonignorable forms of missingness. Even when state-of-the-art methods are used to impute scores for missing values in MNAR missing data, potential bias remains a problem that cannot be ignored. Discussion in a journal article must acknowledge the limitations imposed by this bias. For example, if we know that persons who are very depressed are likely to have missing data on the depression question, it follows that the people for whom we do have data represent a sample that is biased toward lower depression. Schlomer, Bauman, and Card (2010) urged researchers to consider the possible existence of MNAR and reasons why this might occur.

The degree to which missing values are problematic depends more on the pattern of missingness than the amount of missingness (Tabachnick & Fidell, 2018). MNAR is most problematic. Researchers should report information about pattern, as well as amount, of missing data. It is possible to find patterns in data that indicate problems with Type A missingness. However, it is impossible to prove that Type A and/or Type B missingness is absent.

2.12.3 Detection of Type A Missingness

Methods for detection of Type A missingness are discussed in the context of an empirical example in upcoming Section 2.13, including pairwise examination of variables and Little's test of MCAR. In this empirical example, Type A missingness occurs because missingness of depression scores is related to sex, neuroticism, socially desirable response bias, and other variables. The SPSS Missing Values add-on module provides all the necessary tests for Type A missingness. I will demonstrate that many of these tests can also be obtained using SPSS Base (the output from analysis using SPSS Base may be easier to understand). State-of-the-art methods for replacement of missing values are thought to correct most of the bias due to this type of missingness (Graham, 2009).

2.12.4 Detection of Type B Missingness

Unfortunately, evaluation of Type B missingness is difficult. It usually requires information that researchers don't have. Consider this example. If a question about school grade point average (GPA) is included in a survey, it is possible that students are more likely not to answer this if they have low GPAs. To evaluate whether Type B missingness is occurring, we need to know what the GPA scores would have been for the people who did not answer the question. Often there is no way to obtain this kind of information. In some situations, outside information can be helpful. Here are three examples of additional information that would help evaluate whether Type B missingness is occurring.

1. The researcher could follow up with the students who did not answer the GPA question and try again to obtain their answers. (Of course, if that information is obtained, it can be used to replace the missing value.)
2. The researcher could look for an independent source of data to find out what GPA answers would have been for people who did not answer the question. For example, universities have archival computer records of GPA data for all students. (Usually researchers cannot access this information.) If the researcher could obtain GPAs for all students, he or she could evaluate whether students who did not answer the question about GPA had lower GPA values than people who did answer the question. In this situation also, the values from archival data could be used to replace missing values in the self-report data.

3. An indirect way to assess Type B missingness would be to look at the distribution and range of GPA values in the sample of students and compare that with the distribution and range of GPA values for the entire university. Assume that the sample was drawn randomly from all students at the university. If the sample distribution for GPA contains a much lower proportion of GPAs below 2.0 than the university distribution, this would suggest that low-GPA students may have been less likely to report their GPAs than high-GPA students. This would indicate the presence of Type B missingness but would not provide a solution for it.

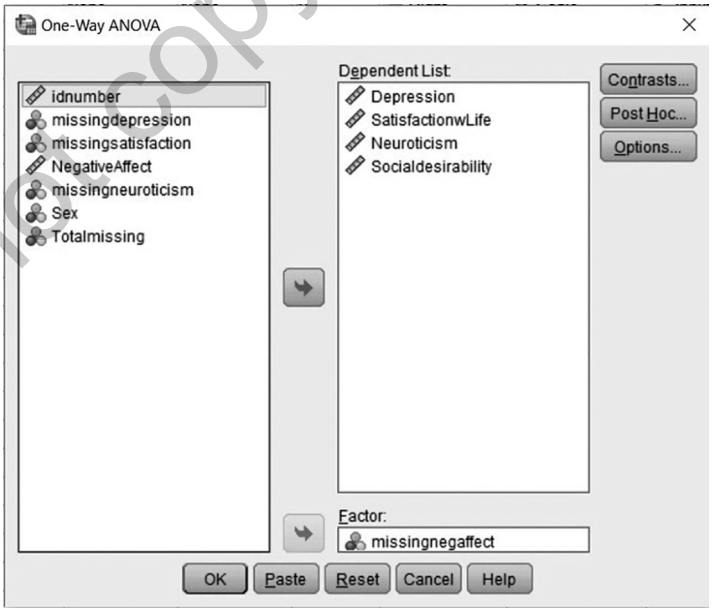
In the data set used as an empirical example, I know that neuroticism had Type B missingness (because, when I created my missing data file, I systematically turned higher scores on neuroticism into missing values). When I created Type B missingness for neuroticism, my new missing data file underrepresented people high in neuroticism, compared with the complete data set. Even after replacement of values using methods such as MI, generalization of findings to persons high on neuroticism would be problematic in this example.

Researchers often cannot identify, or correct for, Type B missingness. When Type B missingness is present (and probably it often is), researchers need to understand the bias this creates. Two types of bias may occur: Parameters may be over- or underestimated, and the sample may not be representative of, or similar to, the original population of interest. (For example, the sample may underrepresent certain types of persons, such as those highest on depression.) A researcher should address these problems and limitations in discussion of the study.

2.13 EMPIRICAL EXAMPLE: DETECTING TYPE A MISSINGNESS

To assess Type A missingness, we need to know whether missing versus nonmissing status for each variable is related to scores on other variables. This information can be obtained using the SPSS Missing Values add-on module. However, when first learning about missing values, doing a similar analysis in SPSS Base may make the underlying ideas clearer.

Figure 2.16 One-Way ANOVA Dialog Box: Assess Associations of Other Variables With Missingness on Negative Affect



Earlier, in Section 2.10, a dummy "missingness" variable was created for each variable in the data set that had one or more missing values. These dummy variables can now be used to test Type A missingness. To see whether missingness on one variable (such as negative affect) is related to scores on other quantitative variables (such as response bias, negative affect, or neuroticism), means for those other quantitative variables are tested to see if they differ across the missing and nonmissing groups. It is convenient to use the SPSS one-way ANOVA procedure for comparison of means. To open the one-way ANOVA procedure, make the following menu selections: <Analyze> → <Compare Means> → <One-Way ANOVA>. The One-Way ANOVA dialog box in Figure 2.16 shows which variables were included. The Options button was used to select descriptive statistics (recall that means and other descriptive statistics are not provided unless requested explicitly). Selected results appear in Figure 2.17.

The groups (groups of persons missing or not missing negative affect scores) did not differ in mean satisfaction with life, $F(1, 148) = .106$, $p = .745$. Missingness on negative affect was related to scores on the other three variables; in other words, there is evidence of Type A missingness. The table of group means (not shown here) indicated that people in the missing negative affect group scored lower on neuroticism, higher in social desirability response bias, and lower on depression. Similar comparisons of means are needed for each of the other missingness dummy variables (e.g., ANOVAs to compare groups of missing vs. not missing status for Depression, SatisfactionwLife, etc.).

To evaluate whether missingness is related to a categorical variable such as sex, or to missingness on other variables, set up a contingency table using the SPSS crosstabs procedure.

The crosstabs results in Figure 2.18 indicate that sex was associated with missingness on depression; 22 of 112 men (almost 20%) of men were missing scores on depression; none of the women were missing depression scores. This was statistically significant, $\chi^2(1) = 27.68$, $p < .001$ (output not shown).

The SPSS Missing Values add-on module provides similar comparisons of group means and crosstabs (not shown here). An additional test available from the Missing Values add-on module is **Little's test of MCAR** (Little, 1988). Little's test essentially summarizes information from the individual tests for Type A missingness just described.

To obtain Little's test, open the Missing Values add-on module by selecting <Analyze> → <Missing Value Analysis> (not either of the two additional menu choices that appear to the right after selecting <Multiple Imputation>; refer back to Figure 2.13). The Missing Value Analysis dialog box appears as shown in Figure 2.19.

Figure 2.17 ANOVA Source Table: Comparison of Groups Missing Versus Not Missing Negative Affect Scores

ANOVA

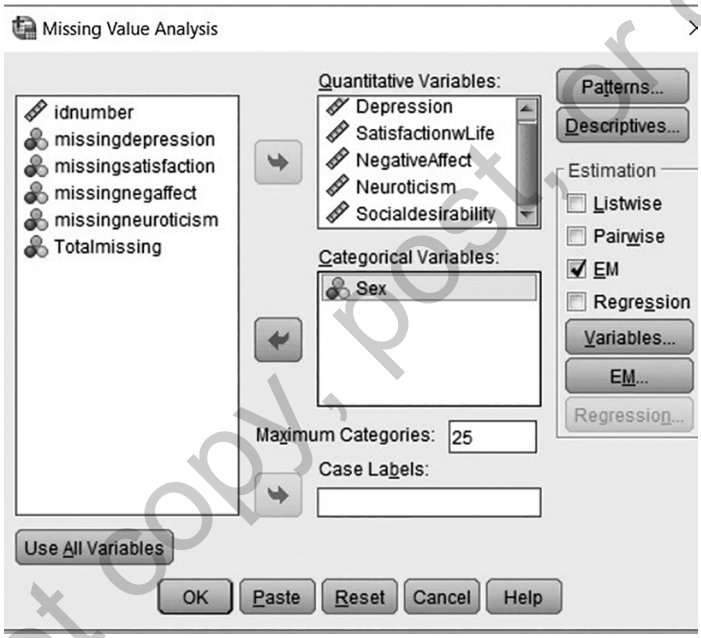
		Sum of Squares	df	Mean Square	F	Sig.
Depression	Between Groups	690.064	1	690.064	6.310	.013
	Within Groups	23620.840	216	109.356		
	Total	24310.904	217			
SatisfactionwLife	Between Groups	1.263	1	1.263	.106	.745
	Within Groups	1760.505	148	11.895		
	Total	1761.768	149			
Neuroticism	Between Groups	201.893	1	201.893	5.991	.015
	Within Groups	7346.884	218	33.701		
	Total	7548.777	219			
Socialdesirability	Between Groups	686.018	1	686.018	104.892	.000
	Within Groups	1556.582	238	6.540		
	Total	2242.600	239			

Figure 2.18 Contingency Table for Missingness on Depression by Sex

Sex * missingdepression Crosstabulation

			missingdepression		
			0	1	Total
Sex	male	Count	90	22	112
		% within Sex	80.4%	19.6%	100.0%
	female	Count	128	0	128
		% within Sex	100.0%	0.0%	100.0%
Total		Count	218	22	240
		% within Sex	90.8%	9.2%	100.0%

Figure 2.19 Missing Value Analysis Dialog Box to Request Little's MCAR Test



In the Missing Value Analysis dialog box, move all quantitative variables to the "Quantitative Variables" pane, and move any categorical variables into the separate "Categorical Variables" pane. Check the box for "EM" in the "Estimation" list, then click OK. (If you also want the *t* tests and crosstabs that were discussed earlier in SPSS Base, click the Descriptives button and use checkboxes in the Descriptives dialog box to request these; they are not included here.)

Little's MCAR test appears as a footnote to the "EM Means" table in Figure 2.20. This was statistically significant, $\chi^2(33) = 136.081$, $p < .001$. The null hypothesis is essentially that there is no Type A missingness for the entire set of variables. This null hypothesis is rejected (consistent with earlier ANOVA and crosstabs results showing that missingness was related to scores on other variables). This is additional evidence that Type A missingness is present. There is no similar empirical test for Type B missingness.

Figure 2.20 "EM Means" Table From SPSS Missing Values Analysis With Little's MCAR Test

EM Means[a]

Depression	SatisfactionwLife	NegativeAffect	Neuroticism	Socialdesirability
22.35	17.40	24.39	30.50	9.85

a. Little's MCAR test: Chi-Square = 136.081, DF = 33, Sig. = .000

2.14 POSSIBLE REMEDIES FOR MISSING DATA

There are essentially three ways to handle missing values. The first is to ignore them, that is, throw out cases with missing data using default methods such as SPSS listwise or pairwise deletion. (Somewhat different terms are used elsewhere; "complete case analysis" is synonymous with listwise deletion; "available data analysis" is equivalent to pairwise deletion; Pigott, 2001.) Listwise deletion is almost universally regarded as bad practice. However, Graham (2009) said that listwise deletion may yield acceptable results if the overall amount of missing data is less than 5%; he stated that "it would be unreasonable for a critic to argue that it was a bad idea" if an analyst chose to use listwise deletion in this situation. However, he recommended the use of missing data replacement methods such as MI even when there is less than 5% missing data.

One obvious problem with listwise deletion is reduction of statistical power because of a smaller sample size. A less obvious but more serious problem with listwise deletion is that discarding cases with missing scores can systematically change the composition of the sample. Recall that when I created a missing values pattern in the data set used as an example, I systematically deleted the cases with the highest scores on neuroticism (this created Type B missingness for neuroticism). If listwise deletion were used, subsequent analyses would not include any information about people who had the highest scores for neuroticism. That creates bias in two senses. First, if we want to generalize results from a sample to some larger hypothetical population, the sample now underrepresents some kinds of people in the population, Second, there is bias in estimation of statistics such as regression slopes, effect sizes, and *p* values (this is known from Monte Carlo studies that compared different methods for handling missing values in the presence of different types of pattern for missingness).

A second way to handle missing values is to replace them with simple estimates based on information in the data set. Missing scores on a variable could be replaced with the mean of that variable (for the entire data set or separately for each group). Missing values could be replaced with predicted scores from a regression analysis that uses other variables in the data set as predictors. These methods are not recommended (Acock, 2005), because they do not effectively reduce bias.

There are several state-of-the-art methods for replacement of missing values that involve more complex methods. Graham (2009) "fully endorses" multiple imputation. Monte Carlo work shows that MI is effective in reducing bias in many missing-values situations (but note that it cannot correct for bias due to Type B missingness). Graham and Schlomer et al. (2010) described other state-of-the-art procedures and the capabilities of several programs, including SAS, SPSS, Mplus, and others. They also described freely downloadable software for missing values.

The empirical example presented in the following section uses MI. Graham (2009) stated that MI performs well in samples as small as 50 (even with up to 18 predictors) and with as much as 50% missing data in the dependent variable. He explained that, contrary to some beliefs, it is acceptable to impute replacements for missing values on dependent variables. He suggested that a larger number of imputations than the SPSS default of 5 may be needed with larger amounts of missing data, possibly as many as 40 imputations.

2.15 EMPIRICAL EXAMPLE: MULTIPLE IMPUTATION TO REPLACE MISSING VALUES

To run MI using the SPSS Missing Values add-on module, start from the top-level menu. Choose <Analyze> → <Multiple Imputation>, then from the pop-up menu on the right, select <Impute Missing Data Values>. The resulting dialog box appears in Figure 2.21. All the variables of interest (both the variables with missing values and all other variables that will be used in later analyses) are included. Note that you can access a list of procedures that can be applied to imputed data in SPSS Help, as noted in this dialog box.

The number of imputations is set to 5 by default (note that a larger number of imputations, on the order of 40, is preferable for data sets with large percentages of missing data; Graham, 2009). A name for the newly created data set must be provided (in this example, Imputed Data).

MI does something comparable with replacement by regression. Each imputation estimates a different set of plausible values to replace each missing value for a variable such as depression; these plausible values are based on predictions from all other variables. The resulting data file (a subset appears in Figure 2.22) now contains six versions of the data: the original data and the five imputed versions. The first column indicates imputation number (0 for the original data).

Figure 2.21 Dialog Box for Impute Missing Data Values

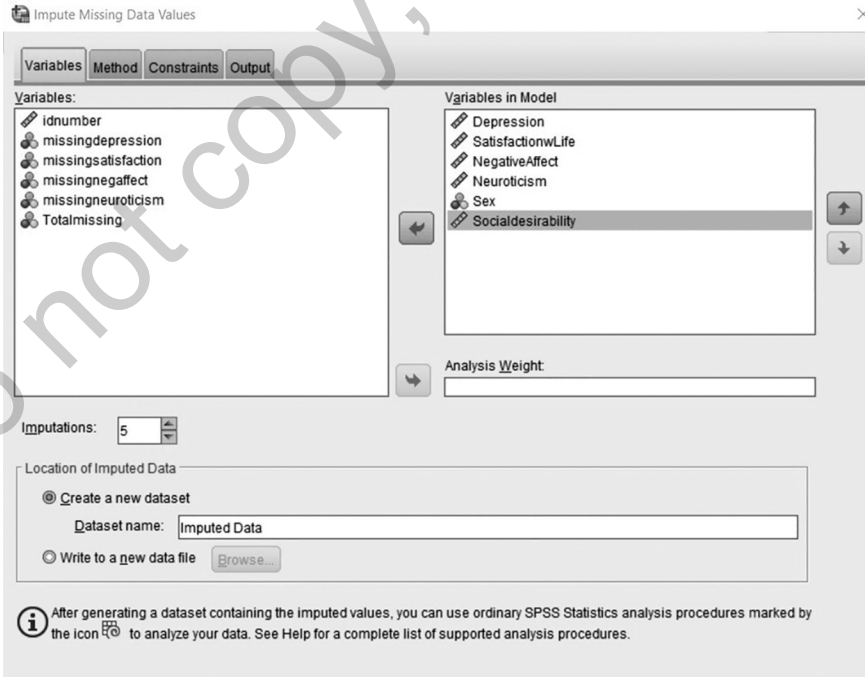
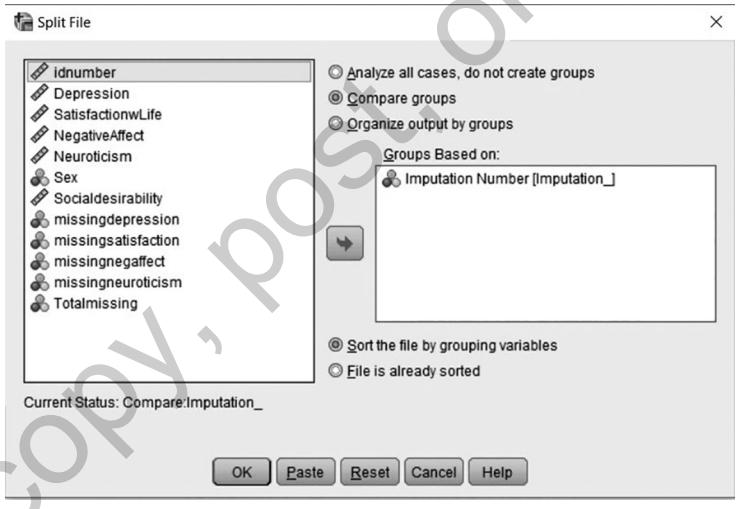


Figure 2.22 Selected Rows From Imputed Data Set

	Imputation_	idnumber	Depression	Satisfaction wLife	Negative Affect	Neuroticism	Sex	Socialdesirabilit
235	0	235	18	16	.	26	1	1
236	0	236	.	20	18	23	1	1
237	0	237	0	20	10	24	1	1
238	0	238	26	18	37	37	2	1
239	0	239	30	9	32	26	2	
240	0	240	20	15	20	26	2	
241	1	1	30	18	24	32	2	
242	1	2	30	14	24	38	1	1
243	1	3	24	22	31	34	1	
244	1	4	44	17	31	35	2	
245	1	5	14	18	15	39	2	1
246	1	6	16	23	17	23	1	

Figure 2.23 Split File Command Used to Pool Results for Imputed Data File



The final analysis of interest (for example, prediction of depression from the other five variables) is now run on all versions of the data (Imputations 0 through 5), and results are pooled (averaged) across data sets. Prior to the regression, select <Data> and <Split File> (not <Split into Files>).

In the Split File dialog box, move the variable Imputation Number into the pane under "Groups Based on" and select the radio button for "Compare groups." You should see a line that says Current Status: Compare:Imputation_ in the lower left corner. The SPSS syntax is: SPLIT FILE LAYERED BY Imputation_.

Now run the analysis of interest. In this example, it was a multiple regression to predict scores on Depression from SatisfactionwLife, NegativeAffect, Neuroticism, Sex, and Socialdesirability. Selected results for this regression analysis appear in Figure 2.24.

Figure 2.24 shows the regression coefficients (to predict Depression from SatisfactionwLife, NegativeAffect, Neuroticism, Sex, and Socialdesirability), separately for the original data, for each of the imputed data sets (1 through 5), and for the pooled results. We hope to see consistent results across all solutions, and that is usually what is obtained. For these data, results varied little across the five imputations. Reporting would focus on pooled coefficient estimates and the overall statistical significance of the regressions (in the ANOVA tables, not provided here).

Figure 2.24 Prediction of Depression From SatisfactionwLife, NegativeAffect, Neuroticism, Sex, and Socialdesirability Using Linear Regression: Original and Imputed Missing Values

Coefficients[a]

Imputation Number	Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
Original data	1	(Constant)	14.365	6.686		2.149	.034
		SatisfactionwLife	-.832	.208	-.272	-4.003	.000
		NegativeAffect	.709	.119	.460	5.941	.000
		Neuroticism	.380	.139	.207	2.730	.007
		Sex	-1.840	1.362	-.090	-1.352	.179
		Socialdesirability	-.393	.298	-.091	-1.322	.189
1	1	(Constant)	14.954	4.441		3.367	.001
		SatisfactionwLife	-.780	.130	-.257	-5.981	.000
		NegativeAffect	.676	.073	.485	9.270	.000
		Neuroticism	.387	.089	.225	4.364	.000
		Sex	-2.295	.897	-.109	-2.559	.011
		Socialdesirability	-.348	.154	-.101	-2.263	.025
Pooled	1	(Constant)	12.420	4.888		2.541	.013
		SatisfactionwLife	-.851	.184		-4.625	.000
		NegativeAffect	.665	.090		7.397	.000
		Neuroticism	.428	.100		4.294	.000
		Sex	-1.442	1.105		-1.305	.200
		Socialdesirability	-.221	.185		-1.198	.238

a. Dependent Variable: Depression

2.16 DATA SCREENING CHECKLIST

Decisions about eligibility criteria, minimum group size, methods to handle outliers, plans for handling missing data, and so forth should be made prior to data collection. For longitudinal studies that compare treatment groups, **Consolidated Standards of Reporting Trials (CONSORT)** guidelines may be helpful (Boutron, 2017). Document what was done (with justification) at every step of the data-screening process. The following checklist for data screening and handling covers many research situations.

Some variation in the order of steps is possible. However, I believe that it makes sense to consider distribution shape prior to making decisions about handling outliers and to deal with outliers before imputing missing values. These suggestions are not engraved in stone. There are reasonable alternatives for most of the choices I have recommended.

1. Proofread the data set against original sources of data (if available). Replace incorrect scores with accurate data. Replace impossible score values with system missing.
2. Remove cases that do not meet eligibility criteria.
3. If group means will be compared, each group should have a minimum of 25 to 30 cases (Boneau, 1960). If some groups have smaller *n*'s, additional members for these groups might be obtained prior to other data analyses. Alternatively, groups with small *n*'s can be dropped, or combined with other groups (if that makes sense).
4. Assess distribution shapes by examining histograms. If groups will be compared, distribution shape should be assessed separately within each group. Some distribution shapes, such as Poisson, require different analyses than those covered in this book (see Appendix 2A).
5. Possibly apply data transformations (such as log or arcsine), but only if this makes sense. If distribution skewness is due to a few outliers, it may be preferable to deal with those outliers individually instead of transforming the entire set of scores.

6. Screen for univariate, bivariate, and multivariate outliers. Decide how to handle these (for example, convert extreme scores to less extreme values, or replace them with missing values).
7. Test linearity assumptions for associations between quantitative variables. If nonlinearity is detected, revisit the possibility of data transformations, or include terms such as X^2 in later analyses.
8. Assess amount and pattern of missing values. If there is greater than 50% missingness on a case or a variable, consider the possibility that these cases or variables provide such poor-quality data that they cannot be used. If cases or variables are dropped, this should be documented and explained.
9. Use multiple imputation to replace missing values (or use another state-of-the-art missing value replacement method, as discussed in Graham, 2009).

2.17 REPORTING GUIDELINES

At a minimum, the following questions should be answered. Some may require only a sentence or two; others may require more information. For additional suggestions about reporting, see Johnson and Young (2011), Recommendations 9 and 10, and Manly and Wells (2015).

In the "Introduction": What types of analyses were done and why were these chosen?

In the "Methods" section: Details about initial sample selection, measurements, group comparisons (if any), and other aspects of procedure.

In the "Results" section: Data screening and handling procedures should be described at the beginning of the "Results" section. This should address each of the following questions:

1. What were the final numbers of cases for final analysis, after any respondents were dropped because they declined to participate or did not meet eligibility criteria (or presented other problems)? For longitudinal studies, a CONSORT flowchart may be helpful (see Section 2.1).
2. If any variables were dropped from planned analyses because of poor measurement quality or if groups were omitted or combined because of small *n*'s, explain.
3. Explain rule(s) for outlier detection and the way outliers were handled, and note how many changes were made during outlier evaluation. Explain any data transformations.
4. Report the amount of missing values, such as the percentage of scores missing in the entire data set, the percentage missing for each variable, or the percentage of participants missing one or more scores.
5. Describe possible reasons for missing values.
6. Explain pattern in missing values. Type A missingness is present if Little's MCAR test is significant; details about the nature of missingness are found in the *t* tests and crosstabs that show how missingness dummy variables are related to other variables. It may not be possible to detect Type B missingness unless additional information is available beyond the data set; this possibility should be discussed. (Type A missingness is ignorable; Type B is problematic.)
7. Provide specific information about the imputation method used to replace missing values, including software, version, and commands; number of imputations; and any notable differences among results for different imputations and original data.

In the "Discussion" section: Be sure to explain the ways in which data problems, such as sample selection and missing values, may have (a) created bias in parameter estimates and (b) limited the generalizability of results.

2.18 SUMMARY

Before collecting data, researchers should decide on rules and procedures for data screening, outliers, and missing values, and then adhere to those rules. This information is required for preregistration of study plans. Open Science advocates preregistration as a way to improve completeness and transparency of reporting and calls for making data available for examination by other researches through publicly available data archives. Some journals offer special badges for papers that report preregistered studies. For further discussion, see Asendorpf et al. (2013) and Cumming and Calin-Jageman (2016).

In addition, professional researchers often seek research funding from federal grant agencies (e.g., the National Science Foundation, the National Institutes of Health). These agencies now require detailed plans for data handling in the proposals, for example, decisions about sample size on the basis of statistical power analysis, plans for identification and handling of outliers, and plans for management of missing data. A few professional journals (for example, *Psychological Science* and some medical journals) provide the opportunity to preregister detailed plans for studies including this information. Journal editors are beginning to require greater detail and transparency in reporting data screening than in the past. The requirement for detailed reporting of data handling is likely to increase.

For many decisions about outliers and missing value replacement, there is no one best option. This chapter suggests several options for handling outliers, but there are many others (Aguinas et al., 2013). This chapter describes the use of MI for replacement of missing values, but additional methods are available or may become available in the future.

The growing literature about missing values includes strong arguments for the use of MI and other state-of-the-art methods as ways to reduce bias. However, even state-of-the-art methods for replacement of missing values does not get rid of problems due to Type B missingness.

It is important to remember that many other common research practices may be even greater sources of bias. Use of convenience samples rather than random or representative samples limits the generalizability of findings. Practices such as *p*-hacking and hypothesizing after results are known to greatly inflate the risk for Type I error. Quality control during data collection is essential. Nothing that is done during data screening can make up for problems due to poor-quality data.

Numerous missing values situations are beyond the scope of this chapter, for example, imputation of missing values for categorical variables (Allison, 2002), attrition in longitudinal studies (Kristman, Manno, & Côté, 2005; Muthén, Asparouhov, Hunter, & Leuchter, 2011; Twisk & de Vente, 2002), missing data in multilevel or structural equation models, and missing values at the item level in research that uses multiple-item questionnaires to assess constructs such as depression (Parent, 2013).

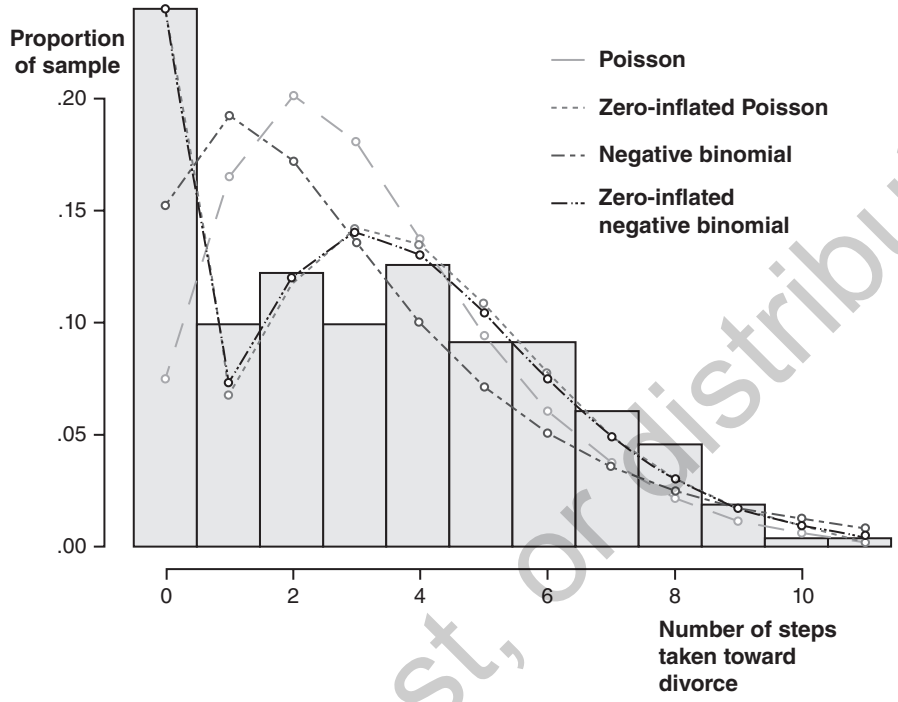
Subsequent chapters assume that all appropriate data screening for generally required assumptions has been carried out. Additional data-screening procedures required for specific analyses will be introduced as needed.

APPENDIX 2A

Brief Note About Zero-Inflated Binomial or Poisson Regression

The following empirical example provides an illustration. Figure 2.25 is adapted from Atkins and Gallop (2007). The count variable in their study (on the *X* axis) is the number of steps each person has taken toward divorce, ranging from 0 to 10. The distribution is clearly non-normal; it has a mode at zero and positive skewness (a very small proportion of persons in the sample had taken 8 or more steps).

Figure 2.25 Four Models for Distribution Shape of Frequency Count Variable



Source: Adapted from Atkins and Gallop (2007).

Note: Variable on the *X* axis is the number of steps or actions taken toward separation or divorce, ranging from 0 to 10.

Atkins and Gallop (2007) evaluated the fit of four mathematical distribution models to the empirical frequency distribution in Figure 2.25: Poisson, zero-inflated Poisson (ZIP), negative binomial, and zero-inflated negative binomial (ZINB). Quantitative criteria were used to evaluate model fit. They concluded that the ZIP model was the best fit for their data (results were very similar for the ZINB model). The regression analysis to predict number of steps toward divorce from other variables would be called zero-inflated Poisson regression; this is very different from linear regression.

It is possible to ask two questions about analyses in these models applied to behavior count variables. Consider illegal drug use as an example (e.g., Wagner, Riggs, & Mikulich-Gilbertson, 2015). First, we want to predict whether individuals use drugs or not. For those who do use drugs, a zero frequency of drug use in the past month is possible, but higher frequencies of use behaviors can occur. The set of variables that predicts frequency of drug use in this group may differ from the variables that predict use versus nonuse of drugs. This information would be missed if a data analyst applied ordinary linear regression.

The SPSS generalized linear models procedure can handle behavior count dependent variables. (Note that this is different from the GLM procedure used in Volume I [Warner, 2020].) For an online SPSS tutorial, see UCLA Institute for Digital Research & Education (2019). Atkins and Gallop (2007) provided extensive online supplemental material for their study. Note that count data should not be log transformed in an attempt to make them more nearly normally distributed (O'Hara & Kotze, 2010).

COMPREHENSION QUESTIONS

1. What can you look for in a histogram for scores on a quantitative variable?
2. What can you look for in a three-dimensional scatterplot?
3. What quantitative rule can be used to decide whether a univariate score is an outlier?
4. Are there situations in which can you justify deleting a case or participant completely? If so, what are they?
5. Under what conditions might you convert a score to system missing?
6. What is the point of running an analysis once with outliers included and once with outliers deleted?
7. What is a way to identify multivariate outliers using Mahalanobis distance?
8. Describe two distribution shapes (other than normal) that you might see in actual data (hint: any other distribution graph you have seen in this chapter, along with any strange things you might have seen in other data).
9. When can log transformations be used, and what potential benefits do these have? When should log transformations not be used?
10. If you have a dependent variable that represents a count of some behavior, would you expect data to be normally distributed? Why or why not? What types of distribution better describe this type of data? Can you use linear regression? What type of analysis would be preferable?
11. Which do authorities believe generally pose more serious problems in analysis: outliers or non-normal distribution shapes?
12. What problems arise when listwise deletion is used to handle missing values?

NOTE

[1]Chapter 7, on moderation, explains that when forming products between predictor variables, the correlation between X^2 and X can be reduced by using centered scores on X to compute the squared term. A variable is centered by subtracting out its mean. In other words, we can calculate $X^2 = (X - M_X) \times (X - M_X)$ where M_X is the mean of X. The significance of the quadratic trend is the same whether X is centered or not; however, judgments about whether there could also be a significant linear trend can change depending on whether X was centered before computing X^2.

DIGITAL RESOURCES

Find **free study tools** to support your learning, including **eFlashcards, data sets, and web resources,** on the accompanying website at **edge.sagepub.com/warner3e.**