Missing data and imputation 101

E-mail distributed on 08-11-2022

Dear all,

Missing data are a common problem in data analysis. Today's stat support will cover the basics of when and how to deal with this issue. If data are missing at the trial-level (multiple trial experiment), or at the item-level (multi-item scales), then this usually poses no problem. Missing trials can be ignored by averaging, or analyzed with models that do not require balanced data (e.g., multilevel regression). Missing questionnaire items can be compensated by adjusting total scores, or by using estimation methods that do not require all values to be observed (e.g., full-information maximum likelihood in structural equation modelling).

Missing data at higher levels of measurement can be problematic, depending on the type of variable, the type of analysis, and the mechanism that generated missingness. Regarding the latter, three such mechanisms have been proposed in the literature:

- **Missing Completely at Random (MCAR):** Missingness is unrelated to the data. Every case has the same probability of being missing.
- **Missing at Random (MAR):** Missingness depends on observed information in the data. Missing values can be reconstructed based on this information. The standard assumption for multiple imputation and multilevel regression.
- **Missing Not at Random (MNAR):** Missingness depends on unobserved information in the data. Missing values cannot be reconstructed.

In most software, the default handling of missing values is to delete that case entirely from the analysis (=**complete case analysis**). However, this is only valid under the MCAR assumption, and thus recommended only when that assumption is plausible, and the total fraction of missing values is extremely low. For substantial and complex missingness with many variables, MAR will be a more realistic assumption. As an alternative to complete-case analysis, researchers may also consider imputing missing data. In my opinion, the imputation method should have either minimal complexity or maximal complexity:

- **"Zero" imputation:** Replace missing values with control values which would be expected under the assumption of "no effect". Probably the safest choice for experimental data.
- **Mean/mode imputation:** Replace missing values with the observed mean of that variable for continuous data, and with the observed mode for categorical data. The best choice for simple imputation but assumes MCAR.

- **Multiple stochastic imputation:** Generate multiple imputed data sets with model-based stochastic imputation and pool analysis results using appropriate rules. Assumes MAR and reconstructs missing data as accurately as possible using observed information.

Multiple imputation (MI) can be run easily nowadays with R packages such as MICE. However, the procedure should be rigorously justified and executed! **Never** impute values that are missing by design (e.g., unassigned conditions in partial within-subject designs) or that are missing because of an experiment failure (e.g., a questionnaire item was forgotten for half the participants).

For MI, never create just a single imputed data set! The procedure requires that multiple imputed data sets are created, so that the impact of missingness on the analysis result can be estimated. For the use of MI in a paper, you need to argue persuasively that missingness could not be prevented, and that complete-case analysis would have led to an unacceptable loss of information.

Best,
Ben

Reference: Van Buuren (2012). *Flexible Imputation of Missing Data*. CRC Press.

Ben Meuleman, Ph.D.
Statistician
Swiss Center for Affective Sciences
University of Geneva | Campus Biotech
Chemin des Mines 9 | CH-1202 Genève
ben.meuleman@unige.ch | +41 (0)22 379 09 79