בס"ד

Handling Missing Data and Imputation Methods

Yehonatan Hayoun
Bar-Ilan University

Abstract

Behind every statistical analysis, the idea is usually inference from a sample to a given population. For almost any kind of research, whether we use Administrative Data or conduct a survey or even use Big-Data files, missing values will probably occur for many reasons.

Missing data are observations that we intended to make but could not. Missing data can cause a problem if they cause that the distribution of the sample is different from the distribution of the target population from which it is sampled.

When we have missing data, our goal remains the same with what it was if we have the complete data. So, the analysis is now more complex. It is important to note that missing data are not un-exist data. Treating them as such, and using any kind of Deletion methods of missing values can cause biased estimates, create a loss of data, change summary statistics for other variables too, reduce statistical power (due to sample size), and other severe errors in statistical inference, analysis and general conclusions.

We review three types of assumptions on Missingness Mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). Accordingly, we review different methods handling missing data and Imputation Methods since the 1970s till innovative Multiple-Imputation methods which use high computing power. We also show the advantages and disadvantages of every method depending on the missingness mechanism in question.

In addition, we present examples and research relevance of missing data and statistical Computer Softwares which provides solutions handling the problem according to a selected Imputation Method.