### ****Overview of MCAR, MAR and MNAR****

Datasets can be difficult to work with if you are using real-life data - one of the ways in which problems arise is with missing data. Missing data is a common problem which can occur for a variety of reasons, and can appear as blank or empty cells in a dataset, as NaN or null values, or as an incomplete/non-random sample (amongst others!). Understanding how and why your data is missing is a very important sept in your data cleaning and validation, which is required in carrying out a proper analysis.

Missing data can occur because:

* Participants did not provide answers to certain questions for whatever reason
* Mistakes occurred in the data entry process
* There were issues with data collection instruments
* Data was removed to anonymise entries, or sensitive information was omitted to protect participants
* Environmental factors did not allow data collection to take place or continue

Data can be missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR). An explanation if these types is provided in this guide.

### ****Guide contents****

The tabs of this guide will support you in understanding missing data. The sections are organised as follows:

* **MCAR**- missing completely at random
* **MAR** - missing at random
* **MNAR**- missing not at random
* **Dealing With Missing Data** - some ways to deal with missing data to get the best results for your analysis

### ****MCAR****

MCAR stands for 'missing completely at random' and is a non-systematic form of missing data. This is when the data is randomly distributed and constitute a random subsample of your sample, meaning that it is less likely to introduce bias into the data distribution. Of course, MCAR data reduced the analysable population of the study, which will reduce the statistical power.

In practice, MCAR only very rarely happens as the conditions for its occurrences are very strict. However, when your data gaps are MCAR you can either ignore them or listwise or pairwise delete them.

#### Scenario 1

"You *are a meteorologist using previous weather data to predict tomorrow's forecast. The computer glitches when you are inputting your data, and some entries are deleted."*

This is an example of MCAR because it was a complete accident that this data was missing, and is not yours nor anyone's fault that this has happened.

#### Scenario 2

" You *are an oceanographer measuring fluctuations in seawater salinity using unmanned water vehicles placed sporadically in the area. After a few months and without anyone's knowledge, a sea creature crawls into one of the vehicles and breaks the sensor, so no data is collected after this point."*

Here, missing data comes from the fact that only a small amount of data was gathered rather than what was intended due to external circumstances beyond anyone's control, and so is an example of MCAR.

#### Scenario 3

"You are *a healthcare researcher whose research involves participants submitting a formative questionnaire, receiving a treatment and completing a summative questionnaire. One participant's summative answers are not collected because they forgot to click 'Submit' after they filled it out."*

In this case it was a complete accident that this participant's answers were not collected: the missingness is not due to any of the values, any relationship between your variables or the survey itself and therefore is considered to be MCAR.

### ****MAR****

Missing at random (MAR) data is data which is again non-systematic, is not randomly distributed but is unrelated to the variables. In other words, this missing data is related to the unobserved variables but not to the observed ones, which preserves its randomness.

#### Scenario 1

"You are *an optometrist surveying people's experiences at your brand's high street shop. You notice that many older people do not realise that the last question is on the next page and accidentally skip it."*

This missing data is an example of MAR because it affects survey participants of a different, unobserved variable rather than the observed variables involved in the analysis. The difference between this scenario and MCAR scenario 3 is that in this example this missingness is related due to the unobserved variable, and not a fluke like in the MCAR example.

Scenario 2

"You are a consultant commissioned to *investigate the worth of businesses having hybrid working options (working in the office and at home) for their employees. The results show that no participants who answered 'retired' as their working status answered the question about hybrid working."*

The context of this is particular: the variable you are concerned with has not been answered by a certain group of people...but retired people do not work, so of course there are no hybrid working options for them! The probability of completing the survey is therefore fully dependent on their work status (which is fully observed) but not due to their hybrid working abilities.

In other words, although the missing values for one question are dependent on another, they are still missing at random.

#### Scenario 3

"You are *a psychiatrist examining the link between depression and burnout. You notice that very few men signed up to take part in the study."*

Here, the sample size you have collected is not representative of the population as a significantly greater number of women took part in the study than men did. This is an example of MAR as the missing data is not related to the variables you have chosen (which here is depression and burnout) but an unobserved variable not part of the analysis (the gender of the participant).

### ****MNAR****

Data which is considered to be missing not at random (MNAR) is problematic. It is systematic, not randomly distributed and is related to the variables. MNAR can occur when there are problems in the dataset, and can indicate that there was an issue with the data source, data collection methods, the sampling methods, or otherwise, which can introduce significant bias in your data's distribution.

Sometimes it is difficult to determine if your missing data is MNAR or MAR, as the reasons why a participant did not answer a question may be unknown to you.

#### Scenario 1

"You are*a psychologist researching people's attitudes towards mental health. Some interviewees think one of the questions you ask is offensive and deliberately do not answer it."*

Here the missing data occurred when the interviewees did not answer a certain question, which would lead to blank cells appearing in your dataset: this has happened intentionally by the participants and shows that the interview you designed was poor.

#### Scenario 2

"You are *a video game developer who has created a new game and want to find people to test it. You decide to only ask your friends to test the game."*

This is similar to the third MAR scenario because there would be no blank cells in the dataset...however, this is an example of a poor sample selection. By only limiting your sample to a few deliberately chosen people, you are missing the randomness which will provide a more fair and true-to-life result, which leads to this missing data being MNAR.

By only asking your friends to be the testers you introduce bias as your friends would be less likely to provide honest answers, tending to provide more positive answers.

#### Scenario 3

"You *are a criminology student looking to see the relationship between illegal drug use and household income. None of your respondents declare that they have used illegal drugs before."*

In this scenario, you have no data on illegal drug use because even if your respondents do partake in illegal drug use, they will not disclose this as they may get into trouble. This data is MNAR as it is related to your variables for analysis and is not systematically excluded from your dataset.

### ****Dealing With Missing Data****

There is no perfect 'one-size-fits-all' approach to missing data, but there are a few techniques to help you handle and process it: you can either remove the participants' contributions (deletion) ore replace them with a suitable best guess (imputation)…or just leave them alone!

There are some instances where missing data can only be dealt with by discarding your dataset and starting again: this is definitely the case when it comes to poor sampling techniques.

#### Do Nothing!

Yes, for MCAR data you can (most of the time) simply leave it by, because in the missing data is truly random the analysis will not produce biased estimates.

#### Complete Case Analysis (Listwise Deletion)

Complete case analysis, also known as listwise deletion, involves deleting any responses removing completely any response with missing data in them, leading to your dataset containing only participants with complete data. This method is okay for MCAR data, or if we do not have a lot of missing data, however you must be careful when doing this because you may unintentionally introduce bias into your dataset by no longer making your sample representative of the population. This in turn makes it difficult to make inferences about your sample or the population.

#### Available Case Analysis (Pairwise Deletion)

Another way to remove missing data is to use available case analysis, also known as pairwise deletion. This involves removing participants who had missing data in some variable analysis, but not their other non-missing data. This approach is often favoured over complete case analysis because it reduces the number of participants you remove.

Like complete case analysis, available case analysis can also introduce bias so you must be careful when implementing this.

#### Unconditional Mean Imputation

This is a process which involves replacing the missing data with an estimated value - more specifically, you take the arithmetic mean of your observed data and replace any missing entries with it. Subsequent analyses treat these imputed data as real data entries.

However, this method is limited in its use, as replacing many missing data with the mean will have a zero correlation with the variables, and in extreme cases can dramatically influence the correlation coefficient and significance. Replacing many missing data points with a constant value also impact the variability of your sample, which may be problematic when it comes to interpreting your results and answering your research question.

#### Regression Imputation

Another way to replace missing data is with an estimated value using regression. This provides a better estimate of the missing data than using unconditional mean imputation as tit uses the observations you have already to predict what observations you most likely should have according to the model, and therefore will produce a new unbiased estimate of the true mean when the missing data is MCAR or MAR.

Understandably, regression imputation has its limitations as well: the imputed data will be biased in a linear regression as the predicted values will be based on that of the generated straight-line equation, which introduces new concerns of correlation and variability within the new dataset.

#### Stochastic Regression Imputation

Stochastic regression imputation accounts for the lost variability in unconditional mean imputation or regression imputation by generating a predicted value to replace the missing data but this time with an error term included (the replacement data point is the sum of the prediction from the regression with this error term).

A limitation this process has is that this method provides no adjustment for the standard error and makes these very small, which may lead to an increased change of a Type 1 error (false positive) occurring in your analysis.