Missing value jargon

James Carpenter & Mike Kenward

Department of Medical Statistics
London School of Hygiene & Tropical Medicine

James.Carpenter@lshtm.ac.uk
https://missingdata.lshtm.ac.uk

June 2005

Table of Contents

Some notation

The data

We denote the *data we intended to collect*, by Y, and we partition this into

$$\mathsf{Y} = \{\mathsf{Y}_o, \mathsf{Y}_m\}.$$

where Y_o is observed and Y_m is missing.

Note that some variables in Y may be outcomes/responses, some may be explanatory variables/covariates. Depending on the context these may all refer to one unit, or to an entire dataset.

Some notation

Missing value indicator

Corresponding to every observation Y, there is a missing value indicator R, defined as:

$$R = \begin{cases} 1 & \text{if } Y \text{ observed} \\ 0 & \text{if } Y \text{ missing} \end{cases}$$

with R corresponding to Y.

Missing value mechanism

The key question for analyses with missing data is, under what circumstances, if any, do the analyses we would perform if the data set were fully observed lead to valid answers?

As before, 'valid' means that effects and their SE's are consistently estimated, tests have the correct size, and so on, so inferences are correct.

The answer depends on the *missing value mechanism.*
This is the *probability that a set of values are missing given the values taken by the observed and missing observations*, which we denote by

$$\mathrm{Pr}(\mathrm{R} \mid \mathrm{y}_o, \mathrm{y}_m)$$

Examples of missing value mechanisms

1. The chance of nonresponse to questions about income usually depend on the person's income.
2. Someone may not be at home for an interview because they are at work.
3. The chance of a subject leaving a clinical trial may depend on their response to treatment.
4. A subject may be *removed* from a trial if their condition is insufficiently controlled.

Missing Completely at Random (MCAR)

Suppose the probability of an observation being missing does not depend on observed or unobserved measurements. In mathematical terms, we write this as

$$\mathrm{Pr}(\mathrm{r} \mid \mathrm{y}_o, \mathrm{y}_m) = \mathrm{Pr}(\mathrm{r})$$

Then we say that the observation is *Missing Completely At Random*, which is often abbreviated to MCAR.

Note that in a sample survey setting MCAR is sometimes called *uniform non-response*.

Missing Completely at Random (MCAR)

If data are MCAR, then consistent results with missing data can be obtained by performing the analyses we would have used had their been no missing data, although there will generally be some loss of information. In practice this means that, under MCAR, the analysis of only those units with complete data gives valid inferences.

An example of a MCAR mechanism would be that a laboratory sample is dropped, so the resulting observation is missing.

Missing Completely at Random (MCAR)

However, many mechanisms that initially seem to be MCAR may turn out not to be. For example, a patient in a clinical trial may be lost to follow up after 'falling' under a bus; however if it is a psychiatric trial, this may be an indication of poor response to treatment. Likewise, if a response to a postal questionnaire is missing because the questionnaire was lost or stolen in the post, this may not be random but rather reflect the area in which the sorting office is located.

Missing Completely at Random (MCAR)

As we have already said, under MCAR analyses of *completers* only (a short hand for including in the analysis only units with fully observed data) give valid inferences.

So do analyses based on moment based estimators (for example, generalised estimating equations), and other estimators derived from *consistent estimating equations*.

By *consistent estimating equations* we mean functions of the data and unknown parameters whose expectation, taken over the complete data at the population parameter values, is zero. Under MCAR, they still have expectation zero, and so still lead to valid inferences.

Missing Completely at Random (MCAR)

Saying the same thing mathematically, an estimating equation can be written as $U(\mathrm{y}, \theta)$, and at the estimate $\hat{\theta}$, $U(\mathrm{y}, \hat{\theta}) = 0$.

The estimating equation is consistent because $\mathbf{E}\left[U(\mathrm{Y}, \theta)\right] = 0$ (where θ is the population parameter value). It remains consistent if the data are missing completely at random (MCAR) because, even then, still $\mathbf{E}\left[U(\mathrm{Y}_o, \theta)\right] = 0$.
bigskip
A simple example of a consistent is estimating equation is the sample mean, $U(\mathrm{y}, \theta) = \bar{y} - \theta$.

Missing At Random (MAR)

After considering MCAR, a second question naturally arises. That is, what are the most general conditions under which a valid analysis can be done using only the observed data, and no information about the missing value mechanism, $\Pr(r \mid y_o, y_m)$?

The answer to this is when, *given the observed data, the missingness mechanism does not depend on the unobserved data*. Mathematically,

$$\Pr(r \mid y_o, y_m) = \Pr(r \mid y_o).$$

This is termed *Missing At Random*, abbreviated MAR.
This is equivalent to saying that the behaviour of two units who share *observed values* have the same statistical behaviour on the other observations, whether observed or not.

Missing At Random (MAR)

For example:

	Variables					
Unit	1	2	3	4	5	6
1	1	3	4.3	3.5	1	4.6
2	1	3	?	3.5	?	?

As units 1 and 2 have the same values where both are observed, given these observed values, under MAR, variables 3, 5 and 6 from unit 2 have the same distribution (NB not the same value!) as variables 3, 5 and 6 from unit 1.

Missing At Random (MAR)

Note that under MAR the probability of a value being missing will generally depend on observed values, so it does not correspond to the intuitive notion of 'random'. The important idea is that the missing value mechanism can expressed solely in terms of *observations that are observed.*

Unfortunately, this can rarely be definitively determined from the data at hand!

Examples of MAR mechanisms

- A subject may be *removed* from a trial if his/her condition is not controlled sufficiently well (according to pre-defined criteria on the response).
- Two measurements of the same variable are made at the same time. If they differ by more than a given amount a third is taken. This third measurement is missing for those that do not differ by the given amount.

Examples of MAR mechanisms

A special case of MAR is *uniform non-response within classes*. For example, suppose we seek to collect data on income and property tax band. Typically, those with higher incomes may be less willing to reveal them. Thus, a simple average of incomes from respondents will be downwardly biased.

However, now suppose we have everyone's property tax band, and *given property tax band* non-response to the income question is random. Then, the income data is missing at random; the reason, or mechanism, for it being missing depends on property band. Given property band, *missingness does not depend on income itself.*

Examples of MAR mechanisms

Therefore, to get an unbiased estimate of income, we first average the observed income within each property band. As data are missing at random given property band, these estimates will be valid. To get an estimate of the overall income, we simply combine these estimates, weighting by the proportion in each property band. In this example, a simple summary statistic (average of observed incomes) was biased. Conversely, a simple model (estimate of income conditional on property band), where we condition on the variable that makes the data MAR, led to a valid result.

Examples of MAR mechanisms

This is an example of a more general result. Methods based on the *likelihood* are valid under MAR. However, in general *non-likelihood* methods (e.g. based on completers, moments, estimating equations & including generalised estimating equations) are not valid under MAR, although some can be 'fixed up'. In particular, ordinary means, and other simple summary statistics from observed data, will be *biased*.

Finally, note that in a likelihood setting the term *ignorable* is often used to refer to and MAR mechanism. It is the mechanism (i.e. the model for $\Pr(R \mid y_o)$) which is ignorable - not the missing data!

Missing Not At Random (MNAR)

When neither MCAR nor MAR hold, we say the data are *Missing Not At Random*, abbreviated MNAR.
In the likelihood setting (see end of previous section) the missingness mechanism is termed *non-ignorable*.
What this means is

1. Even accounting for all the available observed information, the reason for observations being missing *still depends on the unseen observations themselves*.
2. To obtain valid inference, a *joint model* of both Y and R is required (that is a joint model of the data and the missingness mechanism).

Missing Not At Random (MNAR)

Unfortunately

1. We cannot tell from the data at hand whether the missing observations are MCAR, NMAR or MAR (although we can distinguish between MCAR and MAR).
2. In the MNAR setting it is very rare to know the appropriate model for the missingness mechanism.

Hence the central role of sensitivity analysis; we must explore how our inferences vary under assumptions of MAR, MNAR, and under various models. Unfortunately, this is often easier said than done, especially under the time and budgetary constraints of many applied projects.

Summary

We have defined, in non-technical language, the commonly used terms MCAR, MAR and NMAR, together with *ignorable* and *non-ignorable*.

Summary

We have seen that

1. The implications of missingness for the analysis depend on the *missing value mechanism* , which is rarely known.
2. The intuitive notion of randomness for the missing value mechanism is called *Missing Completely at Random (MCAR).* A wide range of analyses are valid under the assumption of MCAR.
3. A special intermediate case between 'missing completely at random' and 'not missing at random' is *Missing at Random (MAR)*.
   Assuming MAR, particular analyses that ignore the missing value mechanism are valid under MAR (e.g. likelihood) and others can be fixed up (e.g. estimating equations can be fixed up by weighting).
4. In most situations, the true mechanism is probably MNAR.

Summary

Important

1. We cannot tell from the data at hand whether the missing observations are MCAR, NMAR or MAR (although we can distinguish between MCAR and MAR).
2. In the MNAR setting it is very rare to know the appropriate model for the missingness mechanism.

Hence the central role of sensitivity analysis; we must explore how our inferences vary under assumptions of MAR, MNAR, and under various models. Unfortunately, this is often easier said than done, especially under the time and budgetary constraints of many applied projects.