Cornell University
Cornell Statistical Consulting Unit

StatNews #47

Limitations of Common Solutions to Missing Data
November 2001

A previous newsletter, StatNews #46, discussed some of the causes of missing data and some of the consequences of analyzing only complete cases. This newsletter will discuss some other common ways of dealing with missing data, with a discussion of their advantages and disadvantages.

Available case analysis (pairwise deletion) calculates each step of the analysis separately using the cases that have data available for that step. Therefore, a case with data missing on one variable will be used only in steps that do not involve that variable. The advantage is that the sample size for each individual analysis is generally higher than with complete case analysis, but the results are unbiased only if the data are MCAR. It can also lead to mathematical problems in computing estimates of some parameters, and is not recommended.

Most other methods involve imputation-replacing the missing values with an estimate, then analyzing the full data set as if the imputed values were actual observed values. There are many ways to choose an estimate. The following are common methods:

- Mean: the mean of the observed values for that variable
- Substitution: the value from a new individual who was not selected to be in the sample
- Hot deck: a randomly chosen value from an individual who has similar values on other variables
- Cold deck: a systematically chosen value from an individual who has similar values on other variables
- Regression: the predicted value obtained by regressing the missing variable on other variables
- Stochastic regression: the predicted value from a regression plus a random residual value.
- Interpolation and extrapolation: an estimated value from other observations from the same individual.

Imputation is popular because it is conceptually simple and because the resulting sample has the same number of observations as the full data set. It can be very tempting when complete-case analysis eliminates a large proportion of the data set. But it has limitations. Some imputation methods result in biased parameter estimates, such as means and correlations, unless the data are MCAR. The bias is often worse than with complete-case analysis, especially for mean imputation. The extent of the bias depends on many factors, including the missing data mechanism, the proportion of the data that is missing, and the information available in the data set.

Moreover, all of these imputation methods underestimate standard errors. Since the imputed observations are themselves estimates, their values have corresponding random error. Despite this, imputed values are treated as actual observations in analyses. The extra source of error is ignored, resulting in too-small standard errors and too-small p-values. Furthermore, although imputation is conceptually simple, it is usually difficult to do well in practice. Therefore, these imputation methods are not satisfactory in most circumstances.

Two alternate methods maintain the full sample size and can result in unbiased estimates of parameters and standard errors for ignorable missing data: multiple imputation and maximum likelihood estimation. These techniques are just beginning to appear in common statistical software. Subsequent newsletters will describe these methods and discuss their availability in software packages.

Author: Karen Grace-Martin

Back to StatNews Table of Contents

(This newsletter was distributed by the Cornell Statistical Consulting Unit. Please forward it to any interested colleagues, students, and research staff. Anyone not receiving this newsletter who would like to be added to the mailing list for future newsletters should contact us at cscu@cornell.edu. Information about the Cornell Statistical Consulting Unit and copies of previous newsletters can be obtained at http://www.cscu.cornell.edu).