1.3: Missing Data

Learning Objectives

At the end of this section you should be able to answer the following questions:

- How would you explain the difference between the concepts of MCAR, MAR, and missing not at random?
- How would you explain the purpose and interpretation of Little's MCAR test?
- What are the two main ways of dealing with missing data?

There are a number of things you will want to check before you commence any serious analyses with your data. The first thing is to check if you have any missing data. Missing data occurs when a response opportunity is missed by someone responding to your survey or questionnaire. Your participants might not respond to an item on your survey for different reasons. They may miss a question on your survey, or they may not want to answer a particular question, or they may get bored and stop filling the survey in! This lack of response to an item or items, creates a bit of a problem for the analyses. Missing data can be missing completely at random (*MCAR*), missing at random (*MAR*), or *missing not at random*.

Missing completely at random (MCAR) means the probability of a respondent missing data point is the same for all respondents. Someone might randomly miss a question on your survey, making that missing data point completely random. If the probability of a value being missing is the same only within groups defined by the observed data, then the data are missing at random (MAR). Because of this, MCAR and MAR are closely related concepts.

If the missing data is found not to be MCAR or MAR, it is missing not at random. For example, if a sizeable number of participants decide to skip one particular question on a survey, then that is not random. There is likely a reason for that question being skipped. It could be poorly worded, too personal, or just hidden at the bottom of the page.

There is a test to see if data is missing at random or not, which is called *Little's MCAR test*. Basically, if the test is not significant, any missing data is likely to have occurred at random. If the test is significant, there might be systematic or non-random reason the data is missing.

There are two main ways to deal with missing data.

First, there is a procedure of *mean replacement*. In this instance, you can replace the missing data points with a mean of that variable, though this technique is only recommended if the data is missing at random and is less than 5% of the variable in question.

There is also a technique called *multiple imputation*. This method is when the statistical program you are using, goes through the data and assigns a value for the missing variable of a particular case. This value is based upon previous responses to related variables and other non-missing responses for that variable in other cases. It is recommended that you use this if you have data missing at random of between 5-10% of the total responses of the variable.