Introduction: Issues Raised by Missing Data

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

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Table of Contents

Introduction

Missing data are common.

However, they are usually inadequately handled in both epidemiological and experimental research.

For example, Wood, White and Thompson (2014) reviewed 71 recently published BMJ, JAMA, Lancet and NEJM papers.

- 89% had partly missing outcome data.
- In 37 trials with repeated outcome measures, 46% performed complete case analysis.
- Only 21% reported sensitivity analysis.

What do we mean by missing data?

Missing data are simply observations that we intended to be made but did not. For example, an individual may only respond to certain questions in a survey, or may not respond at all to a particular wave of a longitudinal survey.

In the presence of missing data, our goal remains making inferences that apply to the population targeted by the complete sample - i.e. the goal remains what it was if we had seen the complete data.

What do we mean by missing data?

However, both making inferences and performing the analysis are now more complex. We will see we need to make assumptions in order to draw inferences, and then use an appropriate computational approach for the analysis.

We will avoid adopting computationally simple solutions (such as just analysing complete data or carrying forward the last observation in a longitudinal study) which generally lead to misleading inferences.

What do we mean by missing data?

In practice the data consist of: (a) the observations actually made (where "?" denotes a missing observation):

	Variables							
Unit	1	2	3	4	5	6	7	.....
1	1	4	1	3.4	5.67	A	8.251	.....
2	1	3	?	?	5.67	B	9.253	.....
3	1	2	1	2.7	5.72	B	12.812	.....
4	1	1	1	3.6	5.13	?	13.614	.....
5	2	?	1	?	?	A	11.4422	.....
6	2	2	1	3.4	5.61	A	9.241	.....
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

Figure 1: Typical partially observed data set

What do we mean by missing data?

and (b): the pattern of missing values:

	Variables							
Unit	1	2	3	4	5	6	7	------
1	1	1	1	1	1	1	1	------
2	1	1	0	0	1	1	1	------
3	1	1	1	1	1	1	1	------
4	1	1	1	1	1	0	1	------
5	1	0	1	0	0	1	1	------
6	1	1	1	1	1	1	1	------
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

Figure 2: Pattern of missing values for the data in Figure 1. A '1' indicates that an observation is seen, a '0' that it is missing

Inferential framework

When it comes to analysis, whether we adopt a frequentist approach (Figure 3) or a Bayesian approach (Figure 4), the likelihood is central.

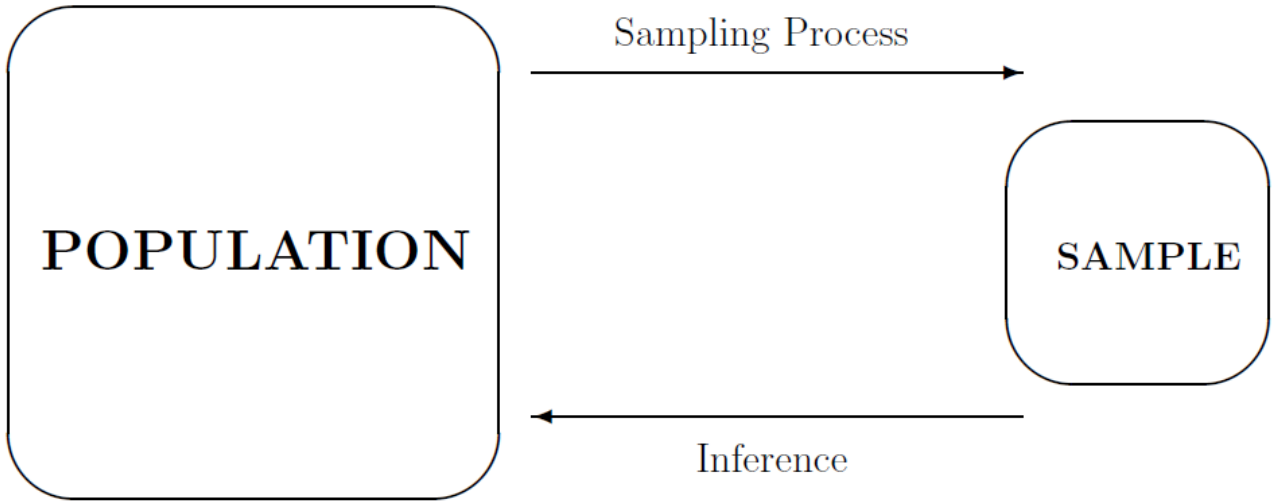

Figure 3: Schematic for frequentist (sometimes termed traditional) paradigm of inference

Inferential framework

In these notes, for convenience, we discuss issues from a frequentist perspective, although often we use appropriate Bayesian computational strategies to approximate frequentist analyses.

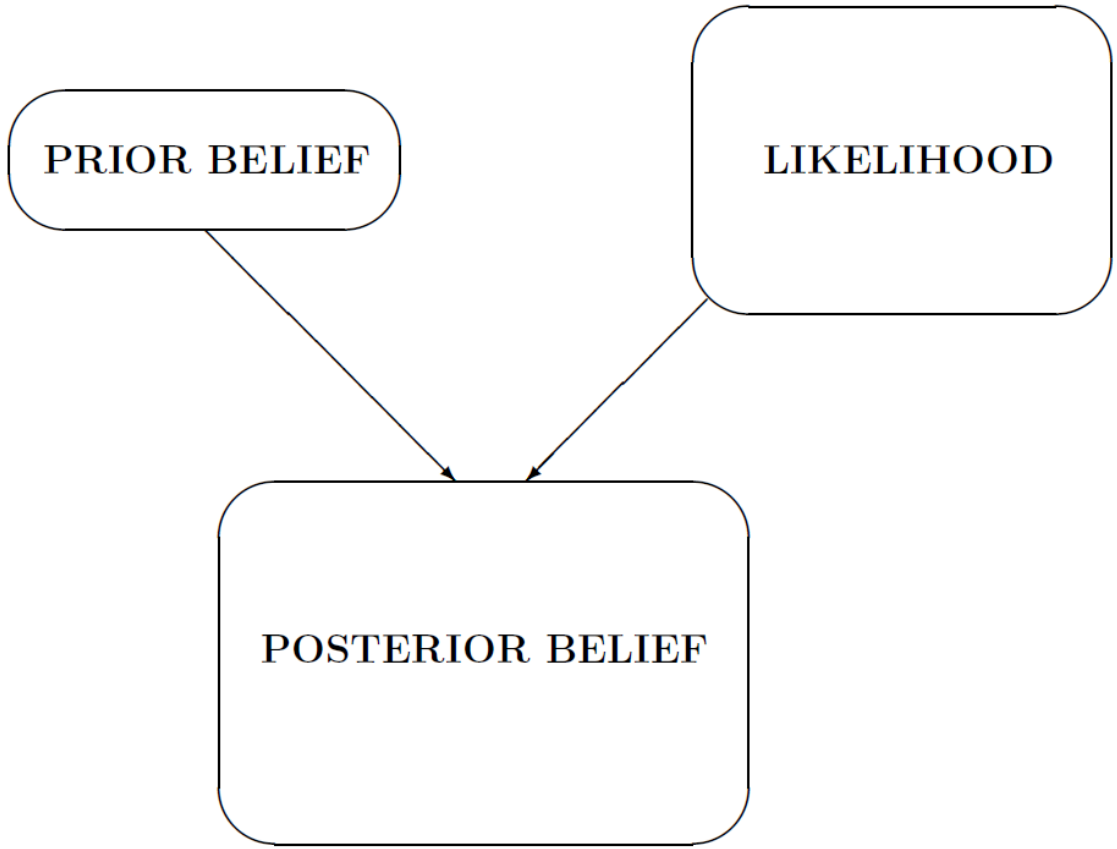

Figure 4: Schematic for Bayesian paradigm of inference

Inferential framework

The actual sampling process involves the 'selection' of the missing values, as well as the units. So to complete the process of inference in a justifiable way we need to take this into account.

The *likelihood* is a measure of comparative support for different models given the data. It requires a model for the observed data, and as with classical inference this must involve aspects of the way in which the missing data have been selected (i.e. the missingness mechanism).

What do we mean by valid inference when we have missing data?

We have already noted that missing data are observations we intended to make but did not.

Thus, the sampling process now involves both the selection of the units, AND ALSO the process by which observations become missing - the *missingness mechanism*.

It follows that for valid inference, we need to take account of the missingness mechanism.

What do we mean by valid inference when we have missing data?

By *valid inference* in a frequentist framework we mean that the quantities we calculate from the data have the usual properties. In other words, estimators are consistent, confidence intervals attain nominal coverage, p-values are correct under the null hypothesis, and so on.

Assumptions

We distinguish between item and unit nonresponse (missingness). For item missingness, values can be missing on *response* (i.e. outcome) variables and/or on *explanatory* (i.e. design/covariate/exposure/confounder) variables.

Missing data can effect properties of *estimators* (for example, means, percentages, percentiles, variances, ratios, regression parameters and so on). Missing data can also affect inferences, i.e. the properties of tests and confidence intervals, and Bayesian posterior distributions.

Assumptions

A critical determinant of these effects is the way in which the *probability* of an observation being missing (the *missingness mechanism*) depends on other variables (measured or not) and on *its own value*.

In contrast with the sampling process, which is usually known, the missingness mechanism is usually unknown.

The data alone cannot usually definitively tell us the sampling process.

Likewise, the missingness pattern, and its relationship to the observations, cannot definitively identify the missingness mechanism.

Assumptions

The additional assumptions needed to allow the *observed* data to be the basis of inferences that would have been available from the *complete* data can usually be expressed in terms of either

1. the relationship between selection of missing observations and the values they would have taken, *or*
2. the statistical behaviour of the unseen data.

These additional assumptions are *not* subject to assessment from the data under analysis; their plausibility cannot be definitively determined from the data at hand.

Assumptions

The issues surrounding the analysis of data sets with missing values therefore centre on **assumptions**. We have to

1. decide which assumptions are reasonable and sensible in any given setting;
   - contextual/subject matter information will be central to this
2. ensure that the assumptions are transparent;
3. explore the sensitivity of inferences/conclusions to the assumptions, and
4. understand which assumptions are associated with particular analyses.

Getting computation out of the way

The above implies it is sensible to use approaches that make weak assumptions, and to seek computational strategies to implement them.

However, often computationally simple strategies are adopted, which make strong assumptions, which are subsequently hard to justify.

Classic examples are completers analysis (i.e. only including units with fully observed data in the analysis) and last observation carried forward. The latter is sometimes advocated in longitudinal studies, and replaces a unit's unseen observations at a particular wave with their last observed values, irrespective of the time that has elapsed between the two waves.

Conclusion

Missing data introduce an element of ambiguity into statistical analysis, which is different from the traditional sampling imprecision. While sampling imprecision can be reduced by increasing the sample size, this will usually only increase the number of missing observations!

As discussed in the preceding sections, the issues surrounding the analysis of incomplete datasets turn out to centre on *assumptions* and *computation*.

The assumptions concern the relationship between the reason for the missing data (i.e. the process, or mechanism, by which the data become missing) and the observations themselves (both observed and unobserved).

Conclusion

Unlike say in regression, where we can use the residuals to check on the assumption of normality, these assumptions cannot be verified from the data at hand.

Sensitivity analysis, where we explore how our conclusions change as we change the assumptions, therefore have a central role in the analysis of missing data.

Simple 'ad-hoc' methods, discussed in the next document, should therefore usually be avoided in practice.