Incomplete Data Analysis

V. Inácio de Carvalho & M. de Carvalho

University of Edinburgh

Missing data mechanisms

- The **missing data pattern** provides insight about the location of the missing values in the dataset but not about the reasons for missingness.
- To decide how to handle missing data, we must carefully consider the reasons for missingness.
- The **missing data mechanism** provides insight about the underlying reasons for missingness and, generally speaking, can be thought of as a model for the probability that a given variable is observed or missing.
- We consider three general missing mechanisms:
  1. Missing completely at random (MCAR).
  2. Missing at random (MAR).
  3. Missing not at random (MNAR).
- The type of missing data mechanism determines the appropriateness of different methods of analyses.

Missing data mechanisms

- To make the discussion more concrete, we consider the following example from Fitzmaurice (Nutrition, 2008).

- Let us consider the setting where it is of interest to relate an outcome variable Y_1, say blood glucose level, to another variable Y_2, say body mass index (BMI).

- Suppose that values of Y_2 are not always measured. That is, for some individuals we obtain their blood glucose level, but do not obtain their BMI.

- Then, the missing data mechanism can be thought of a statistical model for the probability that Y_2 is missing (or observed).

Missing data mechanisms

MCAR

- Data on a variable are said to be **missing completely at random (MCAR)** if the probability that a value is missing is unrelated to either the specific values that, in principle, should have been obtained or to the other observed (or unobserved) variables.

- In the context of our example, MCAR implies that the probability that a BMI value is missing is the same for all individuals, regardless of their BMI and glucose levels. That is, subjects with missing BMI values are no more likely to be obese or underweight or to have extreme blood glucose levels than those subjects with observed BMI values.

- In a certain sense, under the MCAR assumption, missingness can be thought of as being the result of a chance mechanism that does not depend on what was observed or on what happens to be missing.

Missing data mechanisms

MCAR

- The essential feature of MCAR is that the observed data can be thought of as a random sample of the complete data (i.e., the data that would have been obtained if there were no missing values). As a result, the distributions of the data actually observed and of the complete data are similar. This implies that the missing and observed values will have similar distributions.

- An MCAR mechanism has important consequences for data analysis.

- In particular, any method of analysis that is valid in the absence of missing data, will also be valid when missing data are MCAR and the analysis is restricted to those individuals with no missing data (known as a complete case analysis).

- Therefore, under the MCAR assumption, a complete case analysis provides a valid, although inefficient, analysis of the data.

Missing data mechanisms

MCAR

- ↪ It must be emphasised that MCAR is a very strong assumption and should be made only in cases where there is a strong rationale for it being tenable.
- ↪ Violations of the assumption of MCAR are, to a certain extent, testable from the data at hand.
- ↪ For example, if the sample is stratified on the basis of missingness in Y_2, the two groups should not differ in terms of their values for Y_1.
- ↪ Concretising, if we divide the glucose levels in two groups, one for those with observed BMI and another one for those whose BMI measurement is missing, the two groups should not differ. Because glucose levels are fully observed, it is possible to compare the two groups for systematic differences in the glucose levels.

Missing data mechanisms

MCAR

- ↪ If, for instance, the distribution of the two groups of glucose levels differ, this provides compelling evidence that the data are not MCAR and suggests a possible relationship between glucose levels and the probability of missing data.

- ↪ However, we must be aware, that there could still be some relationship between missingness on BMI and BMI values themselves and the aforementioned procedure cannot take this into account.

Missing data mechanisms

MAR

- Most missingness is not completely at random.
- A less restrictive assumption than MCAR is that the probability that a value for a variable is missing depends (only) on observed/available information but it is further unrelated to the specific missing values that, in principle, should have been obtained – **missing at random (MAR)** assumption.
- Specifically, in the context of our running example, MAR assumption implies that the probability that a BMI value is missing varies with the blood glucose levels but does not depend on the BMI values themselves. E.g., individuals with extreme glucose levels may have a higher propensity for having their BMI value not recorded.

Missing data mechanisms

MAR

- ↪ Under the MAR assumption, the probability of missing data on BMI depends on the individual glucose level, but within groups defined by individuals with similar glucose levels, then the probability of a subject having a missing BMI value is the same as for any other subject, i.e., within groups (or strata) of similar blood glucose levels, missing is MCAR (i.e., does not change over the BMI values).

- ↪ Roughly speaking, while MCAR is the process of total randomness of missingness, MAR is randomness only within the levels of what may be called the conditioning variables (in our example, the conditioning variable is the glucose level).

- ↪ In this sense, MCAR is a special type of MAR with one stratum only.

Missing data mechanisms

MAR

- ↪ Note that it is not possible to verify the MAR assumption from the data at hand because it concerns the missing values.
- ↪ For instance, in our example, within each glucose level strata, we would need to know the distribution of the BMI values among those with no recorded BMI, in order to compare it with the distribution of the observed BMI.
- ↪ MAR is an assumption that needs to be justified based on background knowledge and discussion with experts.

Missing data mechanisms

MNAR

- Data are said to be **missing not at random (MNAR)** when the probability that a variable has missing values is related to the specific values that should have been obtained, in addition to (potentially) the ones obtained in the other fully observed variables.

- In the context of the blood glucose/BMI example, the data would be MNAR if, for instance, those subjects with missing values for BMI were more likely to be obese (or underweight). That is, missing in BMI would be related to unobserved obesity.

- MNAR can also occur indirectly trough the relationship of the variable with missing data with another variable that is not available in the dataset. A familiar example from medical studies is that if a particular treatment causes discomfort, a patient is more likely to drop out from the study. If discomfort is not measured in the study, the missing data is MNAR.

Missing data mechanisms

↪ Quoting Fitzmaurice (2008):

> "*In closing, it should be emphasized that assumptions about missing data are inherently difficult, if not possible, to verify from the data at hand. Consequently, whenever possible, researchers should go to great lengths to minimize the amount of missing data in their studies. In general, the potential for bias is somewhat greater when the proportion of missing data is relatively large.*"

Missing data mechanisms

Exercise (from Raghunathan, 2016, p. 23)

aleatória

↪ A survey is being conducted based on a random sample of firms from the population list which has the name and size of the firm (number of employees). A key survey variable of interest is whether or not the firm offers health insurance to its employees and the number of health plans offered. Consider the following missing data mechanisms:

(a) All firms exceeding some certain firm size refuse to participate.

(b) Firms that do not offer health plans and/or very limited number of plans are more likely to be nonrespondents.

State whether the plans are MCAR, MAR, or MNAR.

Missing data patterns

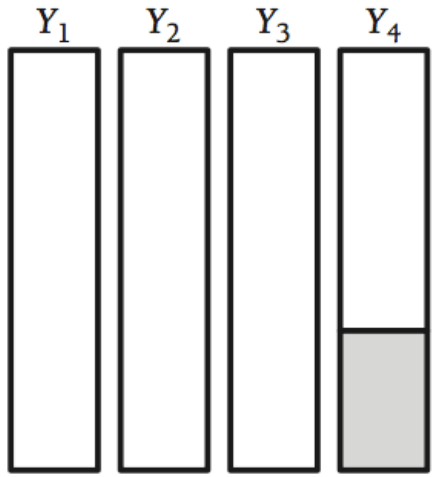
- A pattern of missing data describes the location of the missing values in a dataset.
- The missing data pattern describes the location of the 'holes' in the data but says nothing about why the data are missing.
- The pattern of missing values plays an important role with respect to the theoretical justification and the application of techniques for dealing with missing values.

Missing data patterns

- In a **univariate** pattern, data are missing only in one variable.
- For example, one could be interested in the relationship between the number of children living in a household and hourly wage.
- Suppose further that all households report the number of children but hourly wage is not observed for all households.

Missing data patterns

↪ A simple of extension of the previous bivariate example, as depicted in the figure below, includes the case where there are more than two variables, but only one variable is not completely observed.

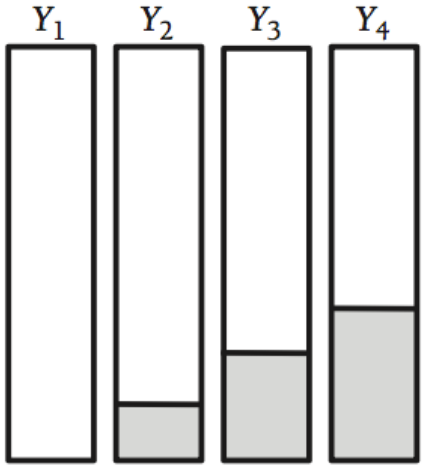


The shaded areas represent the location of the missing values in the data set. We are assuming a rectangular data matrix with rows representing subjects and columns representing variables. Figure from Enders, 2010, p. 4.

↪ The univariate pattern was one of the earliest missing data problems to receive attention in the literature.

Missing data patterns

↪ The figure below shows a **monotone** missing pattern.



The shaded areas represent the location of the missing values in the data set. Figure from Enders, 2010, p. 4.

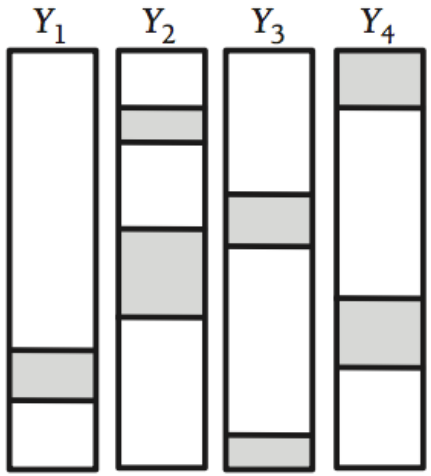
↪ A missing data pattern is called monotone if the dataset can be arranged by sorting rows and/or columns such that going from left to right if a missing value occurs in a row, all the following values in that row are missing as well.

Missing data patterns

- Visually, a monotone pattern resembles a staircase.
- A monotone missing data pattern is typically associated with a longitudinal study where participants drop out and never return.
- For example, consider a clinical trial for a new medication in which participants quit the study because they are having adverse reactions to the drug.

Missing data patterns

↪ An **arbitrary/general** pattern in which any set of variables may be missing for any subject is shown in the figure below.



The shaded areas represent the location of the missing values in the data set. Figure from Enders, 2010, p. 4.

↪ This pattern corresponds to the most common configuration of missing data and cannot be reduced to a univariate or monotone pattern.

Missing data patterns

- As a simple example, consider again the two variable example involving the number of children living in a household and hourly wage.

- The missing data pattern would be arbitrary if for some households the number of children is missing but hourly wage is observed and also, in contrast, for other households the number of children is missing but hourly wage is observed.