

Chapter 4

STATISTICAL METHODS IN CLINICAL RESEARCH

ALBERT D. FARRELL, PH.D.

Numerous exciting developments have occurred in the application of statistical methods to clinical research problems. Models for applying sophisticated statistical tools to a variety of clinical research problems have been presented (e.g., Gottman, 1995; Hoyle, 1994; Newman & Howard, 1991). The advent of increasingly powerful and accessible personal computers has led to user-friendly programs that greatly facilitate the data analysis process. At a more general level, the debate over the merits of null hypothesis significance testing has reemerged and has begun to receive serious attention (e.g., Abelson, 1997; Cohen, 1994; Cortina & Dunlap, 1997; Schmidt, 1996). Although these developments have the potential to improve the quality of clinical research, they have also had some negative consequences.

Many clinical researchers lack sufficient familiarity with some of the newer statistical techniques such that they are not only unable to use them in their own research, but they may have difficulty understanding published research in which these techniques have been used. Keeping up with advances in statistical methods in addition to a substantive research area can be a daunting task, and everyone cannot know everything. I suspect that some researchers find themselves longing for the "old days," when one simply needed a solid grasp of analysis of variance (ANOVA) techniques coupled with a basic understanding of correlational analyses. This may be characteristic not only of researchers who received their training in statistical methods some years ago, but also of recent graduates of doctoral programs. The graduate statistics courses taught in many psychology departments have been slow to incorporate new developments in statistical methods (Aiken, West, Sechrest, & Reno, 1990). Nonetheless, a minimal understanding of a fairly wide array of statistical techniques has become essential for both contributors to and consumers of the research literature in clinical psychology.

Whereas the increasing user-friendliness of statistical software has generally been a positive development, it too has had some negative consequences. Many statistical packages enable individuals with little understanding of statistics (sometimes far too little) to conduct complex data analyses. Such individuals are often able to rely on default settings rather than specifying the model that is most appropriate for their particular research problem (Estes, 1997). Although the default settings may be inappropriate for a given problem, this may not be evident from casual examination of the resulting

printout. This situation exemplifies the old adage "A little knowledge can be a dangerous thing." Another negative side effect of these packages is that they can distance the researcher from the data. Growing concern over this problem has led some to argue for a return to the basics, including graphic displays and stem-and-leaf plots that summarize the distribution of individual points in a sample (Cohen, 1990).

Continuing debate over the value of null hypothesis significance testing has created further confusion for some clinical researchers. Although the practice of deciding the outcome of a study based on rejection of a null hypothesis at the $p < .05$ significance level has been controversial since its inception (e.g., see Bakan, 1966; Nunally, 1960; Rozeboom, 1960), most clinical researchers have been well indoctrinated into this practice. Recent challenges to this practice (e.g., Cohen, 1990, 1994; Hunter, 1997; Loftus, 1996; Rosenthal, 1995; Schmidt, 1996) have led to renewed debate. Although this debate has yet to be settled, it has left researchers who are unfamiliar with alternative approaches uncertain how to proceed.

Although advances in statistical methods and computer technology have increased the variety of tools available to clinical researchers, the basic principles that guide the selection of an appropriate statistical procedure and its interpretation remain largely unchanged. Increasing the use of sophisticated statistical techniques will do little to advance the field of clinical research unless these methods are used appropriately. What is needed is a better appreciation of the basic principles that should guide the selection, use, and interpretation of statistics in clinical research.

In this chapter, I discuss some basic principles of statistical analysis with an emphasis on their application to clinical research. Focusing on clinical research problems does little to narrow the focus of this chapter. Clinical researchers address a wide variety of problems representing both experimental and nonexperimental designs and there are a wide range of statistical methods available. Although I describe some specific statistical techniques, these are used primarily as examples. Ideally, interested readers will refer to some of the material referenced in each section to explore some of these procedures in greater depth. My primary focus is on basic strategies for screening data prior to conducting analyses, the considerations that should guide clinical researchers in the selection of a statistical procedure appropriate for their research questions, and factors that influence the interpretation of findings.

GETTING ACQUAINTED WITH YOUR DATA

Because data collection is often a challenging and time-consuming process, researchers are understandably eager to run their primary analyses and determine the fate of their experimental hypotheses. A well-thought-out plan of analysis, however, must always begin with preliminary analyses designed to provide the researcher with a basic understanding of the nature of the observed data. This process includes a thorough examination of the data to determine the pattern of missing data, presence of outliers, distribution properties of the variables, and validity of the measurement model. These preliminary analyses are essential because they may identify problems that need to be addressed before the primary analyses can be conducted. In other cases, they may lead to changes in the analysis plan (Cohen, 1990).

Creating a Data Set

The most tedious and usually least rewarding aspect of the data analysis involves creating the data set and preparing it for analysis. Care at this stage can often save much aggravation later. The specific steps in creating a data set depend on the nature of the data. Researchers increasingly rely on computer applications such as computer-based interviews, observational coding systems, and scanners to record data directly into a computer database (Farrell, 1991). Other clinical applications involve entering data from coding sheets or paper-and-pencil instruments into a computer file by hand. Regardless of the method used, some system should be employed to verify the accuracy of the data. This includes checking some percentage of the original data (e.g., coding sheets or instruments) against the values recorded in the data set. For particularly sensitive and low base-rate data (e.g., sexual abuse or drug use in elementary school children), one incorrect entry can dramatically change the findings. In such circumstances, it may be prudent to verify each recorded occurrence. At a minimum, researchers need to examine the distribution of each variable to determine if the recorded values appear plausible. Are the means and standard deviations in line with expectations? Are all the values within the expected range (i.e., none exceed the maximum possible value)? Are missing data being identified and handled appropriately? Are there patterns of data that indicate random responding (see Farrell, Danish, & Howard, 1991)? Running descriptive analyses that report the means, standard deviations, and minimum and maximum observed values can be useful for examining the data. More generally, a table reporting group means and standard deviations is an essential part of any research report. Such data can help other researchers understand how samples may differ across studies on key variables. These are, however, only summary statistics, and more can be gained by examining stem-and-leaf plots that report all observed values (Cohen, 1990). Researchers should understand the distribution of each variable before they attempt to examine relationships among variables.

Multiple-item measures merit special attention during data entry. In general, it is preferable to enter data from such instruments at the item level and use the computer to do any necessary scoring. When done properly, computer scoring tends to be more accurate. Moreover, entering item-level data makes it possible to conduct item analyses to check the internal consistency of the scales. This also makes it possible to test the researcher's measurement model of the relationships between the items and the constructs they are purported to measure, and to change the scoring if alternative scoring strategies are identified after the data have been entered. In cases where selected items require reverse coding, they can be recoded into new variables that are then used in calculating scores (I add an "r" to the end of the names of such variables to indicate they have been recoded). It is often helpful to calculate item-total correlations for each item on a scale to verify that all the items are positively correlated with the total score. It is important that researchers who use the computer to score their data understand how their software package treats missing data when performing calculations. For example, a researcher using SPSS (Norušis, 1993) might use the *Sum* command within a compute statement to calculate the total score for a 20-item scale. A potential problem is that this command will calculate the total based on however many items are present. For example, if an individual completes only 5 of the 20 items, the total score will be the sum of

those 5 items. This can obviously lead to some artificially low scores. One option is to specify that all 20 items must be present to calculate the sum. However, this will cause individuals who are missing only a single item to have the entire scale coded as missing. A generally better option is to estimate or impute missing item values and include the imputed values in the calculations. This topic is addressed in the following section.

How Will You Handle Missing Data?

Missing data is a common occurrence in clinical research. Participants may choose to omit their responses to particular items or give responses that cannot be clearly coded. Some participants may not be able to complete a measure within the allotted time. Observational data may not be available because of equipment malfunctions. Data on some participants may be obtained from parents but not from teachers, or from therapists but not from clients. Absentees or transfers within a school system may result in data from students being available for some time points but not for others. Attrition may result in incomplete follow-up data for clinical trials. Indeed, the collection of a data set in which complete responses are obtained from every participant at each time point would seem to be a relatively rare occurrence. Because of the frequency with which missing data occurs in clinical research, it is important that clinical researchers develop appropriate strategies for addressing this problem.

A key consideration in examining the impact of missing data concerns the extent to which the data are missing at random (Little & Schenker, 1995). If data are missing from a random subsample of participants, the primary effect will be a reduction in the net sample size and a corresponding reduction in the efficiency of the analyses. This may be reflected in larger confidence intervals that result from larger standard errors and in reduced power for significance tests. A more serious problem occurs when participants from whom data are missing differ in some ways from those that provide complete data. In such cases, elimination of these participants will result in biased estimates. This can occur in clinical research. For example, it is to be expected that students with poor school attendance will be less likely to be present at multiple time points in a longitudinal study, that clients with high levels of depression who are assigned to a no-treatment control group will be more likely to seek treatment elsewhere, that participants in a weight reduction program who are not seeing any progress may be more likely to drop out of treatment, and so on. In each of these examples, analysis of cases for whom complete data are available is likely to produce biased results. The most critical factor in developing a strategy for dealing with missing data is the extent to which it minimizes the impact of this bias.

Several commonly used strategies for dealing with missing data involve the deletion of cases with missing data. Analysis of complete cases, sometimes referred to as listwise deletion, eliminates cases with missing data on one or more of the variables included in the analysis (Cohen & Cohen, 1983; Little & Rubin, 1990). This commonly used approach is the default selection for many statistical packages. Although many researchers find this strategy appealing, it has serious drawbacks. Excluding cases missing even a single variable can substantially reduce the sample size and will produce biased results when data are not missing at random (Acock, 1997; Little & Rubin, 1990). A related strategy involves pairwise deletion, or calculating the correlation or

covariance between two variables based on all cases that have values for both variables. Covariances reflect the degree of relationship between two variables, but unlike correlation coefficients, they are not adjusted for differences in scale units (i.e., a correlation is essentially the covariance between two variables that have been standardized). A missing-data correlation or covariance matrix based on these coefficients may then be subjected to analysis. This approach is appealing because it appears to make use of all available data (i.e., cases missing only some variables are included in some calculations). However, because each coefficient is based on different subsets of the data, the overall sample size is not clear, and it can result in an overall pattern of coefficients that is mathematically impossible and therefore not amenable to analysis (Cohen & Cohen, 1983). Where analyses of such data are possible the results cannot be generalized to any specific subpopulation. For longitudinal designs, researchers may be able to take advantage of procedures such as hierarchical linear modeling, which estimates parameters based on all available data (Bryk & Raudenbush, 1992; Hedeker & Gibbons, 1997; Nich & Carroll, 1997). Although this approach does not require that all participants be present at every time point, it will also result in biased estimates when data are not missing at random.

Other approaches to handling missing data involve estimating or imputing missing values. One of the most commonly used methods is mean substitution, which involves substituting all missing values for a variable with that variable's mean. The mean represents the best single estimate of a participant's score in the absence of any other information. However, replacing missing values with means will tend to underestimate the population variance because all missing values are placed at the center of the distribution. It will also produce biased estimates of population means if missing cases are not random. Other methods of imputing missing values have been designed to take advantage of whatever data are not missing for an individual case. These include substituting the appropriate subgroup mean rather than the overall mean, or constructing regression estimates in which missing scores are estimated based on all available scores for that individual (i.e., an individual missing a score on 1 variable out of 10 has that score estimated based on his or her scores on the other 9 variables). The success of the regression approach will depend on how well scores on each variable can be predicted from the other variables. This approach can also result in overestimates when correlations are calculated between variables with missing values and the variables used to estimate them. These problems have led to further enhancements in which regression estimates are augmented with an error term based on a randomly selected residual obtained from a complete case or randomly selected residual value from an appropriate distribution (Acock, 1997). A more sophisticated approach, known as expectation maximization, takes this further by conducting an iterative process to improve the prediction of missing values as more and more missing values are imputed (Little & Schenker, 1995). Although several of these approaches require specialized computer programs, standard statistical packages have begun to incorporate these procedures (Acock, 1997).

The best strategy for handling missing data is to do whatever is possible to minimize the problem during the data collection stage. In most clinical research, a certain amount of missing data is inevitable. A reasonable starting point for most researchers is to determine the extent of missing data within their sample and to use whatever data are

available to examine differences between individuals with and without missing data (see Cohen & Cohen, 1983, especially Chapter 7, for a discussion of appropriate data analytic strategies). In psychotherapy outcome research, this may involve comparing individuals who complete treatment to those who drop out (e.g., see Kendall, Flannery-Schroeder, & Ford, 1999). When the amount of missing data is small and fairly random, most procedures for handling missing data will produce similar results. In other cases, researchers will need to carefully select an approach most likely to reduce this source of bias. In some instances, it may be informative to compare several different strategies (Flick, 1988).

What Are the Distribution Properties of Your Variables?

Preliminary analyses are also needed to examine the distribution properties of variables. Many statistical techniques are based on assumptions about the underlying distributions of the variables. For example, ANOVA is based on the assumption that the variables are normally distributed. Although it has been established that many methods are fairly robust to some deviations from these assumptions (Keppel, 1982; Kirk, 1982), investigators need to assess the extent to which their data may represent an extreme deviation. In such cases, it may be possible to address this problem by using transformations (Cohen & Cohen, 1983; Tabachnick & Fidell, 1996) or by recoding the data. Another option is to use alternative analyses that make fewer assumptions about the variables' underlying distributions (e.g., Hu, Bentler, & Kano, 1992).

Examining the distribution properties of the data may also identify outliers or cases with extreme high or low values. Outliers are sometimes referred to as influential data points because their presence can produce substantial changes in the findings. For example, in a least squares regression analysis, an extreme score will have a powerful impact on the slope of the regression line because parameters are estimated so as to minimize the sum of squared differences between observed values and predicted values. As a result, the further a case is from the regression line, the more it will influence the slope of the line (i.e., a case 5 points above the regression line will have a squared residual of 25 versus a squared residual of 1 for a case 1 point below the regression line). The identification of outliers must include both univariate and multivariate outliers. Univariate outliers are cases with extreme values on a single variable. The identification of univariate outliers is a straightforward process that involves examining the distribution of each variable using stem-and-leaf plots (Tabachnick & Fidell, 1996). Multivariate outliers represent cases whose scores on any single variable may not be extreme, but that have an unusual pattern of scores. For example, examination of outliers in a recent study of risk and protective factors for adolescent drug use (Farrell & White, 1998) identified an individual who reported a high frequency of peer pressure for drug and alcohol use (e.g., being offered drugs, feeling pressured to drink), but who indicated that none of her friends used alcohol or drugs. Neither score by itself was that extreme, but the combination of scores represented a clearly unusual (and in this case implausible) pattern. Detection of multivariate outliers may be accomplished using a variety of statistical and graphical analyses (see Tabachnick & Fidell, 1996).

The most appropriate procedure for handling an outlier depends on the factors responsible for its occurrence (Tabachnick & Fidell, 1996). Outliers can reflect data

entry errors. For example, a 111 is entered instead of a 1. In other instances, outliers can reflect a failure to specify missing value codes for a variable. For example, using 999 to represent missing values for age, but neglecting to screen out these values in the data analysis. Errors of this sort can be common and can be corrected once they are identified. Other reasons for the presence of an outlier may be harder to assess. One possibility is that the individual case is not a member of the population you intended to sample. For example, inclusion of a person with a paranoid thought disorder in a study designed to investigate the relationship between perceived social support and response to stressful life events could result in an extreme pattern of scores that could have a strong influence on the findings. If the researchers did not intend to have their findings generalize to such individuals, they could justify excluding such a case. Some cases with extreme values may be representative of the population of interest, and excluding them from the analysis based on their scores would be inappropriate. Indeed, there are always outliers to some degree and elimination of one extreme data point will inevitably lead to another case being labeled extreme. Tabachnick and Fidell outline procedures that may be followed to reduce the impact of extreme scores in cases where the sample appears to have more extreme scores than would be expected in a normal population. In any case, it is important that researchers specify and justify whatever approaches they used to address outliers in any given study.

Does Your Measurement Model Fit the Data?

Another focus of preliminary analysis is to test the extent to which the researcher's measurement model fits the data. Because most constructs examined in clinical research are not directly observable, the majority of measures assess constructs indirectly, by measuring observable indicators of each construct or latent variable (Hoyle, 1991). For example, a researcher may assess marital satisfaction based on each partner's ratings, social skills by obtaining judges' ratings of responses to a role-play test, school problems by collecting teachers' ratings, and job performance by obtaining supervisors' ratings. Each example assumes a measurement model that specifies a link between the observed variables and the construct of interest. The accuracy of the measurement model is critically important because researchers are generally not interested in making inferences about the specific measures, but rather wish to draw conclusions about the underlying constructs the measures are assumed to measure. For example, a researcher who finds a substantial correlation between a measure of marital satisfaction and a supervisor's ratings of job performance may wish to make inferences about the relationship between marital satisfaction and job performance. The extent to which such inferences are warranted depends on the validity of the underlying measurement model.

Clinical researchers often make implicit assumptions about their measurement models without attempting to verify them. For example, a researcher interested in the relationship between exposure to community violence and frequency of violent behavior among adolescents might examine this relationship by calculating the correlation between scores on self-report measures of both constructs. This implies a measurement model that assumes the existence of two distinct constructs that are each represented by the items included in the observed measures, and that both measures are perfectly reliable. The first assumption is sometimes evaluated by calculating alpha coefficients

