Methods for Addressing Missing Data in Psychiatric and Developmental Research

CALVIN D. CROY, PH.D. AND DOUGLAS K. NOVINS, M.D.

ABSTRACT

Objective: First, to provide information about best practices in handling missing data so that readers can judge the quality of research studies. Second, to provide more detailed information about missing data analysis techniques and software on the *Journal's* Web site at www.jaacap.com. **Method:** We focus our review of techniques on those that are based on the "Missing at Random" assumption and are either extremely popular because of their convenience or that are harder to employ but yield more precise inferences. **Results:** The literature regarding missing data indicates that deletion of observations with missing data can yield biased findings. Other popular methods for handling missing data, notably replacing missing values with means, can lead to confidence intervals that are too narrow as well as false identifications of significant differences (type I statistical errors). Methods such as multiple imputation and direct maximum likelihood estimation are often superior to deleting observations and other popular methods for handling missing data problems. **Conclusions:** Psychiatric and developmental researchers should consider using multiple imputation and direct maximum likelihood estimation rather than deleting observations with missing values. *J. Am. Acad. Child Adolesc. Psychiatry,* 2005;44(12): 1230–1240. **Key Words:** missing data, statistical methods, deletion, maximum likelihood, expectation maximization, multiple imputation.

Data collection for research is seldom perfect. Information about past experiences may not have been collected. Study participants skip or refuse to answer questions. Measurement devices malfunction. Data are accidentally coded into impossible values that must later be deleted. All of these situations produce missing data, data values that should have been collected but were not.

Because missing data affect statistical analyses, researchers have spent decades developing methods to address this problem. Indeed, the different approaches to handling missing data can lead to different values of key statistics such as means, proportions, correlations, and regression coefficients. For example, Pigott (2001) showed that in a study of perceived ability to control asthma, the multiple regression coefficient for reading ability was reduced by half, from 0.409 to 0.201, when all cases were analyzed after replacing missing data with estimated values rather than analyzing only cases with complete data. Changes of this magnitude can influence the results of statistical testing, changing the researcher's conclusions about the relationships of variables. Thus, when analyzing data, researchers need to make informed decisions about how to deal with missing values and not simply rely on the default settings of their favorite statistical software.

Our goal here is to briefly cover the most commonly used options for handling missing data. For readers of psychiatric and developmental research, we discuss best

Accepted July 15, 2005.

Drs. Croy and Novins are with the National Center for American Indian and Alaska Native Mental Health Research, University of Colorado at Denver and Health Sciences Center, Aurora, Colorado.

Supported in part by National Institute of Mental Health grant MH42473 (Spero M. Manson, Ph.D., principal investigator). The authors gratefully acknowledge the many helpful comments from Drs. Robert J. Harmon, Janette Beals, Diane L. Fairclough, Lori L. Jervis, Christina M. Mitchell, and Joan M. O'Connell on earlier versions of this manuscript. They also gratefully acknowledge the contribution of the anonymous reviewers of this article for their valuable comments and suggestions.

Reprint requests to Dr. Calvin D. Croy, National Center for American Indian and Alaska Native Mental Health Research, University of Colorado at Denver and Health Sciences Center, Mail Stop F800, P.O. Box 6508, Aurora, CO 80045; e-mail: calvin.croy@uchsc.edu.

Article Plus (online only) materials for this article appear on the Journal's *Web site: www.jaacap.com.*

DOI: 10.1097/01.chi.0000181044.06337.6f

practices in handling missing data so that they can judge the quality of research studies. For psychiatric and developmental researchers, we provide sources for detailed information about missing data analysis techniques and software in our Resource Appendix. More in-depth discussion of several technical topics can be found in the Technical Appendix. Both appendixes are available through the Article Plus feature on the *Journal's* Web site at www.jaacap.com.

We focus our discussion on techniques that are based on the assumption that data are "Missing at Random" (see below) and that are either extremely popular because of their convenience or are harder to employ but yield more precise inferences. We make no attempt to review all of the methods for coping with missing data; several authors provide more comprehensive accounts (Allison, 2002; Little, 1998; Little and Rubin, 2002; Schafer and Graham, 2002).

KEY ASSUMPTIONS

Many major techniques for handling missing data other than deleting observations require that the missing data be Missing at Random (Allison, 2003; Pigott, 2001; Rubin, 1976) to ensure the calculation of accurate statistics. Indeed, Little and Rubin (2002) have observed that "essentially all the literature on multivariate incomplete data assumes that the data are MAR…" (p. 22). Researchers should carefully consider the theory that underlies their investigation to determine whether their data are likely to meet this assumption, as discussed below. For studies in which the Missing at Random condition often does not occur, the imputation and maximum likelihood (ML) methods based on this condition that we discuss can lead to incorrect inferences. For these studies, we guide researchers to more appropriate approaches.

When data are Missing at Random, the probability of a data value being missing is unrelated to its value after controlling for other variables in the analysis (Allison, 2002). Consider the simple example in which researchers collect data only about the occurrence of depressive symptoms (present versus absent) and the sex of each adolescent in a study. The Missing at Random assumption would be met if the probabilities of missing values for the depressive symptoms were not related to experiencing depressive symptoms themselves after controlling for the sex of the adolescent or to other characteristics for which data were not collected (e.g., physical health status). Unfortunately, without obtaining additional data from nonrespondents, it is impossible to check whether the Missing at Random assumption is true (Allison, 2003; Schafer and Graham, 2002; Sinharay et al., 2001). In the above example, it is impossible to know whether the probability of providing answers about depressive symptoms is related to those symptoms because we do not know the level of depressive symptoms for individuals who did not answer these questions. When additional data from the nonrespondents cannot be collected, researchers must consider whether strong assumptions regarding the distribution of values in the population are reasonable in determining whether their data meet the Missing at Random assumption.

When Is the Missing at Random Assumption Not Warranted?

The assumption that data are Missing at Random is not true whenever the probability of data being missing is correlated with the unknown missing values or with unobserved covariates after controlling for the observed data (Schafer and Graham, 2002). This would occur, for example, when people with particular data values are more likely not to answer the question, even after controlling for all of the other observed variables (Allison, 2002). In our previous example, the symptoms of major depression would not be Missing at Random if there was a relationship between the occurrence of depressive symptoms and the probability of reporting them, after controlling for sex. Also, the Missing at Random assumption does not hold in studies of the same people over time if the reason why participants drop out is related to the values they would have reported after controlling for the collected data (informative dropout [Diggle and Kenward, 1994]).

When deciding how to handle missing data, researchers should also consider whether variables are outcome/dependent variables or covariates/predictors (Little, 1992; Pigott, 2001). As Little (1992) shows, observations with missing data for single outcome variables should not necessarily be omitted. They can be of value in likelihood methods (see below) when values of predictors are also missing, but add no information to regressions