CHAPTER 8

How Do I Deal With Missing Values, Outliers, and Non-normality?

Chapter 8 is a response to three frequently asked questions about data. The first section discusses the implications of missing data. What do you do when some items of information are unobtainable or are mysteriously missing from your data pool? There is a certain amount of controversy about the treatment of missing data, and the answers range from the simple to the complex. The second section confronts the issue of outliers, those data points or responses from research participants that seem to lie outside the range of the bulk of the data. Here, too, many solutions are possible. Our recommendations rest on making an adequate assessment of the nature and seriousness of the problem. Finally, we address the question of what to do with data that are not normally distributed. This issue was broached initially in Chapter 5 in the context of the assumptions that provide the foundation for statistical testing. In practice, however, data do not always behave as they do in theory. When violations of the assumption of normality warrant special procedures, the researcher needs to know what to do. In this chapter, we offer our suggestions.

HOW DO I DEAL WITH MISSING VALUES?

As we gather data for each case in a research endeavor, some items of information almost certainly will be unobtainable. These incomplete or lost responses are referred to as **missing values** and result in an incomplete set of data for some cases. Values may be missing for numerous reasons. Following is a list of only a few of the possible explanations:

- The respondent refused to answer one or more questions on a survey.
- The respondent is not home when the interviewer calls.
- During a phone interview, the respondent hangs up.
- An interviewer inadvertently skips one or more questions.
- An experimental animal dies halfway through the study.
- A recording instrument fails.
- A survey is collated improperly, and a page is left out of some surveys.
- Data collected by other persons or organizations—the US Census, for example—may not have certain pieces of information available for every case.

Missing data are an unavoidable reality in research. The implications of missing data include the possibility of making inferences on the basis of sample data that are inadvertently biased in unknown directions, as well as being forced to rely on reduced sample sizes for statistical analysis. What does the researcher do? Ignore missing data? Fill in arbitrary values to complete the distribution of scores?

The nature of statistical inference, generalizing from sample observations to conclusions about populations, presupposes that data are missing; that is, the sample differs from the population precisely because it does not contain all the observations within the population. The logic of statistical inference, however, presumes that the sample is *randomly* drawn from the population. Similarly, whether the missing data within a sample are random is an important consideration. When data are missing in a random fashion, there is no systematic difference between the available data and the missing data; they are both random subsets of the data composing the entire sample.

There is no acknowledged way of making this determination, but some guidelines are available to the researcher. In practice, we become suspicious if a large number of data are missing from a certain variable, because it cannot be assumed that the missing data are representative of the remaining data. The missing values may imply, for example, that certain types of subjects had difficulty responding to an item. Excising the scores of these subjects would be prejudicial. For example, people with certain characteristics may be very willing to answer questions about their sexual activity, whereas people with other characteristics may be equally unwilling to do so. These different sorts of people are probably different on a number of variables related to their attitudes and values about sexuality and morality.

As a first step, the researcher seeks to determine the extent of the missing data. The procedures suggested for the examination of data in Chapter 2 will make the problem very clear. Note that in some cases, data *should* be missing, as when a question is the result of a filter from a previous question. For example, it makes no sense to ask people the age at which they were married when a previous question has established that they have never been married. All single, never-married persons should have data missing on such a question.

The researcher then seeks to determine the reasons behind the missing data to help determine if the missing data are the result of random oversights or systematic bias. Sometimes, data are missing because of data collection or data entry problems. More frequently, data are not available on particular variables from certain research participants, or the respondents themselves neglected to answer or refused to answer especially taxing or intrusive questions. In the latter instance, you may be able to avoid having substantial missing data by designing questions that are more benign or by collecting the data more sensitively. For example, rather than ask people directly what their annual income is, survey researchers often hand respondents a card with income categories represented by letters and ask the respondents to give the letter of the category that comes closest to their annual income. Thus, a gauge of annual income can be obtained without respondents having to provide a direct dollar figure.

Three Patterns of Missing Data: MAR, MCAR, and MNAR

The types of “missingness” generally come in three forms:

1. Data missing at random (MAR). The distribution of the missing data is similar to the distribution of the observed data.
2. Data missing completely at random (MCAR). The distribution of the missing data does not depend on the distribution of the observed data either
3. Data that are not missing at random (MNAR)

The distinction among MAR, MCAR, and MNAR may not be obvious. Consider, for example, a longitudinal study in which a large group of men receive prostate-specific antigen (PSA) testing for possible prostate cancer. If some of the men who were initially tested were randomly selected for retesting 6 months later, the result would be MCAR. If all of the men were invited for retesting but only those with a PSA score above 3.00, let's say, showed up 6 months later, the result would be MAR. However, if all the men showed up for the second testing but only those men whose scores on retest were above 3.0 were retained, the result would be MNAR. The important implication is that as the conditions of the missing data move from MCAR to MAR to MNAR, the distribution of scores becomes increasingly different from that of the population: The mean increases, and the standard deviation decreases.

If a review of the research procedures and an overview of the missing data are not sufficient to identify distinct patterns, you may need to take more formal steps to make this determination. If you cannot be sure that the missing entries are random, and if the distributions of included variables are affected by the missing data, then any statistical results based on these data will be biased.

One method to determine if the process resulting in missing data is random is to form two groups, one consisting of cases that contain missing values on a particular variable and the other group consisting of cases without the missing values (Hair, Black, Babin, & Anderson, 2009). These groups are then compared for patterns of significant difference. If these patterns are found, they suggest a nonrandom process of missing data acquisition. For example, let's say we are conducting a survey on the number of sexual partners acknowledged by men and women. Not surprisingly, some participants leave this question unanswered. We now compare the percentage of men and women who answered the question on sexual partners with those who left it blank. If these percentages are almost equal, it appears that the data are randomly missing; if these percentages are significantly different, the data are not randomly missing, at least with respect to the variable of gender. With continuous variables rather than categorical variables, the comparison would be made using a *t*-test rather than percentages. Note, however, that drawing conclusions about randomly missing data on the basis of nonsignificant differences between groups is akin to confirming the null hypothesis, a practice we took issue with in Chapter 3.

A second method uses dichotomized correlations to evaluate the correlation of missing data between variables (Hair et al., 2009). With this approach, all valid values of a variable are coded "1," and all missing data are coded "0," essentially creating a new dummy variable with two codes: zero for "missing" and 1 for "valid." All remaining variables are now correlated with this dummy variable. The correlations show the level of association between being missing on the variable in question and all other variables. Randomness within a pair of variables is indicated by low correlations. A statistical significance test of the correlations offers a conservative estimate of randomness. Note that lack of randomness in this context means that the observed values for a variable may still represent a random sample of values for each value of the paired variable but not necessarily a random sample of all values on that variable. With regard to the previous example, missing data on prior sexual partners might occur randomly among both men and women but much more frequently for women than for men. Consequently, any remedial procedure to accommodate the missing data on sexual partners must consider the gender of the respondent.

Adjusting for MCAR Data

After researchers obtain a clearer understanding of the scope of the problem, they can institute a number of alternative procedures to deal with it. If data are believed to be MCAR, one technique is to simply delete subjects with missing data. Called **case deletion** or **listwise deletion**, this is the default option in many statistical programs. It is a straightforward method but has the significant disadvantage of reducing power (i.e., increasing the standard error) through subject loss. Only in cases where a few subjects account for a substantial portion of the missing data, or where a large number of subjects are available and very few data are missing, is this strategy recommended. An exception, noted by Hair et al. (2009), is to delete cases in which the missing values occur in the dependent variable of a statistical analysis. In general, however, whenever missing values are distributed throughout cases and variables in a multivariate study, deletion of entire cases leads to considerable loss of data. Furthermore, when the data are organized in an experimental design, losing a single case may result in unequal cell sizes and lead to more complicated data analyses.

We have found that in research, no matter how carefully designed the data collection instruments or how carefully the potential participants are screened, the instruments typically will be largely incomplete in a few cases. We typically assume that the respondent was not interested, was unmotivated, or, worse yet, made a conscious attempt to sabotage the research. In such situations, it may make sense to eliminate all data for that case.

Adjusting for Missing, Non-MCAR Data

Case Deletion

When the missing data are not MCAR, then the results from using a case deletion approach may be biased, especially if the violations of MCAR are substantial, because the complete sample data set may not accurately represent the population. Case deletion is apt to be especially inefficient for multivariate analyses with large sample sizes because a few missing items on several variables can result in many cases being eliminated. The beauty of case deletion is in its simplicity, but as Schafer and Graham (2002) advised, it is best to explore the data set before proceeding to ascertain that the discarded cases aren't overly influential.

A related approach is to eliminate variables that are associated with considerable data loss from the study. Taking this approach may in fact be unavoidable, because items that are left unanswered by many subjects are likely to be untrustworthy. On the other hand, no investigator wants to lose key variables from a study.

Imputation

A second major approach to handling missing values is to use "imputation" techniques. **Imputation** refers to estimating missing values and then using the estimates in subsequent statistical analyses, that is, proceeding as if there are no missing data. Imputation techniques have become increasingly sophisticated over the last several years, initially stimulated by Rubin's (1976) algorithmic framework for inferring the values of incomplete data. The necessary statistical computations can be difficult, but, fortunately, contemporary statistical program packages make imputation more accessible. One of the first programs to be adopted is "SPSS Missing Value Analysis®," which assesses the magnitude of each pattern of missing data within a table. The procedure uses different missing value methods (listwise, pairwise, regression, or expectation-maximization) to estimate means, standard deviations, covariances, and correlations and then fills in (imputes) the missing values with estimated values. The EM (Expectation-Maximization) algorithm and the regression imputation are particularly noteworthy approaches to generating these values.

As indicated above, many forms of imputation are available, including complex, model-based procedures that can be employed with nonrandom missing data (Little & Rubin, 2002). We will address only common ones here. Each approach can be implemented either by using data from observations with no missing data to estimate missing data on the remaining cases (the "complete case approach") or by using data from all available valid observations to make these estimates (the "all-available approach"; Hair et al., 2009).

The most straightforward method is **mean imputation** (also called **mean substitution**), which consists of entering the mean value of a variable for any subject with missing data on that variable. Mean imputation is a conservative procedure, in that the mean of the distribution of that variable does not change. The procedure will, however, artificially reduce the variance of the distribution and thus may reduce the correlation of the variable with other variables. Although the approach is certainly easy to administer, it is no longer recommended (Allison, 2002).

A second technique is known as **random imputation** or **sequential hot deck imputation** (Little & Rubin, 2002). The idea is to replace the missing value with a value chosen randomly from the available cases. The term *sequential hot deck imputation* comes from the process of arranging a data file randomly and utilizing the case adjacent to the case with the missing value to provide that score. This approach does not systematically affect the variance of the distribution in the way that mean imputation does, but it does introduce more random variability. As such, it is no longer viewed as a desirable approach either.

Regression

A more sophisticated method for estimating missing values is to rely on **regression values**. One constructs a regression equation with the other variables as independent variables and the variable with missing data as the dependent variable. The equation is derived from subjects without missing data and is then used to predict the missing values for the remaining cases. Typically, the predicted values from the first round of regression are assigned for missing values, and then all the cases are used in a second regression. The predicted values for the variable with missing data from this round are the basis for a third regression. The process keeps going until the predicted values from one round to the next are similar. The predictions from the last round are then chosen to replace the missing values.

The advantage of the regression approach is that it offers a more accurate estimate of missing values. The disadvantages are that it is computationally complex and that scores taken from regressions fit together better than they should because the estimates have been based on the other variables and are likely to be more consistent with them than actual scores would be (Tabachnick & Fidell, 2007). Thus, the method reinforces the relationships present in the sample data, which then become less generalizable. The SPSS Missing Value Analysis procedure adds some random error to each substitution to reduce the scope of the problem; nonetheless, better strategies are available.

Adjusting for Missing Outcomes Due to Participant Attrition

A particularly aggravating situation is when the data that are missing consist of outcomes for participants who failed to complete a study or procedure.

Ignoring these participants and performing the analysis on those who completed the study not only reduces sample size but also runs the risk of bias, because attrition may be directly related to treatment condition.

Most imputation methods also have limitations here. Simple mean imputation, of course, assumes that attrition is random over the entire study, which usually is not the case. More sophisticated methods of imputation, such as those proposed by Pigott (1994), are better, but they still assume that the mechanisms that account for the missing data can be ignored. That is, the mechanisms may not be completely random, but at least they are not related to the actual values of the missing data, a situation that is more likely to be defensible. Moreover, such methods of imputation assume that there are variables available that are good predictors of the treatment outcomes.

Maximum Likelihood Methods

Modern approaches to imputation include maximum likelihood estimation and multiple imputation. **Maximum likelihood** (**ML**) methods are very popular and include a range of approaches. The overall principle is to estimate values for missing data that, to the extent that these estimated values represent the probable responses of the missing cases, yield distributions that make the observed data the most likely representation of the complete sample without missing values (Allison, 2002). The methods involve drawing inferences from a likelihood function derived from the observed data. Then the missing values are estimated by extrapolating the function using the principles of probability theory. For example, when a distribution is normal, ordinary least squares linear regression can be regarded as an ML method. As long as the data are assumed to come from multivariate normal distributions, a number of linear models, including logistic regression models, can be estimated in this way. For instance, the maximum likelihood estimator of the population mean is the sample mean. Other estimators are a bit more complicated. Nonlinear models, in particular, have not yet been successfully modeled by many computer programs. According to Schafer and Graham (2002), when data are MAR, the marginal distribution of observed data provides the most likely estimates for the missing values, assuming that the model underlying the complete data set is realistic.

Rubin (1976) authored the EM (expectation-maximization) algorithm to compute ML estimates for many different missing data problems. The algorithm estimates parameters from the given data, then estimates the missing values from the parameters, and again estimates the parameters from the enhanced data set. These steps are sometimes referred to as *expectation* and *maximization*, respectively These methods have evolved over the last several years to the point of being more practical because of the advent of sophisticated computer programs with the ability to process large statistical tasks. The missing data are treated as random variables that are deleted from the likelihood function as if they were never part of the sample. Whereas older methods of imputation attempt to predict missing data accurately (and may do so while distorting the distribution variances and correlations), the ML approach focuses on making accurate inferences about a population of interest (Schafer & Graham, 2002).

Multiple Imputation Methods

The second modern approach, **multiple imputation** (**MI**), works well with almost any MAR data. MI uses random data rather than the constant or predicted value used by EM. The randomness is introduced in the imputation process to add an error component, which compensates for the systematic standard errors provided by other methods. Because a single solution is apt to underestimate standard errors, the procedure is repeated several times to create multiple completed data sets. These alternative versions are arithmetically combined to produce overall estimates and standard errors that capture the uncertainty built into missing data. Random MI produces slightly different estimates every time it is used with the same data set, but that should not be much of a problem. More details are available through Enders (2010) and Little and Rubin (2002). SPSS and SAS have modules for MI, as does Shafer's NORM program, which also provides step-by-step procedures. NORM, a free Windows program, forms MIs for data with missing values while assuming an unstructured normal model (see http://sites.stat.psu.edu/~jls/misoftwa.html).

MI has considerable flexibility in terms of data formats and sample sizes. As with other likelihood methods, it assumes that data are MAR, but apparently MNAR applications are available as well (Schafer & Graham, 2002). It is fundamentally a Bayesian approach (see Chapter 4), which requires the assumption of a model at the imputation stage. For two variable data sets, a simple regression model works fine. A multivariate normal model is the most commonly used assumption (all variables have normal distributions, linear relationships with the other variables, and a normal, homoscedastic error term). One approach, cited by Allison (2002), is data augmentation (DA), a method of determining posterior distributions that is common in Bayesian statistics. As a general principle, the number of iterations for DA should be no fewer than those required for EM, and the more data are missing, the more iterations are needed. It should also be kept in mind that MI is designed for continuous variables rather than categorical variables; the latter require some modifications to the procedure (Allison). It is not always necessary to choose an imputation model based on established theory, but the model does need to be sufficiently robust to be valid in subsequent analyses.

Adjusting for Missing Values: Summary

In summary, a rule of thumb in compensating for missing data is to use all of the available data if possible. That's why recommendations that involve arbitrary threshold rules such as "Exclude all cases that are missing more than 15% of the responses," may be ill-advised. A 40% response rate is not necessarily better than a 15% response rate. If the data are MCAR or MAR, low response rates do not imply bias. According to Newman (2009), the problem of missing responses is a function of the "systematic nonresponse parameters" (SNPs) that relate to the constructs being addressed. Therefore, these SNPs must be identified at the outset. For example, the lack of a response to specific questions in a survey can be predicted by constructs such as how favorable respondent attitudes are regarding the topic in question, and how confident they are in their ability to respond. Many of these social or psychological variables can be identified and obviated by careful attention to research methods (e.g., by personalizing surveys, sending reminders, or offering incentives). Obtaining empirical estimates of differences between respondents and nonrespondents on key variables can help you understand and minimize the potential biasing effect of low response rates.

Table 8.1 presents Daniel Newman's (2009) overview of bias and power issues related to common approaches to managing missing data. The techniques

Table 8.1 Parameter Bias and Statistical Power Problems of Common Missing Data Techniques

Missing Data Technique	*Missingness Mechanism*		
	MCAR	*MAR*	*MNAR*
Listwise deletion	Unbiased, low power	Biased, low power	Biased, low power
Pairwise deletion	Unbiased, innacurate power	Biased, inaccurate power	Biased, inaccurate power
Maximum likelihood	**Unbiased, accurate power**	**Unbiased, accurate power**	Biased, accurate power
Multiple imputation	**Unbiased, accurate power**	**Unbiased, accurate power**	Biased, accurate power

SOURCE: Newman (2009, p. 11).

that he recommended are printed in boldface. We recognize that there are disparate opinions among knowledgeable statisticians on this topic, but we believe that his formulations are unusually sound.

HOW DO I CONTROL OR ADJUST FOR OUTLIERS?

Most researchers have been confronted with the dilemma of eyeballing a distribution of data and discovering that a few cases (subjects) have scores that lie far outside the distribution of scores in the sample. Such so-called **outliers** can be problematic because their presence can unduly affect the description of the sample distribution and subsequent inferential statistics.

In Chapter 2, we discussed how to examine a distribution to detect the presence of outliers using box and whisker plots and stem-and-leaf diagrams. Finding outliers in univariate distributions is relatively simple. An outlier arises as an observation that appears to be unattached to the bulk of the distribution, which is typically piled up near the center with fewer cases trailing off to the sides. Then one asks these questions: Do the outliers contribute to an understanding of the phenomenon being studied? Are the extreme scores from the same population as the other cases in the sample? Should they be kept or deleted in terms of computing statistical summaries and tests?

Identifying Outliers

Univariate Distributions

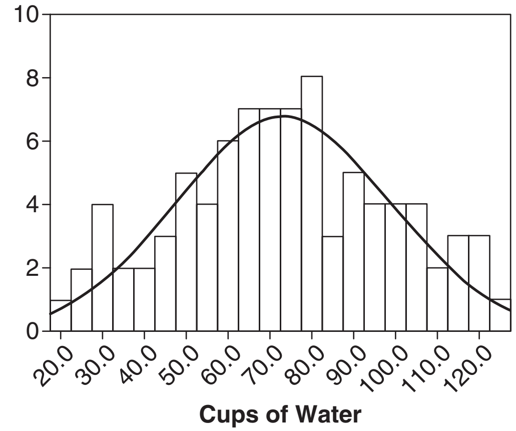
The first task is to identify the presence of an outlier. Convention suggests that scores that are more than 3 standard deviations from the mean may be regarded as outliers on a **univariate** distribution. With smaller sample sizes (fewer than 70), this criterion could be reduced to 2.5 standard scores (z scores); with very large sample sizes, one might anticipate more extreme standard scores (z scores), including a few in excess of 3 standard deviations from the mean, and adjust the criterion upward. A glance at a frequency distribution or a graphic display can give you a quick indication if an outlier exists. An outlier emerges as a case that appears to be unattached to the bulk of the distribution.

The bar graph in Figure 8.1 illustrates the number of 8-ounce cups of water consumed by 87 hikers on a weeklong wilderness expedition. These data represent a continuous distribution with a mean of approximately 73 and a standard deviation of 25.5. According to the rule of 3 standard deviations, an outlier would have a score below 73 – (3 × 25.5) < 0 or above 73 + (3 × 25.5) = 149.5. As shown by the histogram, and by the boxplot in Figure 8.2, there are no outliers in this distribution.

To understand the impact of an outlier on statistical computations, watch what happens to the mean and standard deviation of the water use data with and without inclusion of outliers. We add to this distribution 2 cases who consumed large quantities of water, 167 and 182 8-ounce bottles each. (The *z* scores of these two values, in the new distribution containing 89 cases, are 3.11 and 3.62, respectively.) The new boxplot, shown in Figure 8.3, clearly shows these values as being above the upper fence.

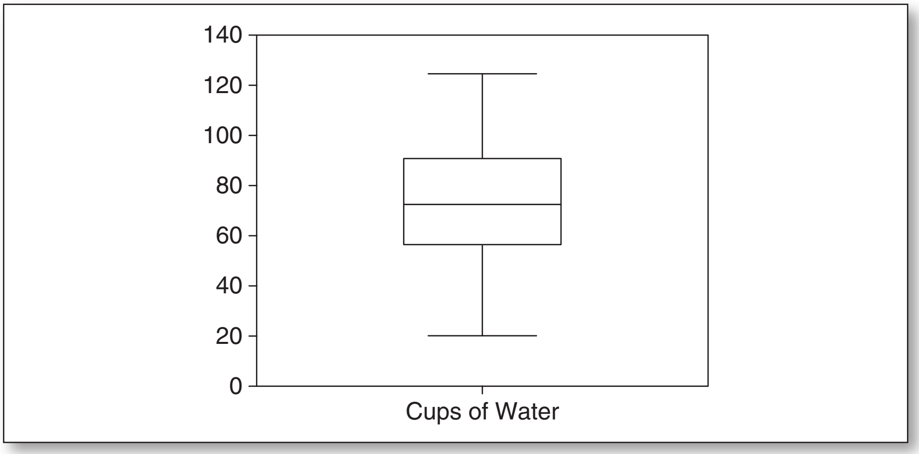
The mean of the new distribution is 75.4, up 2.3 bottles from the previous value, and the standard deviation has increased 4 bottles, from 25.5 to 29.4. Now that we have diagnosed the situation, what do we do about it? Perhaps there is nothing particularly unusual about these data: The two hikers just drink a lot of water. Perhaps the data are in error (e.g., one hiker drank 128 bottles, not 182). Although there may be no reason to exclude these two cases from the sample, the researcher might decide to report a different measure of central tendency, the median as opposed to the mean,

Figure 8.1 Histogram of Number of Cups of Water Used by 87 Hikers



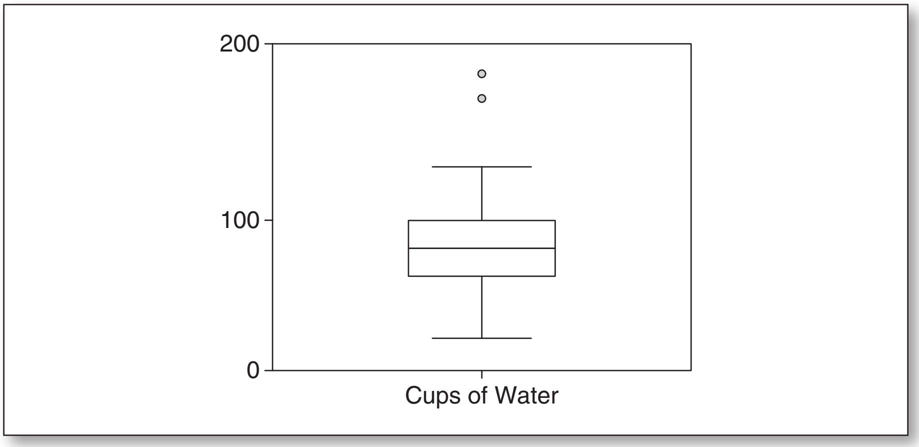
SOURCE: Newton and Rudestam (1999).

Figure 8.2 Boxplot of Cups of Water Used by 87 Hikers



SOURCE: Newton and Rudestam (1999).

Figure 8.3 Boxplot of Cups of Water Use Among 87 Hikers, With Outliers



SOURCE: Newton and Rudestam (1999).

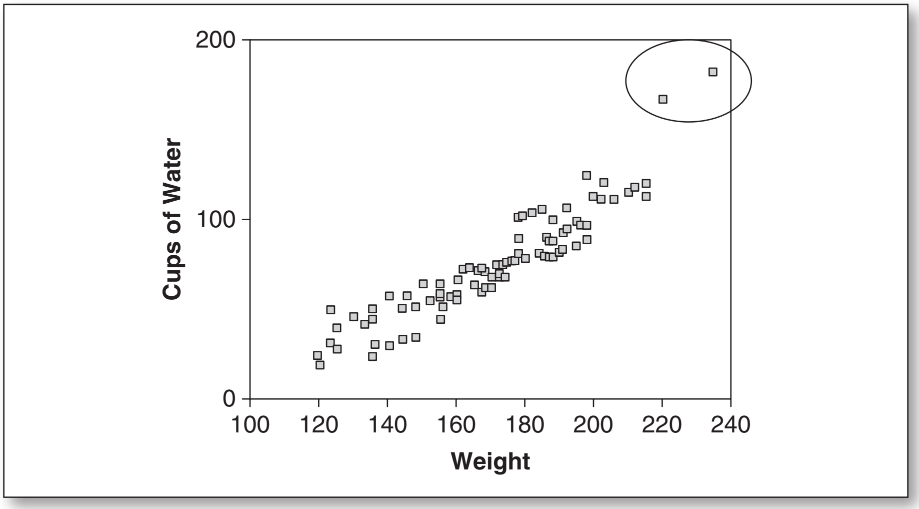
to describe the distribution, because the addition of the two outliers changed the median only from 73 to 74. Our conclusion, based on a comparison of these statistics, probably would be that the outliers were not particularly problematic.

Bivariate and Multivariate Distributions

Extreme cases can also exist in a **bivariate** distribution, even when there is no outlier on either single variable that makes up the distribution. Consider the situation in which the data indicate that a 17-year-old subject has had three divorces. Three divorces are not uncommon, nor is the presence of a 17-year-old subject, but the combination certainly is unusual. Whereas the appropriate measure with univariate distributions is the standard deviation from the mean, with two variables the measure is the **standardized residual** (greater than 3) from the regression line. Most statistical software packages compute standardized residuals.

Scatterplots can help one visually identify bivariate outliers. Superimposing an ellipse that represents a bivariate normal distribution on the scatterplot can help you visually determine the expected range of observations (Hair et al., 2009). The confidence intervals on the ellipse (e.g., 90%) can be adjusted to locate outliers according to whatever criterion you establish. For example, imagine a variable that is likely to be correlated with water use, such as the hiker's weight. We probably would predict a positive relationship between how much a person weighed and how much water was consumed. Figure 8.4 presents the bivariate scatterplot of weight by water consumption. Note that the two

Figure 8.4 Scatterplot of Water Use, by Weight (outliers on use variable circled)



SOURCE: Newton and Rudestam (1999).

outliers do not appear particularly unusual in this representation. We simply happen to have two individuals who weighed a lot and thus drank a lot of water.

Multivariate outliers derived from more than two variables can also be diagnosed, but it is difficult to do so without the aid of statistical software. J. Stevens (2009) has written a detailed account of how outliers may occur in multiple regression as outliers on the criterion variable or on the predictor variables. The analysis requires determining the multidimensional position of each observation from a common point. The usual measure of multivariate outliers is the **Mahalanobis distance**. The Mahalanobis distance is the distance of a case from the centroid of the rest of the cases, where the centroid is a point in space determined by the means of all the variables (Tabachnick & Fidell, 2007). The Mahalanobis distance is computed using a discriminant function analysis by which an equation is determined that best distinguishes one case from the rest of the cases. Whenever a case has an unusual configuration of scores, those scores become heavily weighted in the function, and the Mahalanobis distance of the case from the bulk of the cases is significant. These computations are available in many statistical software packages, including SPSS, SAS, and STATA. A conservative *p* value of .001 or less is recommended to define an outlier using this measure.

Adjusting Data for Outliers

Once outliers have been located, there is still the question of what to do with them. Because a primary cause of the presence of outliers is sloppy data recording, the first recommended antidote is to check data entry and transcription for the involved values. Sometimes, missing values are read as real data because missing value codes have not been specified accurately in the computer analysis. In such a case, the correct missing value codes need to be introduced.

More complex solutions arise if there are no coding errors. Because it is impossible to know if an outlier is an extreme case within a single population or represents a case drawn from a different population, it is not advisable simply to eliminate it. Outliers easily can be observations that represent a unique but valid aspect of the sample population. These outliers, of course, should be retained in the sample. They contribute to a complete understanding of the phenomenon under study. Elimination of them runs the risk of facilitating the statistical analysis but reducing its generalizability. If most of the outliers, however, are due to the presence of one variable, it might make sense to delete that variable from the analysis.

If the extreme cases are not part of the relevant population of cases, they can be deleted with no loss of generalizability of results because the results do not apply to that population. If the outliers *do* belong to the sample population, there are two options. One is to retain the cases but modify their values so they won't be overly influential in determining the statistical results of the study. This involves **transforming** the data (see the following discussion). The transformation is intended to change the distribution of scores to a more normal distribution because the outliers are considered to be part of a non-normal distribution. Transformation allows for easier statistical manipulation, and it retains outliers in the tails of the distribution but allows them to have less impact on the results.

A second, and certainly less drastic, option for dealing with outliers is to run your analyses twice, once with the outliers included and once without. Both sets of results can be reported. With reasonable sample sizes, the results from the two analyses frequently will be similar. The point here is that although it is important to examine and diagnose problems or potential problems with your data distributions, sometimes these result in no appreciable differences in interpretation. If this is the case, we recommend reporting the results with untransformed data.

HOW DO I ADJUST FOR NON-NORMAL DATA?

We first addressed the issue of the normal distribution in Chapter 2 and suggested strategies for assessing the "normality" of data. In Chapter 5, we discussed the "assumption of normality" as a basic criterion for the conduct of some statistical tests. In this chapter, we consider the question of what strategies to invoke when data do not appear to be normally distributed. This follows from the material in Chapters 2 and 5 and from the discussion of how to handle missing data and outliers in the previous two sections of this chapter. Once we have dealt with missing data and outliers, the problem of non-normal distributions may have fixed itself. However, when this is not the case, the researcher needs to consider using some method to adjust the distribution.

Even though many parametric statistical procedures require normally distributed population distributions, the researcher does not, and cannot, know for certain whether or not the population from which a sample came is normally distributed. He or she can, however, examine the *sample* distributions for evidence about the population's structure. The larger the sample is, the more confidence we are likely to have in what the sample distributions suggest; small samples are likely to tell us much less. As shown in Chapter 2, the easiest way to get a sense of the shape of a distribution is simply to plot it. The four types of plots that we suggested (histograms, boxplots, stem-and-leaf diagrams, and normal probability plots) are likely to indicate whether the distributions contain outliers and/or extreme skew and whether the case for the normality of the population distribution cannot reasonably be made.

Data Transformation

Not all distributions are normal. Some would argue that few are, and sometimes extreme cases do belong in the sample, resulting in distributions that are seriously skewed. One option, then, is to modify ("transform") the distribution in such a way that (a) extreme cases won't be overly influential in determining the statistical results of the study and (b) the distribution assumes a more "normal" shape.

Data transformation is a larger topic that has generated considerable interest among researchers at least since John Tukey's (1977) groundbreaking volume, *Exploratory Data Analysis,* which proposed a host of graphic and numeric ways of looking at data in order to understand them better. Transforming data is also useful for responding to a number of distribution problems, such as lack of normality, and, in bivariate and multivariate distributions, lack of homoscedasticity, nonlinearity, and lack of bivariate and multivariate normality.

The justification for changing or transforming data goes back to the goals of the researcher. In data analysis, we are interested in *describing* data and in making *inferences* from the data. At the descriptive level, we have focused on measures of central tendency (mean, median, model) in order to identify the typical score and on measures of dispersion (range, standard deviation, variance) and association (correlation coefficients). By far the most fundamental way of identifying the typical value of a distribution of scores is to cite the mean. But some distributions, such as the skewed or bimodal distributions described in Chapter 2, are not as well behaved; the mean, median, and mode, for example, may not be in the same location, making it more difficult to identify the typical value. In that case, we may transform the data to a different metric to create more symmetry and establish a more functional relationship among the variables (Fink, 2009).

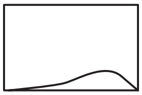
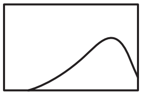
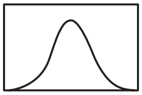
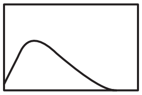
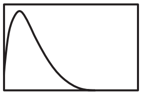
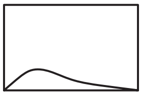
Power Transformations

In order to transform skewed, unimodal data one might use a **single-bend**, or **one-bend**, transformation (also called a **power transformation** because it involves raising the value of a variable (X) to some power (q). For example, squaring a variable raises that variable to the power of 2 (i.e., X^2), which makes the distribution more symmetric and the mean become a reasonably typical value. In a skewed distribution, scores are closer to each other at one end of the distribution than at the other end. For instance, it has been suggested that happiness is positively correlated with income, but only up to about $75,000 per year

(Kahneman & Deaton, 2010). Thereafter, the relationship is marginal. This would be a negatively skewed distribution, such that the same amount of income difference (e.g., $10,000) at the lower end of the distribution would yield greater increases in happiness than the same change in income at the upper end of the distribution. Transforming the data by converting the scores logarithmically would compress the data so that the differences between the scores would be similar at both ends of the distribution (i.e., the skewness disappears, and the mean and median become more or less identical).

Power transformations are particularly helpful in reducing skew, condensing outliers, and conditioning the distribution to approximate a normal curve. Tukey (1977) gave the moniker "ladder of powers" to the set of steps that applies different powers to bring non-normal distributions toward normality. We provide some illustrations in Figure 8.5 and an example afterward.

Figure 8.5 The Effects of Power Transformations on Distribution Shape

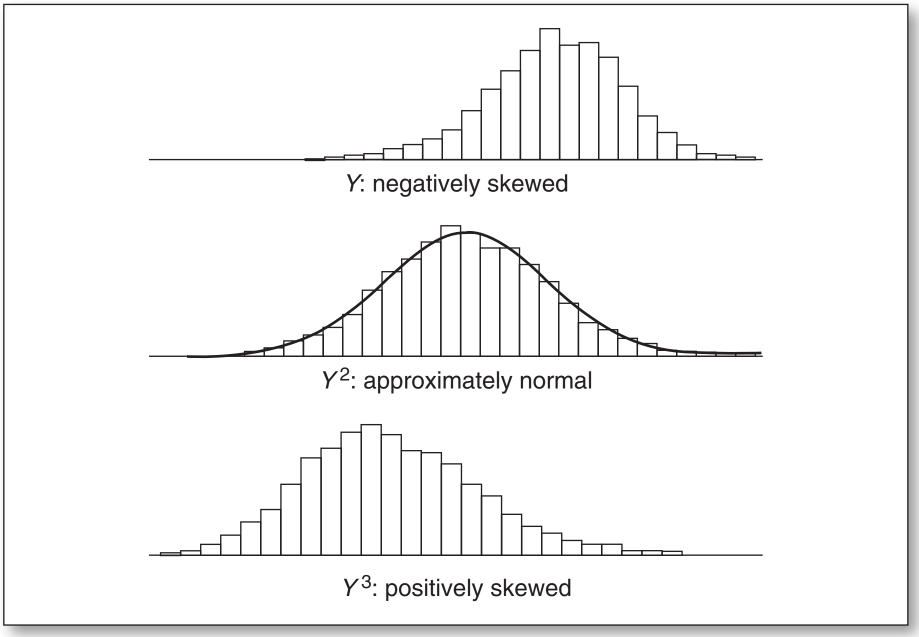
Problem	Transformations	Name	Effect
	$X = X^3$ $(q = 3)$	Cube	Reduces extreme negative skew
	$X = X^2$ $(q = 2)$	Square	Reduces negative skew
	$X = X^1$ $(q = 3)$	Raw	No effect
	$X = X^{(1/2)}$ $(q = 1/2)$	Square Root	Reduces Positive skew
	$X = \log_{10}(X)$ $(q = 0)$	Log	Reduces Positive skew
	$X = -[X^{(-1/2)}]$ $= -1/\sqrt{X}$ $(q = -1/2)$	Negative reciprocal root	Reduces extreme Positive skew

SOURCE: Newton and Rudestam (1999).

As can be seen by examining the values of *q* in Figure 8.5, powers of *q* that are greater than 1 are used to adjust for problems of negative skew. This is because they tend to change the distribution by shifting the area of the distribution to the upper tail. In contrast, powers less than 1 change the distribution by shifting the area out of the upper tail, thus reducing positive skew. In general, if a distribution differs moderately from normal and is positively skewed, a square root transformation should be tried first, and if a distribution is substantially different from normal, a **logarithmic transformation** is recommended (Tabachnick & Fidell, 2007). Logarithmic transformations are particularly good for stabilizing the variance between different sets of data. This can be useful when you are comparing batches of data from different populations. We provide examples in Figures 8.6 and 8.7 using distributions that represent positive and negative skew.

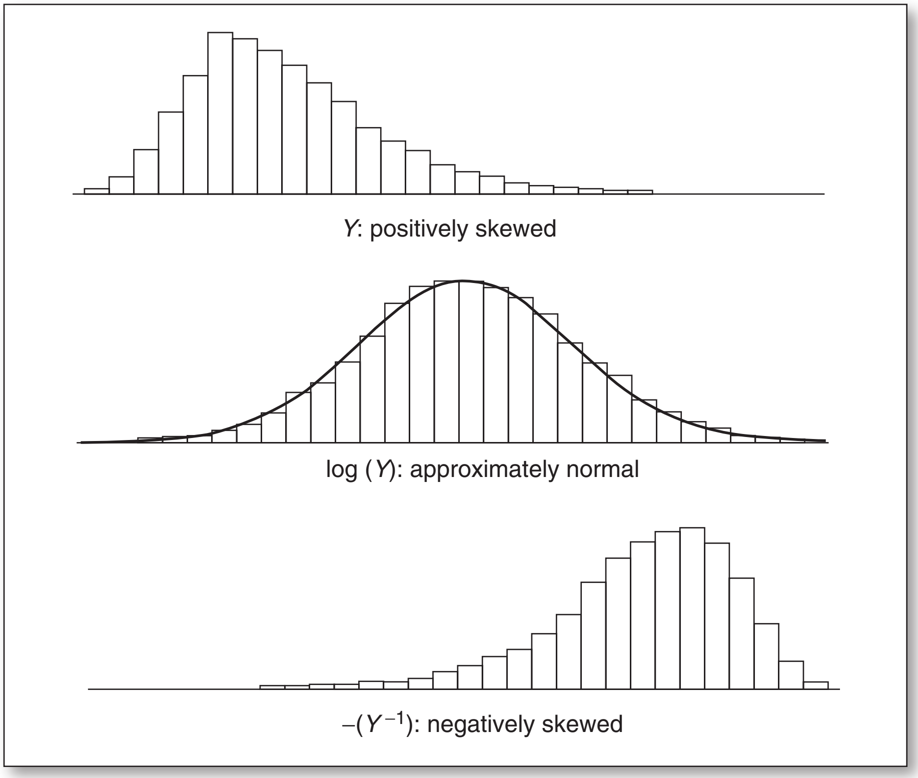
As can be seen from the examples, a positively skewed distribution can be made approximately normal by applying a log transformation, but we can also

Figure 8.6 The Effect of Square and Cube Transformations on Negative Skew



SOURCE: Adapted from Hamilton (1992).

Figure 8.7 The Effect of Log and Negative Reciprocal Transformations on Positive Skew



SOURCE: Adapted from Hamilton (1992).

go too far and morph a positively skewed distribution into a negatively skewed one, essentially leaving us no better off than when we started. Similarly, "over-correcting" a negatively skewed distribution by applying a cube transformation changes a negatively skewed distribution into a positively skewed one. A square transformation works much better in this case.

So how does the researcher go about selecting the right power transformation? The answer usually is found by trial and error, but some software packages can help with the job, not only by making it easy to assess the degree of non-normality but also by making suggestions for an appropriate transformation. For example, Stata® will plot the histogram of a variable using the ladder of powers to let the user select the transformation that appears most reasonable. SPSS® provides a dialog box that contains various transformation options; the user only needs to select one that seems appropriate. Finally, the ladder of powers offers a range of values that may be suitable, but it may be necessary to select a power that is between two of these values. For example, if a transformation overcorrects, such as a cube, this does not imply that a square is appropriate. It may be necessary to select a power such as 2.3 or 2.5.

Data transformation can also be used to deal with nonlinearity. Most common statistical tests assume a general linear model, and applying them to nonlinear data can violate this assumption. Moreover, your theory or hypothesis is likely to assume a linear relationship between an independent and dependent variable. In such cases, it may be necessary to transform your data to create linearity and then to test the linear relationship in accordance with your hypothesis. Standard transformations for these purposes can be found in most experimental design texts. Prior to adopting a procedure for transforming data, eyeball your data to examine a variable's skewness, equality of spread, and linearity. In addition, you may use statistical procedures to evaluate these attributes (e.g., Whistler, White, Wong, & Bates, 2007).

Finally, power transformations can be important in meeting the assumptions of multivariate techniques; they can aid in dealing with problems of multivariate normality in regression analysis and heteroscedasticity in both analysis of variance and regression analysis. We recommend Lawrence C. Hamilton's intermediate text, *Regression with Graphics* (1992), for an excellent treatment of these applications.

Limitations of Data Transformation

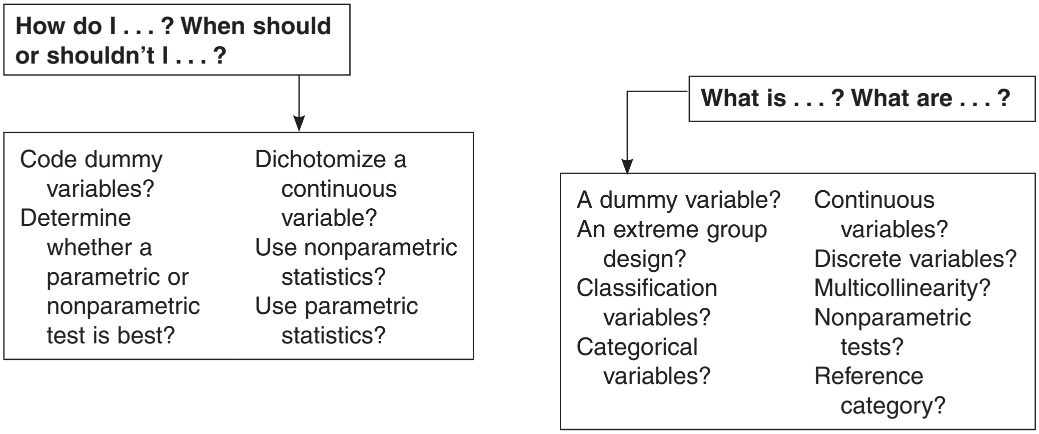
Edward Fink (2009) has raised the question of whether violations of the assumptions of common inferential tests are to be taken seriously, that is, whether most of these tests are not sufficiently robust to absorb such violations to normal distributions. He argued that although minor violations of test assumptions will probably not affect rejecting or failing to reject the null hypothesis, if the purpose of the research includes understanding the functional form of the relationships among the variables, then transformation comes into play even if the statistical tests are robust. The transformation of a variable offers feedback to theory as well as to how measurement of the variables should proceed.

The biggest limitation of transforming data is the possible difficulty of interpreting the new scores. For instance, if raw scores refer to an inherently meaningful scale such as income, the transformed scores may be harder to interpret.

As Fink (2009) put it, a transformed variable can always be untransformed subsequently for the sake of clarity. He suggested that, in addition to understanding enough theory to generate a problem worthy of study, having the skills to measure the relevant variables, and knowing how to design the data collection procedure, a researcher should be able to examine data visually and statistically and figure out if and how the data need to be transformed.

On the other end of the spectrum are behavioral scientists such as Erceg-Hurn and Mirosevich (2008), who are very dubious about the value of transforming data for violations of normality and homoscedasticity. They assert that these violations are indeed widespread and potentially damaging. For example, the likelihood of a Type I error at a *p* value of .05 can jump to over 50% when data are non-normal and heteroscedastic (Wilcox, 2003). Moreover, the power of many commonly used statistical tests (e.g., *t*-test) can be significantly reduced when these assumptions are violated. What makes matters worse, claimed Erceg-Hurn and Mirosevich, is that the so-called assumption tests (e.g., Kolmogorov-Smirnov test, Levene's test) found in SPSS and other statistical software don't work well either when distributions stray from normality and homoscedasticity. Unfortunately, they are no more sanguine about the use of data transformation, arguing that transformation may not restore normality and homoscedasticity; may reduce overall power; and may, as we mentioned earlier, make the interpretation of results more difficult. Their solution? Rely on so-called **modern robust statistical methods,** a topic we pursue further in Chapter 12.

Types of Variables and Their Treatment in Statistical Analysis



How do I . . . ? When should or shouldn’t I . . . ?

Code dummy variables?
Determine whether a parametric or nonparametric test is best?
Dichotomize a continuous variable?
Use nonparametric statistics?
Use parametric statistics?

What is . . . ? What are . . . ?

A dummy variable?
An extreme group design?
Classification variables?
Categorical variables?
Continuous variables?
Discrete variables?
Multicollinearity?
Nonparametric tests?
Reference category?

Level: Intermediate

Focus: Instructional