



Katrien Declercq

In practice, the problem of missing values is common: individuals drop out in longitudinal clinical trials, people who don't respond to sensitive questions in a survey, incorrect values that are detected in data checking. Most SAS statistical procedures exclude observations with any missing variable values from the analysis. Analysing only the complete cases has its simplicity, but the information in the incomplete cases is lost and possible systematic differences between the complete cases and the incomplete cases are ignored and the resulting inference may not be applicable to the population of all cases. Some SAS procedures use all cases with available information in an analysis. For example, the CORR procedure estimates a correlation by using all cases with non-missing values for this pair of variables, ignoring the possible missing values in other variables.

Another strategy for handling missing data is simple imputation, which substitutes a value for each missing value (several single imputation methods are available in the Data Replacement Node of the SAS Enterprise Miner). However, single imputation does not reflect the uncertainty about the predictions of the unknown missing values and the resulting estimated variances of the parameter estimates will be biased toward zero. Instead of filling a single value for each missing value, multiple imputation replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute. The multiply imputed data sets are then analysed using standard procedures for complete data. Finally, the results from these analyses are then combined in such a way that valid statistical inferences are obtained that properly reflect the uncertainty due to missing values.

The new MI procedure creates multiply imputed data sets for incomplete multivariate data. Several methods can be chosen, depending on the pattern of missingness. For data sets with monotone missing patterns, either a parametric regression method that assumes multivariate normality or a nonparametric method using propensity scores is appropriate. For data sets with arbitrary missing patterns, a Markov Chain Monte Carlo method that assumes multivariate normality is used to impute all missing values or just enough to make the imputed data sets have monotone missing patterns.

Once the complete (imputed) data sets are analysed using standard SAS procedures, the new MIANALYZE procedure can be used to generate valid statistical inferences about these parameters by combining the results from the separate analyses. These two procedures are available in experimental form in Release 8.1 and 8.2 of the SAS System and will be production in Release 9.

Detailed information for both procedures can be found from <http://www.sas.com/rnd/app/da/new/dami.html>