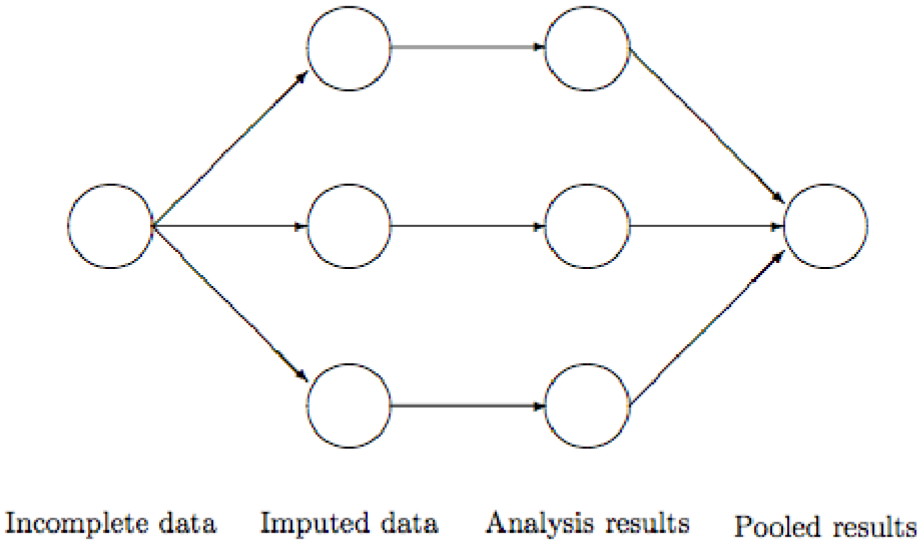
Multiple Imputation



What is multiple imputation?

Multiple imputation is a statistical technique for analyzing incomplete data sets, that is, data sets for which some entries are missing. Application of the technique requires three steps: imputation, analysis and pooling. The figure illustrates these steps.

1. Imputation: Impute (=fill in) the missing entries of the incomplete data sets, not once, but m times (m=3 in the figure). Imputed values are drawn for a distribution (that can be different for each missing entry). This step results is m complete data sets.
2. Analysis: Analyze each of the m completed data sets. This step results in m analyses.
3. Pooling: Integrate the m analysis results into a final result. Simple rules exist for combining the m analyses.

Rubin (1987) has shown that if the method to create imputations is 'proper', then the resulting inferences will be statistically valid.

The most challenging step is Imputation, that is, the construction of the mcompleted data sets. This step accounts for the process that created the missing data. Typical problems are:

1. the fact that something is missing could be related its value (e.g., people with higher incomes tend to skip income questions more often);
2. missing entries can appear anywhere in the data;
3. the method used in the imputation step must foresee the intended complete-data analyses.

The repeated analysis step on the imputed data is actually somewhat simpler than the same analysis without imputation, since there is no need to bother with the missing data.

The pooling step consists of computing the mean over the m repeated analysis, its variance, and its confidence interval or P value. In general, these computation are relatively simple.

Further reading

An up-to-date account of multiple imputation, as well as code and examples using the mice package in R, can be found in Stef van Buuren (2012), Flexible Imputation of Missing Data. Chapman & Hall/CRC, Boca Raton, FL. ISBN 9781439868249. CRC Press, Amazon

Anyone seriously dealing with incomplete data problems should consult the authorative book: Little RJA & Rubin DB (2002) Statistical analysis with missing data (second edition). Wiley, NJ.

A classic introduction to multiple imputation is the article by Joe SchaferMultiple imputation: a primer. Statistical Methods in Medical Research, 8:3-15, 1999. Another popular paper is Schafer JL & Graham JW (2002)Missing data: Our view of the state of the art. Psychological Methods, 7, 147-177.

A slightly broader text appeared in the famous green and cheap SAGE series. Check out Paul D. Allison (2001) Missing Data, which contains a gentle introduction into multiple imputation.

The original work on multiple imputation can be found in the Donald B Rubin (1987), Multiple Imputation for Nonresponse in Surveys. Wiley, NY. The paper Rubin DB (1996) Multiple imputation after 18+ years (with discussion). Journal of the American Statistical Association, 91, 473-489 provides excellent insight into many issues in multiple imputation, as well as a fairly complete account of the literature up to 1996. The discussants provide contrasting views.

The paper Meng X-L (1994), Multiple Imputation with Uncongenial Sources of Input (with discussion). Statistical Science, 9, 538-574.) deals with the important issue of how the imputation model relates to the intended complete-data analysis.

The primary source for the joint modeling approach for multivariate missing data is Schafer JL (1997) Analysis of incomplete multivariate data. Chapman & Hall, London. An early work on fully conditional specification is van Buuren S, Boshuizen HC & Knook DL

(1999) Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine, 18(6), 681–694.