Chapter 19

Incomplete Data: Diagnosis, Imputation, and Estimation

Susanne Rässler
University of Erlangen-Nürnberg

Donald B. Rubin
Harvard University

Nathaniel Schenker
University of Maryland

Survey data can be imperfect in various ways. Sampling, noncoverage, interviewer error, and features of the survey design and administration can affect data quality. In particular, surveys typically have missing-data problems due to nonresponse.

This entry discusses concepts regarding mechanisms that create missing data, as well as strengths and weaknesses of four commonly used approaches to deal with missing data.

These four approaches comprise simple approaches such as complete-case analysis and available-case analysis, weighting procedures, single- and multiple-imputation methods, and, finally, direct analyses using model-based procedures, in which models are specified for the observed data, and inferences are based on likelihood or Bayesian analyses.

GLOSSARY OF KEY CONCEPTS

Unit Nonresponse. A unit fails to provide any data on the questionnaire.

Item Nonresponse. A unit answers some items on the questionnaire but not other items.

Missing Completely At Random (MCAR). Data are missing completely at random if the missingness is unrelated to the (unknown) missing values of that variable as well as unrelated to the values of other variables. For example, the missing values are a random sample of all values. The rate that values are missing can vary across the different items in the questionnaire.

Missing At Random (MAR). Data are missing at random if the missingness is possibly related to the observed data in the data set, but, conditional on these data is not related to any unknown values. In other words, the missing values are a random sample of all values within classes defined by observed values (i.e., conditional on the observed data, the missingness is completely at random).

Not Missing At Random (NMAR). The missingness depends on some unobserved (missing) values, even after conditioning on all observed values.

Ignorable missingness. If the data are MAR (which includes MCAR), and if the parameter governing the distribution of the data is distinct from the parameter governing the missingness mechanism given the data, the missingness is said to be ignorable with respect to likelihood-based or Bayesian inference. In this case, the observed data observed-data likelihood does not depend on the missingness mechanism. Distinct means a priori independent for Bayesian inference and that the joint parameter space is the product of the individual disjoint parameter spaces for likelihood-based inference.

Nonignorable missingness. When the missingness is not ignorable. In this case, a model for the missingness generally must be postulated and included in the analysis to allow valid inferences.

Single imputation. Each missing value in a data set is filled in with one value, yielding one completed data set. To get valid inference from singly imputed data, in general, special variance estimators have to be used to account for the particular imputation method applied and for the particular point estimator used.

Multiple imputation. Each missing value is replaced by a set of *m* (*m*>1) values, resulting in m completed data sets. Each of these is analyzed as if it were the true data, and the results are combined to produce a single final point estimate and its associated sampling variability, which reflects both sampling variance if no data were missing and the uncertainty with which the missing data can be predicted from the observed data. Generally, valid procedures ae obtained without specialized equations.

Note: these definitions were established in Rubin (1976, Biometrika). Also see Little and Rubin (2002).