Random bytes column

Missing data: implications for analysis

Garrett Fitzmaurice, Sc.D.*

Department of Biostatistics, Harvard School of Public Health, Boston, and Laboratory for Psychiatric Biostatistics, McLean Hospital, Belmont, Massachusetts, USA

Introduction

Missing data are a ubiquitous problem that complicates the statistical analysis of data arising from studies in nutrition. The reasons why missing data are so problematic are at least two-fold. First, standard statistical techniques (e.g., *t* tests, chi-square tests, linear regression) assume that all subjects have complete information on all the relevant variables involved in the analysis. Indeed, as many of our readers can confirm, the standard presentation of statistical methods in introductory and intermediate-level courses in statistics implicitly assumes there are no missing data, thereby conveniently sweeping this problem under the proverbial rug. The second and closely related reason why missing data are problematic has to do with how they are handled in analyses implemented by many standard statistical software packages. With few exceptions, the default option for handling missing data in most statistical software programs is to exclude them entirely from the analysis. That is, only individuals with complete information on the relevant variables are included in the analysis, with all others being excluded. This default option is commonly referred to as "listwise deletion" or "casewise deletion" and the subsequent analysis is sometimes referred to as a "complete case analysis." On the surface, this default option is remarkably simple and has the apparently desirable effect of producing a reduced dataset that is ostensibly free of the problems of missing data and therefore amenable to analysis using conventional techniques.

However, there are two direct consequences of listwise deletion that are problematic. First, listwise deletion can result in a very significant loss of information. For example, suppose that the analysis of interest involves a regression analysis with 10 predictor variables and each of these predictors has a relatively small probability of being missing, say 5% chance of being missing. Furthermore, if the chances that data on one of the predictors are missing is

unrelated to the chances that data on another are missing, then we can expect to have complete information on all 10 predictors for only 40% (or $[1 - 0.95^{10}] \times 100\%$) of the subjects in our original sample. That is, listwise deletion would omit over half of the original sample of individuals in our study even though the probability that data are missing on any single predictor is relatively low. It should be no surprise that such a drastic loss of information will adversely impact the analysis, leading to a reduction in the precision of estimation (i.e., larger standard errors, wider confidence intervals, smaller test statistics, and larger *P* values). This reduction in precision is directly related to the amount of missing data. Second, although it should be transparent that the dramatically reduced sample size often associated with listwise deletion results in very inefficient use of the available data, what is less well recognized is that the resulting analysis can produce badly biased estimates of effects of interest. In general, listwise deletion only produces unbiased estimates of effects under very strong, and in many cases unrealistic, assumptions about the missing data.

When there are missing data, the validity of any method of analysis will require that certain assumptions about the reasons for any missingness, often referred to as the "missing data mechanism," are tenable. Consequently, when data are missing we must carefully consider the reasons for missingness. In this column, I discuss a hierarchy of missing data mechanisms [1,2]. In particular, I make an important distinction between missing data mechanisms that are referred to as "missing completely at random" (MCAR) and "missing at random" (MAR). I discuss the validity of conventional methods under these two assumptions about the missing data. In a future column I will discuss alternative, and more principled, methods for handling missing data in an analysis.

A hierarchy of missing data mechanisms

The missing data mechanism can be thought of as a model that describes the probability or chance that a vari-

* Corresponding author. Tel.: +617-855-3689; fax: +617-855-3826.
E-mail address: fitzmaur@hsph.harvard.edu (G. Fitzmaurice).

able is observed or missing. To make the discussion more concrete, let us consider the setting where it is of interest to relate an outcome, *Y*, say blood glucose level, to a predictor *X*, say body mass index (BMI). Suppose that the values of *Y* are fully observed but the values of *X* are not always observed. That is, for some individuals we obtain a measurement of blood glucose level (*Y*) but do not obtain their BMI (*X*). Then, the missing data mechanism can be thought of as a statistical model for the probability or chance that *X* is missing (or observed), denoted Pr(*X* is missing).

Statisticians have taken great care to distinguish a hierarchy of missing data mechanisms. Unfortunately, their choice of terminology for distinguishing the different mechanisms is rather poor and this has led to widespread confusion among statisticians and empirical researchers alike. In this column I outline the hierarchy and provide a non-technical description of these mechanisms.

Missing completely at random

In the example given earlier, there are missing data on *X* only. The data on *X* are said to be MCAR if the probability that *X* is missing is unrelated to the specific value of *X* that, in principle, should have been obtained or to the observed values of *Y*. That is,

MCAR: Pr(*X* is missing | *X*, *Y*) = Pr(*X* is missing)

Specifically, in our example, MCAR implies that those subjects with missing values for BMI are no more likely to be obese (or underweight) or to have extreme values for blood glucose than those subjects with observed values for BMI. In a certain sense, with an MCAR mechanism, missingness in *X* can be thought of as being the result of a chance mechanism that does not depend on what was observed or on what happens to be missing.

The essential feature of MCAR is that the observed data can by thought of as a purely random sample of the "complete data" (i.e., the data that would have been obtained in the absence of missingness). As a result, the means and variances and correlations, and indeed, the entire distribution of the data that have actually been observed do not differ from those of the complete data.

An MCAR mechanism has important consequences for the analysis of the data. In particular, any method of analysis that is valid in the absence of missing data will also be valid when missing data are MCAR and the analysis is restricted to the "completers" (i.e., those individuals with no missing data). Thus, under MCAR, listwise deletion produces a valid, albeit inefficient, analysis of the data.

It must be emphasized that MCAR is a very strong assumption and should be made only in cases where there is strong rationale for it being tenable. Violations of the assumption of MCAR are testable from the data at hand. For example, if the sample is stratified on the basis of missingness in *X*, the two groups should not differ in terms of their values for *Y*. Because *Y* is fully observed, it is possible to compare the groups for systematic differences in *Y* (e.g., using a *t* test or chi-square test as appropriate).

Missing at random

In contrast to MCAR, the data on *X* are said to be MAR if the probability that *X* is missing depends on the observed values of *Y* but is unrelated to the specific value of *X* that, in principle, should have been obtained. That is,

MAR: Pr (*X* is missing *X*, *Y*) = Pr (*X* is missing | *Y*)

Specifically, in the context of my example, MAR implies that those subjects with missing values for BMI may be more likely to have extreme values for blood glucose. However, if we were to stratify the sample on the basis of similar values for blood glucose, then those subjects with missing values for BMI would not be any more likely to be obese (or underweight). That is, within strata defined by levels of blood glucose (*Y*), it can be assumed that missingness is the result of a chance mechanism that is unrelated to BMI (*X*).

Note that, unlike an MCAR mechanism, it is not possible to verify this assumption from the data at hand. That is, because the data on BMI are missing, it is not possible to verify that the probability of missingness does not depend on those values. However, because the probability of missingness now depends on observed glucose levels, this has important consequences for analysis. One is that an analysis restricted to the completers is no longer valid. Put another way, the completers comprise a biased sample from the target population. The sample means, variances, and correlations based on the completers are biased estimates of the corresponding parameters in the target population. As a result, conventional analyses based on listwise deletion are no longer valid and can yield misleading inferences when data are missing at random. In a future column, I discuss more principled methods for handling missing data that yield valid inferences when data are MAR.

Note that there is a third missing data mechanism referred to as "not missing at random" (NMAR), where the probability that *X* is missing depends on the unobserved values of *X*. In the context of my example, the data would be NMAR if those subjects with missing values for BMI were more likely to be obese (or underweight). That is, missingness in BMI is related to unobserved obesity. Subtle statistical and philosophical issues arise when data are NMAR and, in general, almost all standard statistical methods are no longer valid. When data are suspected to be NMAR, it is important to carefully assess the sensitivity of results to a variety of plausible assumptions concerning the missingness process.

Summary

In this column I have reviewed the main distinction between two important missing data mechanisms, MCAR and MAR, and highlighted their consequences for conven-

tional statistical analysis. The distinction between these two missing data mechanisms determines the appropriateness of conventional analyses. Note that the terminology for these two mechanisms incorporates the word "random" and this might lead the unwary reader to assume that such "randomness" in the missingness cannot possibly produce bias in convention analyses. As I have discussed, this is not the case. It is only when data are MCAR that the observed data can by thought of as a purely random sample of the "complete data." When data are MAR, the completers comprise a biased sample from the target population and conventional analyses based on the completers produces biased estimates of effects.

In closing, it should be emphasized that assumptions about missing data are inherently difficult, if not impossible, to verify from the data at hand. Consequently, whenever possible, researchers should go to great lengths to minimize the amount of missing data in their studies. In general, the potential for bias is somewhat greater when the proportion of missing data is relatively large.

References

[1] Rubin DB. Inference and missing data. Biometrika 1976;63:581–92.
[2] Little RJA, Rubin DB. Statistical analysis with missing data. 2nd ed. New York: Wiley; 2002.