Incomplete Data Analysis

Vanda Inácio

University of Edinburgh

Semester 1, 2020/2021

Course's scope

The goal of this course is to provide a comprehensive overview of statistical methods for analysing data in the presence of incomplete/missing data. Application of methods will be done using `R`.

- ↪ Introduction to the problematic of missing data.
- ↪ Traditional/naive methods for dealing with missing data (and their drawbacks).
- ↪ Likelihood based methods for handling missing data.
- ↪ Multiple imputation.

Incomplete (multivariate) data

Gender	**Age**	**Income**	**...**
M	28	80,000	...
F	?	?	...
?	43	?	...
F	?	100,000	...

Remark

↪ Throughout this course we must be aware that we will be looking at **'second-best' solutions** to the missing data problem, as none of the approaches/methods will be better than having the complete dataset intended to be collected.

↪ Quoting Allison (2001):

"*The only really good solution to the missing data problem is not to have any. So in the design and execution of research projects, it is essential to put great effort into minimising the occurrence of missing data. Statistical adjustments can never make up for sloppy research.*"

Introduction to missing data

- ↪ The goal of this course is to learn about statistical approaches and methods for data analysis in the presence of incomplete or missing data.
- ↪ In many practical settings of interest, such as, sample surveys, clinical trials, or agricultural experiments, data are to be collected according to a predetermined plan or design.
- ↪ The aim is, given the collected data, to make inference about some aspect of the underlying population of interest.
- ↪ However, due to a number of different reasons, data that were intended to be collected (and thus to contribute to our inferential objective) are actually not collected or not available.
- ↪ The next three examples, adapted from Professor Davidian's notes (chapter 1), illustrate different situations in which missing data can occur.

Introduction to missing data

Non or partial response in sample surveys

- ↪ Surveys are very much prone to the occurrence of missing data.
- ↪ Suppose, for instance, that a survey is to be conducted to estimate the proportion of the population likely to vote for a certain candidate or to estimate features of the income's distribution in a certain population of households.
- ↪ A random sample of subjects from the population is to be contacted (e.g., personally or via phone) or, alternatively, a questionnaire may be sent (e.g., by mail or e-mail).
- ↪ However, some members of the sample may not answer the phone or refuse to respond, or some may fail to send back the questionnaire or answer only a subset of the questions (partial respond).
- ↪ When this happens, for such individuals the response of interest (candidate preference, household income) will not be available (i.e., will be **missing**).

Introduction to missing data

Dropout and noncompliance in clinical trials

- ↪ A clinical trial may be conducted with the purpose of comparing the efficacy of say, two treatments, in a certain population.
- ↪ Usually, the clinical procedure is as follows: 1) individuals are recruited and enrolled in the study and are assigned (typically in a random fashion) to one of the treatments, 2) should take the treatment as prescribed and should return to the clinic on a regular basis (e.g., weekly), at which times relevant outcomes are measured and recorded.
- ↪ However, some individuals, beyond a certain point in time, may not show up for any clinic visit, thus 'dropping out' of the study. Others may quit taking the prescribed treatment as instructed or simply quit it at all. Still others may miss visits sporadically.
- ↪ In such a context, part of the intended full set of longitudinal outcomes arising from taking the prescribed treatment and visiting the clinic as directed will be missing for those subjects.

Introduction to missing data

Surrogate measurements and missing by design

- In a nutrition study, the daily average percent fat intake in a certain population may be of interest, and a (random) sample of subjects from the population is recruited to participate.
- Accurate measurement of long-term fat intake requires subjects to keep a detailed 'food diary' over a long period of time. As one can imagine, this can be extremely time-consuming.
- A simpler measure is to record all the food subjects have had in the last 24 hours.
- Implicitly, one is assuming that this 24 hour 'diary' may be correlated with the long term fat intake. Obviously, this 24 hour diary is not a perfect measure of the longer fat intake. It is instead an error prone measurement of it. Such measure is referred to be a surrogate for the complete detailed measure.

Introduction to missing data

Surrogate measurements and missing by design (continued)

- ↪ In order to reduce costs and subject burden, a study may be designed so that although all subjects provide a 24 hour recall measurement, only some of them provide the more expensive and time-consuming full diary measurement.

- ↪ Note that unlike the previous two examples where missingness was outside the control of the investigators, in this example the fact that some individuals are missing the full fat intake diary record is deliberate; that is, the missingness is by design.

Introduction to missing data

- ↪ It is universally accepted that most studies involving human subjects will have missing information for some subjects/variables of interest.
- ↪ This might be due to a variety of reasons, ranging from oversight mistakes by the personnel conducting the study to subjects refusing or being unable to provide the required information.
- ↪ As a consequence, incomplete/missing data are a routine challenge that often complicates the analyses.
- ↪ **Problem**: As mentioned earlier, interest usually lies in making inferences about some part of the distribution of the complete data that were intended to be collected and could be observed if no data were missing. When some of the intended full data are missing, depending on how and why they are missing, the validity of the inferences may be compromised (as we will see latter, they can be, for instance, biased). On the top of this, most statistical methods assume the data are completely observed.

Introduction to missing data

- One possible approach to data analysis when some portion of the data is missing, is simply to ignore the problem and analyse the observed data as if they were the intended complete data.
- As we will see later, this can lead to misleading conclusions in most of the situations.
- This motivates the need for statistical methods that acknowledge that some intended data are missing and that try to correct for this somehow.
- This need has led to an enormous literature on statistical models and methods for analysis in the presence of missing data and this continues to be a very active area of research.

Introduction to missing data

- ↪ Interestingly, although missing data have always been an issue in many, if not all, areas of application, it was not until the 1970s that it was properly handled.

- ↪ Without any doubt, we can say that the ‘game changer’ was the article published in the mid 1970s by Donald Rubin (Rubin, 1976), which laid out a framework for thinking about missing data, characterising formally how they can arise and elucidating their possible implications for inference.

- ↪ According to a search made on google scholar (today!!) this paper has 9805 citations, which is an impressive number for a statistical paper.

Introduction to missing data

Biometrika (1976), **63**, 3, *pp.* 581–92
Printed in Great Britain

Inference and missing data

BY DONALD B. RUBIN
Educational Testing Service, Princeton, New Jersey

SUMMARY

When making sampling distribution inferences about the parameter of the data, θ, it is appropriate to ignore the process that causes missing data if the missing data are 'missing at random' and the observed data are 'observed at random', but these inferences are generally conditional on the observed pattern of missing data. When making direct-likelihood or Bayesian inferences about θ, it is appropriate to ignore the process that causes missing data if the missing data are missing at random and the parameter of the missing data process is 'distinct' from θ. These conditions are the weakest general conditions under which ignoring the process that causes missing data always leads to correct inferences.

Some key words: Bayesian inference; Incomplete data; Likelihood inference; Missing at random; Missing data; Missing values; Observed at random; Sampling distribution inference.