Dealing with missing data

1. Introduction

Missing data is a problem because nearly all standard statistical methods presume complete information for all the variables included in the analysis. A relatively few absent observations on some variables can dramatically shrink the sample size. As a result, the precision of confidence intervals is harmed, statistical power weakens and the parameter estimates may be biased. Appropriately dealing with missing can be challenging as it requires a careful examination of the data to identify the type and pattern of missingness, and also a clear understanding of how the different imputation methods work.

2. Missing data mechanisms.

There are different assumptions about missing data mechanisms:

a) Missing completely at random (MCAR): Suppose variable Y has some missing values. We will say that these values are MCAR if the probability of missing data on Y is unrelated to the value of Y itself or to the values of any other variable in the data set.

b) Missing at random (MAR)-a weaker assumption than MCAR: The probability of missing data on Y is unrelated to the value of Y after controlling for other variables in the analysis (say X).

If MAR assumption is not fulfilled: The missing data mechanism is said to be nonignorable and, thus, it must be modeled to get good estimates of the parameters of interest. This requires a very good understanding of the missing data process.

c) Not missing at random (NMAR): Missing values do depend on unobserved values.

Example: The NMAR assumption would be fulfilled if people with high income are less likely to report their income.

3. Patterns of missingness

We can distinguish between two main patterns of missingness. On the one hand, data are missing monotone if we can observe a pattern among the missing values. Note that it may be necessary to reorder variables and/or individuals. On the other hand, data are missing arbitrarily if there is not a way to order the variables to observe a clear pattern.

Assumptions and patterns of missingness are used to determine which methods can be used to deal with missing data

4. Methods for handling missing data

1. **Listwise Deletion:** Delete all data from any participant with missing values. If your sample is large enough, then you likely can drop data without substantial loss of statistical power. Be sure that the values are missing at random and that you are not inadvertently removing a class of participants.

2. **Recover the Values:** You can sometimes contact the participants and ask them to fill out the missing values. For in-person studies, we've found having an additional check for missing values before the participant leaves helps.

Imputation

Imputation is replacing missing values with substitute values. The following methods use some form of imputation.

3. ***Educated Guessing:*** It sounds arbitrary and isn't your preferred course of action, but you can often infer a missing value. For related questions, for example, like those often presented in a matrix, if the participant responds with all "4s", assume that the missing value is a 4.

4. ***Average Imputation:*** Use the average value of the responses from the other participants to fill in the missing value. If the average of the 30 responses on the question is a 4.1, use a 4.1 as the imputed value. This choice is not always recommended because it can artificially reduce the variability of your data but in some cases makes sense.
5. ***Common-Point Imputation:*** For a rating scale, using the middle point or most commonly chosen value. For example, on a five-point scale, substitute a 3, the midpoint, or a 4, the most common value (in many cases). This is a bit more structured than guessing, but it's still among the more risky options. Use caution unless you have good reason and data to support using the substitute value.
6. ***Regression Substitution:*** You can use multiple-regression analysis to estimate a missing value. We use this technique to deal with missing SUS scores. Regression substitution

predicts the missing value from the other values. In the case of missing SUS data, we had enough data to create stable regression equations and predict the missing values automatically in the calculator.

7. ***Multiple Imputation:*** The most sophisticated and, currently, most popular approach is to take the regression idea further and take advantage of correlations between responses. In multiple imputation [pdf], software creates plausible values based on the correlations for the missing data and then averages the simulated datasets by incorporating random errors in your predictions. It is one of a number of examples where computers continue to change the statistical landscape. Most statistical packages like SPSS come with a multiple-imputation feature. More on multiple imputation.

Missing data is like a medical concern: ignoring it doesn't make it go away. Ideally your data is missing at random and one of these seven approaches will help you make the most of the data you have.