1 Listwise deletion

By far, the most common means of dealing with missing data is listwise deletion (also known as complete case), which is when all cases with a missing value are deleted. If the data are missing completely at random, then listwise deletion does not add any bias, but it does decrease the power of the analysis by decreasing the effective sample size. For example, if 1000 cases are collected but 80 have missing values, the effective sample size after listwise deletion is 920.

1.1 What is multiple imputation?

Imputation, the practice of 'filling in' missing data with plausible values, is an attractive approach to analyzing incomplete data. It apparently solves the missing-data problem at the beginning of the analysis. However, a naive or unprincipled imputation method may create more problems than it solves, distorting estimates, standard errors and hypothesis tests, as documented by Little and Rubin (1987) and others.

The question of how to obtain valid inferences from imputed data was addressed by Rubin's (1987) book on multiple imputation (MI). MI is a Monte Carlo technique in which the missing values are replaced by m¿1 simulated versions, where m is typically small (e.g. 3-10). In Rubin's method for 'repeated imputation' inference, each of the simulated complete datasets is analyzed by standard methods, and the results are combined to produce estimates and confidence intervals that incorporate missing-data uncertainty. Rubin (1987) addresses potential uses of MI primarily for large public-use data files from sample surveys and censuses. With the advent of new computational methods and software for creating MI's, however, the technique has become increasingly attractive for researchers in the biomedical, behavioral, and social sciences whose investigations are hindered by missing data. These methods are documented in a recent book by Schafer (1997) on incomplete multivariate data.

2 Imputation of Missing Data

Imputation is the process of replacing missing data with substituted values.

2.1 Multiple Imputation

Multiple imputation provides a useful strategy for dealing with data sets with missing values. Instead of filling in a single value for each missing value, the multiple imputation procedure replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute. These multiply imputed data sets are then analyzed by using standard procedures for complete data and then combining the results from all of these analyses. No matter which complete-data analysis is used, the process of combining results from different imputed data sets is essentially the same. This results in valid statistical inferences that properly reflect the uncertainty due to missing values.

2.1.1 Phases of Multiple Imputation

Multiple imputation inference involves three distinct phases:

- The missing data are filled in m times to generate m complete data sets.

- The m complete data sets are analyzed by using standard procedures.
- The results from the m complete data sets are combined for the inference.