Missing data 3. How to explore missing data

Tra My Pham, Nikolaos Pandis, Ian R White

This article explains how one should explore missing data in preparation for performing an analysis. The first step in an analysis with missing data is to quantify the extent of missingness in our data set, e.g. by tabulating the amount of missing data in each variable and the patterns of missing data across variables.

The percentages of missing values in each variable can be used to identify possible errors in the data collection or processing, and may help us identify variables that we might omit from analysis.

The patterns of missingness describe the location of the missing values. A missingness pattern is said to be univariate if there is only one variable with missing data. If there are several incomplete variables in the data set, the missingness pattern is said to be multivariate. Table 1 provides two examples of multivariate missingness patterns in an analysis involving an outcome variable measured repeatedly at three time points. Table 1a presents a monotone missingness pattern, where the missingness of variables can be ordered in a way that if one variable is missing then all subsequent variables are also missing. This occurs, for example, when an individual drops out of a longitudinal study, after which point all subsequent measurements of their outcome variable are not available. When the missingness of variables cannot be ordered in this way, the missingness pattern is called non-monotone, as presented in Table 1b. The tabulation of the missingness pattern can be used to identify possible errors in the data collection and processing, e.g. when we might expect one variable to be always observed if another one is observed.

Table 1. Examples of multivariate missingness patterns. Observed values are denoted with an x; missing values are denoted with a dot

a. Monotone

% individuals	Outcome (time 1)	Outcome (time 2)	Outcome (time 3)
80	x	x	x
10	x	x	.
10	x	.	.

b. Non-monotone

% individuals	Outcome (time 1)	Outcome (time 2)	Outcome (time 3)
80	x	x	x
10	x	.	x
10	.	x	.

In articles 1 and 2 ***[add refs at proof stage – adjust the refs section accordingly]***, we saw that missing data mechanisms can be described as *missing completely at random* (MCAR), *missing at random* (MAR), or *missing not at random* (MNAR). We cannot verify these assumptions about the missing values from the observed data alone, but it can still be useful to explore the observed data.[1]

In this article, we continue with an example created using data from a randomised controlled trial assessing the evolution of probing depth on the lower six anterior teeth bonded with two types of lingual retainers over time.[2] Here, we look at data on age at baseline, *age25* (≥25/<25 years old, fully observed), and mean probing depth at time 1, *mean_pd1* (partially observed).

The observed data can help us determine whether the missing values in *mean_pd1* are MCAR or not MCAR. Let us suppose that in this example, the trial collects data from 129 individuals, 65 of whom are below 25 years of age, and the other 64 individuals are 25 years or older. Values in mean probing depth at time 1 are missing for some individuals, as presented in Table 2. We refer to the outcome as the mean probing depth because we chose, for reasons of simplicity, to use the average probing depth across the six teeth per individual.

Table 2. Number of individuals with observed and missing mean probing depth by age

Age	Mean probing depth observed	Mean probing depth missing	Total
≥ 25 years	51	13	64
<25 years	35	30	65

We can check whether the proportion of individuals whose mean probing depth is missing differs between the two age groups. We can perform a hypothesis test of an association between missingness in mean probing depth, *mean_pd1_miss*, and age group, *age25*, e.g. with a Pearson χ^2 test. If there is evidence that missingness in mean probing depth varies by age group, we have evidence against mean probing depth being MCAR. Here the Pearson $\chi^2(1) = 9.69$, p=0.002, suggesting there is evidence against the null hypothesis of no association between *mean_pd1_miss* and *age25*. Alternatively, we could fit a logistic regression model with *mean_pd1_miss* as the dependent variable and *age25* as the independent variable, and test for an association (estimated odds ratio=3.36, 95% confidence interval 1.54 to 7.34, Wald p=0.002). We can also use logistic regression to explore the association between missingness and more than one variable in other more complex settings.

In this example, we can see that mean probing depth is more likely to be missing in younger individuals. We could use this observation to motivate improving our data collection procedures in younger individuals. From an analysis point of view, this observation means that the data are not MCAR, and that we need to account for age in the analysis.

While we have seen evidence against MCAR, we still cannot tell whether *mean_pd1* is MAR or MNAR. In order to distinguish between MAR and MNAR, we would need to know if *mean_pd1_miss* depends on *mean_pd1* within each age group, but unfortunately we do not observe *mean_pd1* when *mean_pd1_miss* is equal to 0. Therefore, as alluded to in the first two articles, we cannot

determine if MAR holds or not from the observed data alone. In some instances, such as in studies where the outcome variable is measured repeatedly over time, we could investigate the plausibility of MAR further by cross-tabulating missingness in the outcome at one time point against values of the outcome at a previous time. If individuals with poor outcome values at one time tend to be missing at the next time, this could suggest that their outcome values are even worse at that time, indicating a potential MNAR mechanism.

In practice, MAR can be made more plausible by including in the analysis variables that are predictive of both the missingness and missing values. For example, if age is associated with higher probing depth values and probing depth is more likely to be missing for younger individuals, then including age in the analysis will improve the plausibility of the MAR assumption. Assessing sensitivity of the results to the chosen MAR assumption by considering plausible departures from MAR is key to any analysis with missing data.[3]

The next article will discuss another important aspect of the data that helps to guide the analysis: whether the missing data are in the exposure or the outcome of the model that is to be fitted to the data.

References

1. Carpenter JR, Kenward MG. *Multiple imputation and its application*. 1st ed. Chichester, West Sussex: John Wiley & Sons, Ltd., 2013.
2. Węgrodzka E, Kornatowska K, Pandis N, et al. A comparative assessment of failures and periodontal health between 2 mandibular lingual retainers in orthodontic patients. A 2-year follow-up, single practice-based randomized trial. *Am J Orthod Dentofac Orthop* 2021; 160: 494-502.e1.
3. Cro S, Morris TP, Kenward MG, et al. Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation: a practical guide. *Stat Med* 2020; sim.8569.