STATISTICS NOTES

Missing data

Douglas G Altman[1], J Martin Bland[2]

[1]Cancer Research UK/NHS Centre for Statistics in Medicine, Oxford OX2 6UD
[2]Department of Health Sciences, University of York, York YO10 5DD
Correspondence to: Professor Altman
doug.altman@cancer.org.uk

BMJ 2007;334:424
doi:10.1136/bmj.38977.682025.2C

Almost all studies have some missing observations. Yet textbooks and software commonly assume that data are complete, and the topic of how to handle missing data is not often discussed outside statistics journals.

There are many types of missing data and different reasons for data being missing. Both issues affect the analysis. Some examples are:

(1) In a postal questionnaire survey not all the selected individuals respond;
(2) In a randomised trial some patients are lost to follow-up before the end of the study;
(3) In a multicentre study some centres do not measure a particular variable;
(4) In a study in which patients are assessed frequently some data are missing at some time points for unknown reasons;
(5) Occasional data values for a variable are missing because some equipment failed;
(6) Some laboratory samples are lost in transit or technically unsatisfactory;
(7) In a magnetic resonance imaging study some very obese patients are excluded as they are too large for the machine;
(8) In a study assessing quality of life some patients die during the follow-up period.

The prime concern is always whether the available data would be biased. If the fact that an observation is missing is unrelated both to the unobserved value (and hence to patient outcome) and the data that are available this is called "missing completely at random." For cases 5 and 6 above that would be a safe assumption. Sometimes data are missing in a predictable way that does not depend on the missing value itself but which can be predicted from other data—as in case 3. Confusingly, this is known as "missing at random." In the common cases 1 and 2, however, the missing data probably depend on unobserved values, called "missing not at random," and hence their lack may lead to bias.

In general, it is important to be able to examine whether missing data may have introduced bias. For example, if we know nothing at all about the non-responders to a survey then we can do little to explore possible bias. Thus a high response rate is necessary for reliable answers.[1] Sometimes, though, some information is available. For example, if the survey sample is chosen from a register that includes age and sex, then the responders and non-responders can be compared on these variables. At the very least this gives some pointers to the representativeness of the sample. Non-responders often (but not always) have a worse medical prognosis than those who respond.

A few missing observations are a minor nuisance, but a large amount of missing data is a major threat to a study's integrity. Non-response is a particular problem in pair-matched studies, such as some case-control studies, as it is unclear how to analyse data from the unmatched individuals. Loss of patients also reduces the power of the trial. Where losses are expected it is wise to increase the target sample size to allow for losses. This cannot eliminate the potential bias, however.

Missing data are much more common in retrospective studies, in which routinely collected data are subsequently used for a different purpose.[2] When information is sought from patients' medical notes, the notes often do not say whether or not a patient was a smoker or had a particular procedure carried out. It is tempting to assume that the answer is no when there is no indication that the answer is yes, but this is generally unwise.

No really satisfactory solution exists for missing data, which is why it is important to try to maximise data collection. The main ways of handling missing data in analysis are: *(a)* omitting variables which have many missing values; *(b)* omitting individuals who do not have complete data; and *(c)* estimating (imputing) what the missing values were.

Omitting everyone without complete data is known as complete case (or available case) analysis and is probably the most common approach. When only a very few observations are missing little harm will be done, but when many are missing omitting all patients without full data might result in a large proportion of the data being discarded, with a major loss of statistical power. The results may be biased unless the data are missing completely at random. In general it is advisable not to include in an analysis any variable that is not available for a large proportion of the sample. The main alternative approach to case deletion is imputation, whereby missing values are replaced by some plausible value predicted from that individual's available data. Imputation has been the topic of much recent methodological work; we will consider some of the simpler methods in a separate *Statistics Note.*

Competing interests: None declared.

1 Evans SJW. Good surveys guide. *BMJ* 1991;302:302-3.
2 Burton A, Altman DG. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *Br J Cancer* 2004;91:4-8.

This is part of a series of occasional articles on statistics and handling data in research.