# Dealing with Missing Data*

Denise Chang

March 4, 2024

## 1 What is Missing Data?

Missing data is a factor that every study should consider, no matter how "foul-proof" the data acquisition method is. We say that there are missing data in a dataset when there are values missing in one or many observed variables. In a survey, for example, missing data can be because a respondent chooses not to answer a question or because they forget to answer a set of questions. There are three main categories of missing data, which affect the study differently The following sections will define and discuss data that is Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR).

### 1.1 Missing Data Completely At Random (MCAR)

Data that is Missing Completely At Random is when the missing data is independent of any other of the measured variables. This is considered the best category of missing data, since MCAR data has fewer effect on analysis and inference. The data is unbiasedly missing throughout the problem set, which means that even though the data set has null values, the population is well-reflected in the study. This type of missing data is rare and much less common than the other two categories.

### 1.2 Missing Data At Random (MAR)

Data that is Missing At Random is when there are missing values in a studied variables, and these values are related to another observed variable. For example, let's consider a highschool survey in december to evaluate the students' favourite subject in each grade. The missing data in this situation could be related to the students' age since the older students are busier balancing between academics and their university applications. In this example, the data is

*Code for this paper is available at: https://github.com/DeniseChang9/STA302_ME.git

still missing at random, but there are more missing data in the older students than the younger ones. MAR data is common, but is not ideal. A bias in inference and analysis can happen since the studies population is unevenly represented.

### 1.3 Missing Data Not At Random (MNAR)

Data that is Missing Not At Random is when there are missing values in a studied variable, and these values is related to an unobserved variable, or related to the missing variable itself. For example, let's consider a survey on US citizens' views on scientific methods In this example, there may be missing data from those who opt out of the study since they don't believe in the purpose of the research. The people who don't believe in scientific methods tend to not participate in surveys since they don't trust or believe the validity of these researches. This represents a scenario of MNAR data as the bias is directly related to the studied variable.

## 2 Handling Missing Data

No matter the type of missing data, all missing data affect study conclusions and inferences as the arguments presented become less generalized. When evaluating the data, data scientists must decide on how to deal with missing data based on the context of the study. From a data perspective, common handling methods are dropping observations with missing data, imputing the mean of observations without missing data and using multiple imputations.

### 2.1 Dropping Observations

Dropping observations means to do the data analysis while excluding the missing observations. This method, as well as the next two discussed option, should be used cautiously as it may bias the dataset even further. By dropping observations, it could be the equivalent of purposefully omitting outliers or vulagrizing the study by reducing the sample significantly.

### 2.2 Imputing the Mean

Imputing the mean implies evaluating the mean of a second dataset composed of the observed variables without the missing data, and imputing the selective mean as the value of the missing data in the original dataset. This method preserves the amount of data per variable, but may be biased as it assumes that every missing data necessarily follows the mean trend.

## 2.3 Multiple Imputation

Conducting multiple imputations is similar as the previous method, but involves many more potential datasets and averaging the results of these datasets to imput into the original dataset. This method considers more factors, but is also prone to overestimating of underestimating the actual values due to biases.

Before concluding this paper, I would like to thank Gavin Crooks for taking the time to review the content of this paper and for offering valuable suggestions for improving this work. I would also like to highlight that the contents of this paper is heavily inspired by Alexander's textbook "Telling Stories with Data" [@Rohan].