Incomplete Data Analysis

Vanda Inácio

University of Edinburgh

Semester 1, 2018/2019

General information

- **Lecturer**: Vanda Inácio de Carvalho
- **Email**: Vanda.Inacio@ed.ac.uk
- **Office**: 4601, JCMB.
- **Lectures and location**: Monday, 9:00-10:40, Alrick, classroom 10.
- **Tutorials**: Tuesday, 9:00-9:50 (weeks 3, 5, 7, 9, 11), KB Centre, Lab level 3.

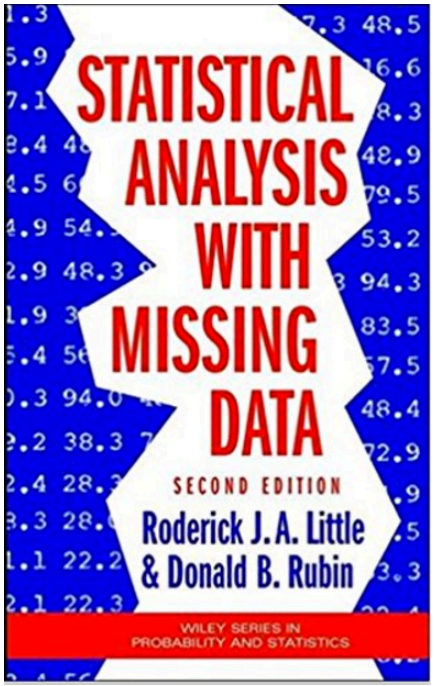
Assessment

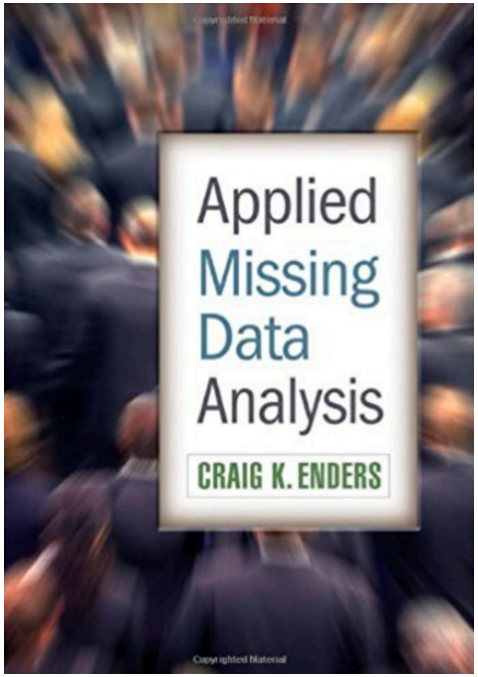
↪ **Exam**: closed-book written exam in December. Worth 95% of the final mark.

↪ **Coursework**: One assignment given in week 7. Solutions to be handed in week 9. Worth 5% of the final mark.

Bibliography

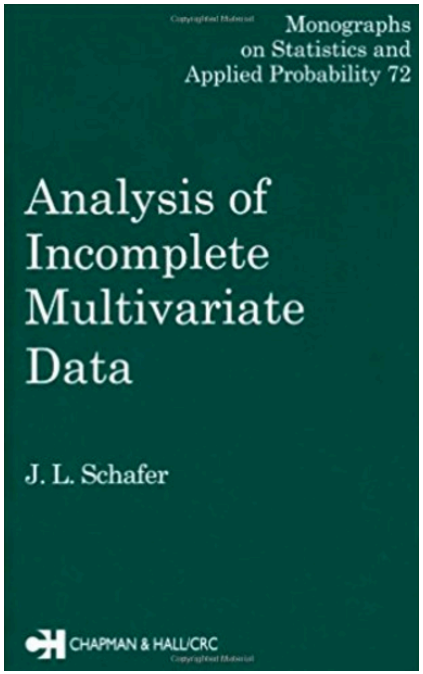
Covers from amazon.uk

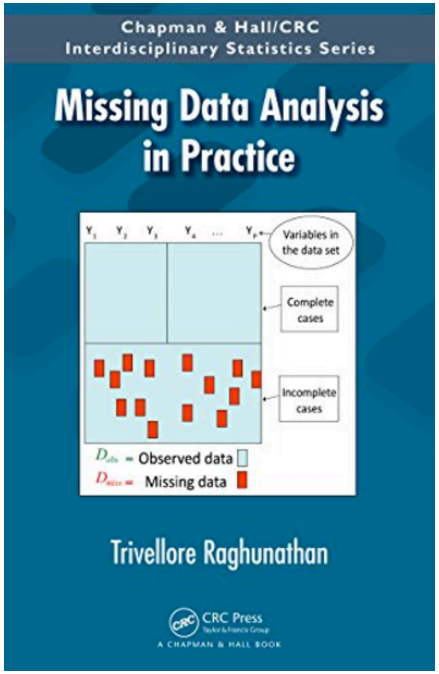


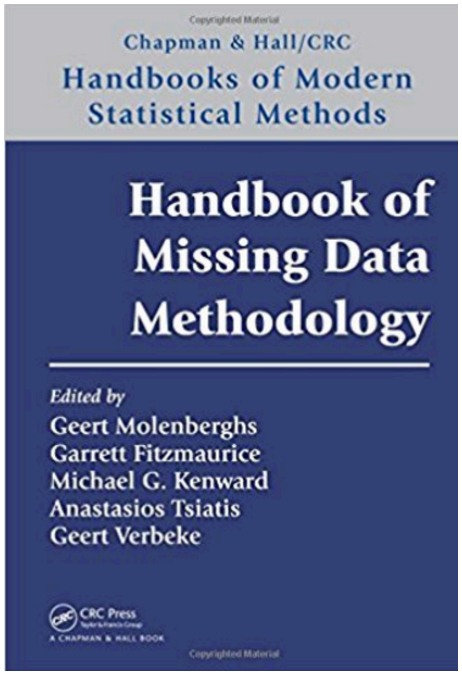


Bibliography

Covers from amazon.uk

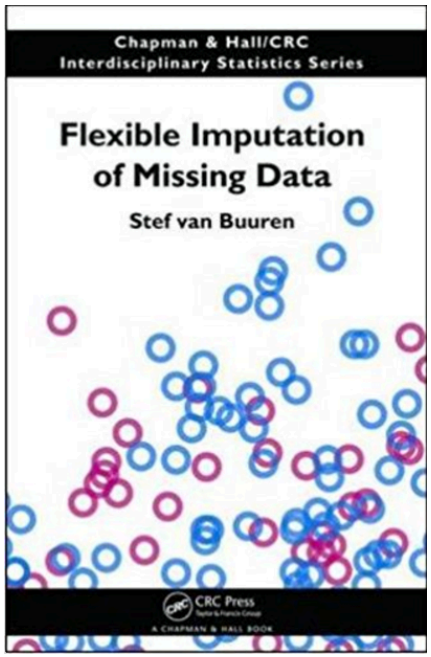


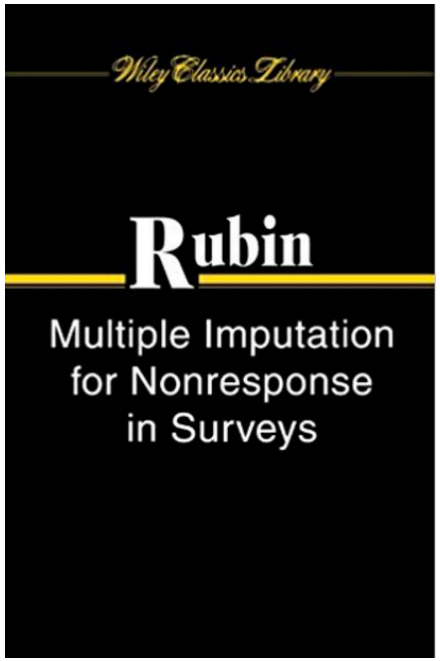


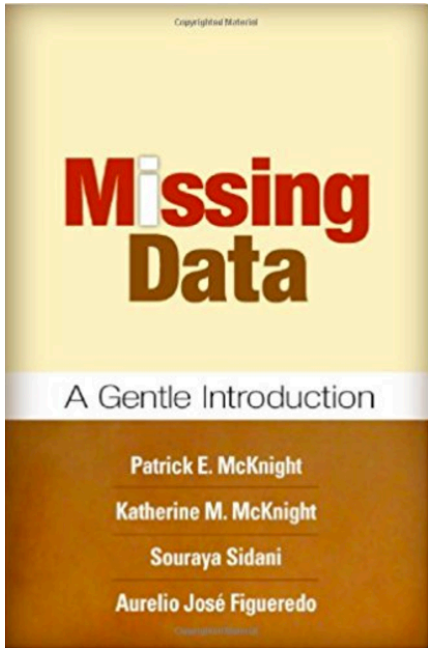


Bibliography

Covers from amazon.uk

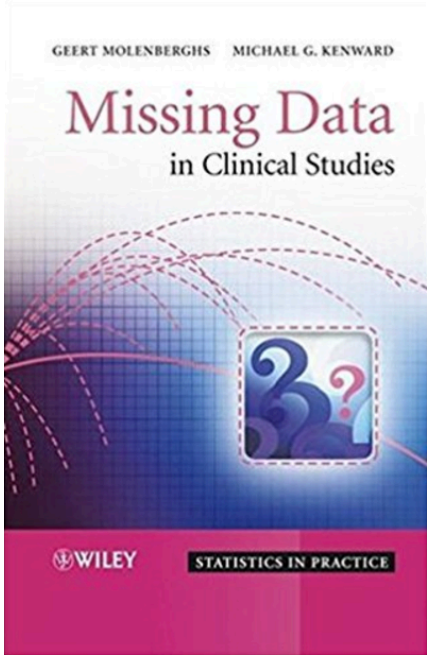


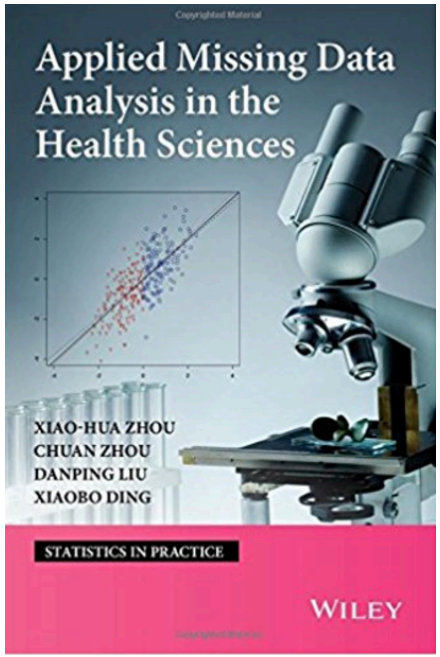


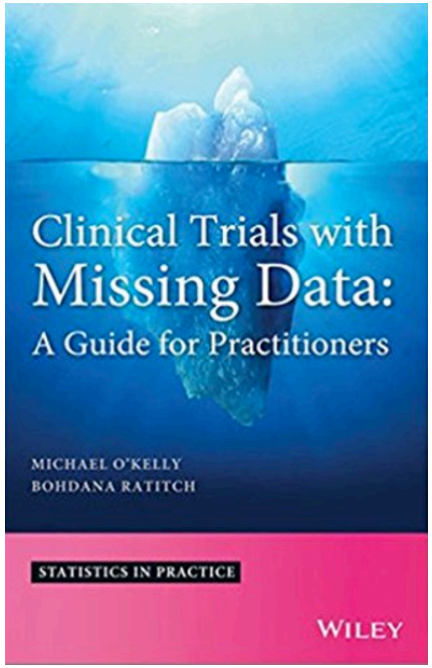


Bibliography

Covers from amazon.uk

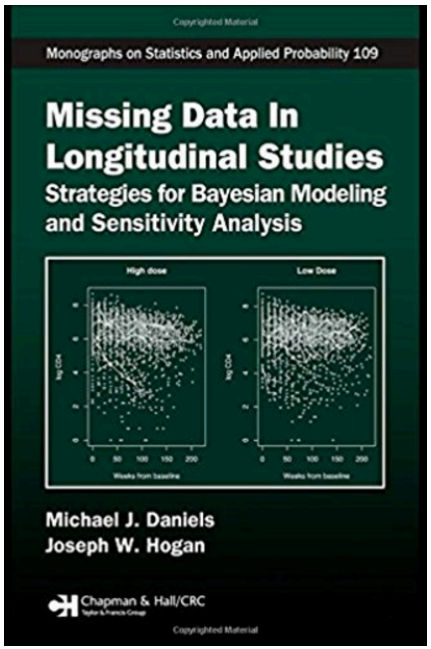


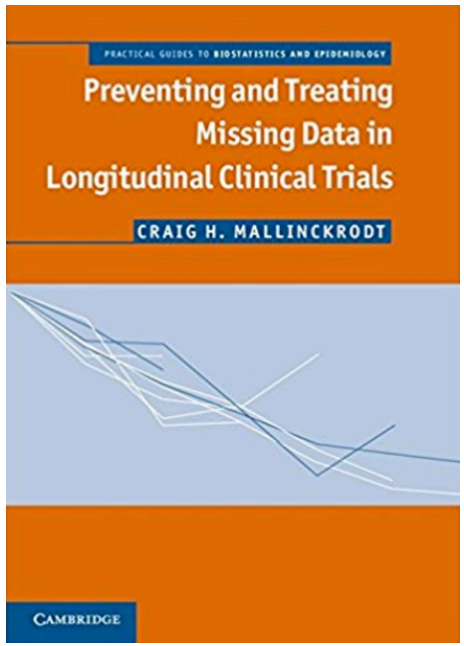


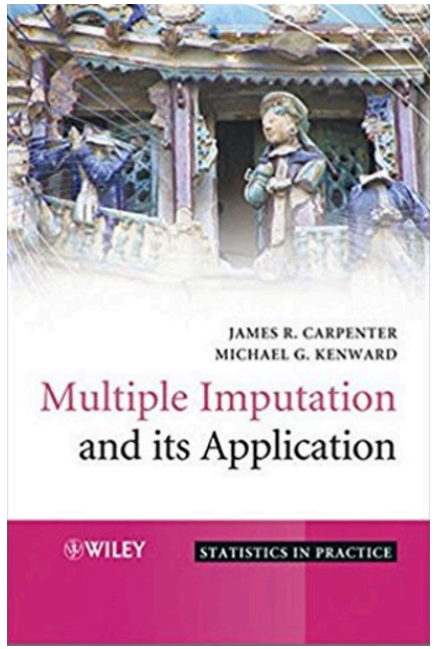


Bibliography

Covers from amazon.uk

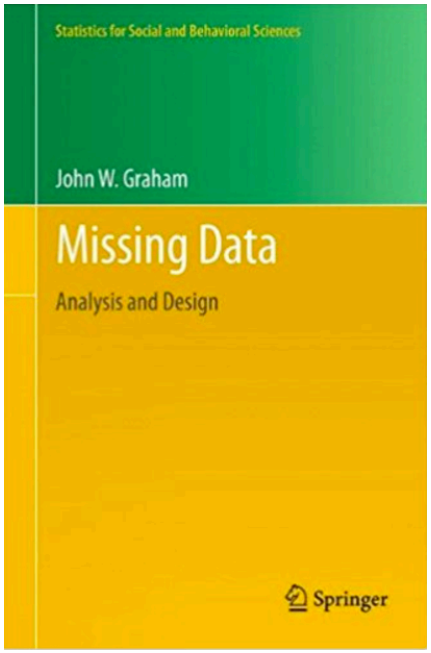


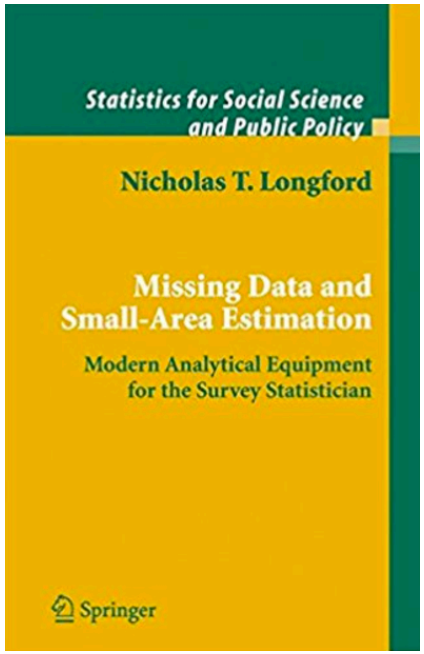




Bibliography

Covers from amazon.uk





Disclaimer: You are, of course, not expected to read all these books. The notes will be self contained, I am just showing what is available.

Bibliography

- Apart from the books shown before, several books have also a chapter on missing data. I will point out some of these as we go along the course.

- Missing data is also a popular topic for short (2/3 days) courses around the world. For some of them the material is freely available on the web. Just G o o g l e!!

- Useful resources may also be found at

  `www.missingdata.org.uk`

- There is also an excellent course taught by Professor Marie Davidian, from North Carolina State University. It is far more theoretical than what we will be doing in this course, but I leave the link here:

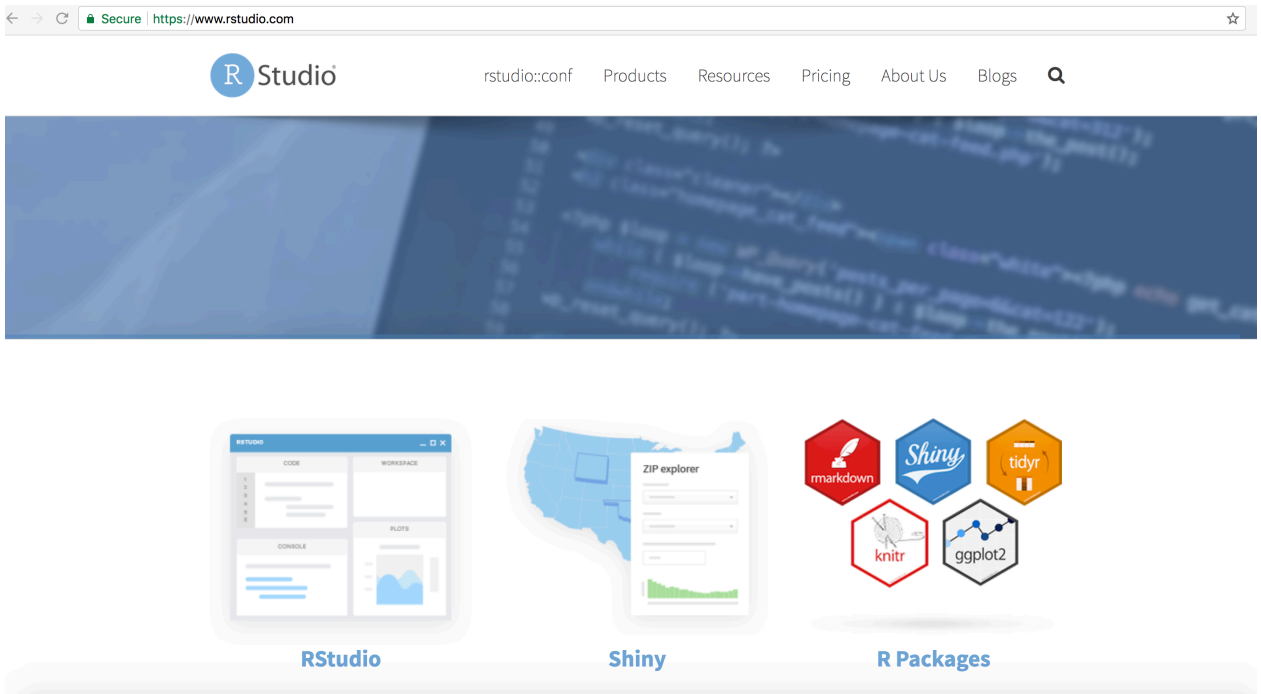
  `www4.stat.ncsu.edu/~davidian/st790/`

Software

↪ We will use `R` extensively in this course.

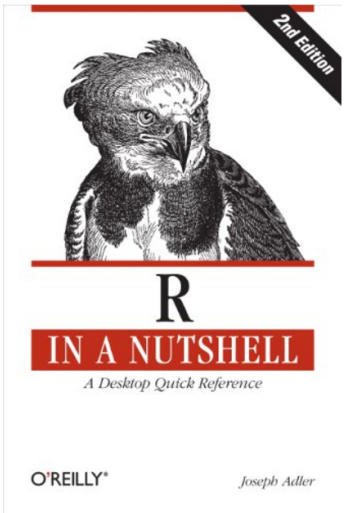
Image from `www.r-project.org`

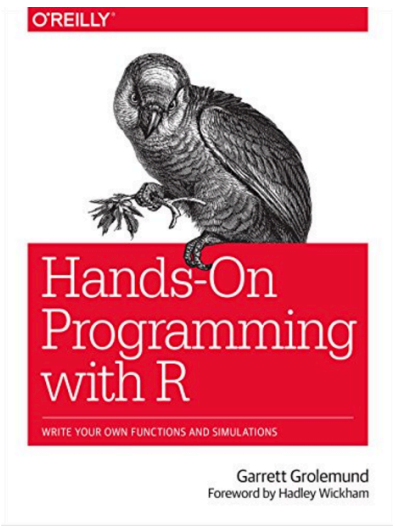
↪ `Rstudio` is a very nice interface.

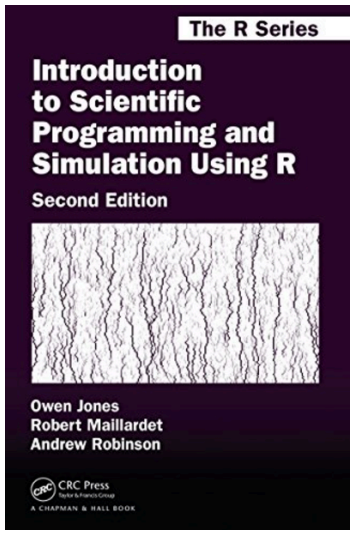


Software

- ↪ If you are not familiar with `R`, you will need to become!
- ↪ There are plenty of materials on the web. On Learn, you can find the file `Rbriefintro.pdf`, kindly provided by Professor Ruth King, which provides you with a first contact with `R`.



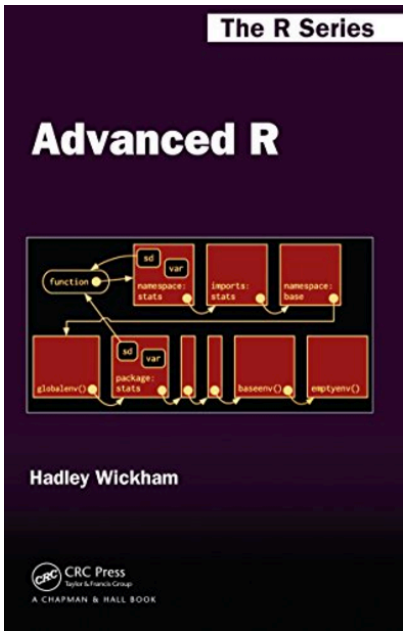


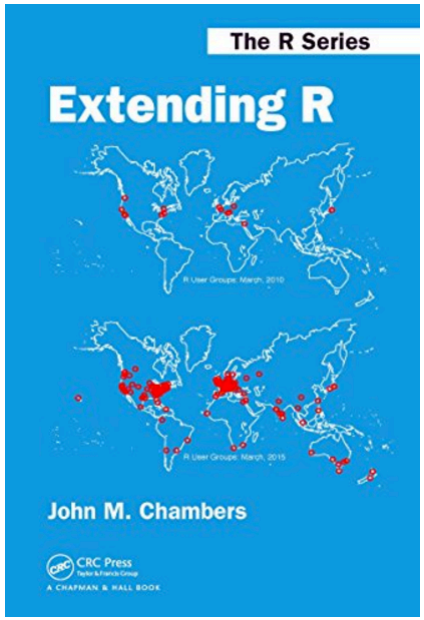


Covers from `www.amazon.co.uk`

Software

↪ For those feeling confident and wanting to learn more, further advanced references are also available.





Covers from `www.amazon.co.uk`

Scope

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

- ↪ Introduction to missing data.
- ↪ Traditional/naive methods for dealing with missing data and their drawbacks.
- ↪ Likelihood based methods for handling missing data.
- ↪ Multiple imputation.
- ↪ Further topics (if time permits).

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 models and methods for data analysis in the presence of incomplete or missing data.

- In many practical settings of interest, such as clinical trials, sample surveys, or agricultural experiments, to name only a few, data are to be collected according to a predetermined plan or design. The aim is to, given the collected data, make inference about some aspect of an underlying population of interest.

- However, due to a number of possible different reasons, data that were supposed to be collected (and thus to contribute to our inferential objective) are actually not collected or not available.

- The next three examples, adapted from Davidian's set 1 of notes, illustrate different situations in which missing data can occur.

Introduction to missing data

Non or partial response in sample surveys

- ↪ Most surveys provide good examples of missing data occurrence. For instance, suppose 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.
- ↪ 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, the response of interest (candidate preference, household income) will not be available (i.e., will be **missing**) for such individuals.

Introduction to missing data

Dropout and noncompliance in clinical trials

- ↪ A clinical trial may be conducted with the purpose of comparing the efficacy of two or more treatments in a certain population.
- ↪ Usually, the clinical procedure is as follows: individuals are recruited and enrolled in the study, are assigned (typically in a random fashion) to one of the treatments, 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 directed 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 we can imagine, this is extremely time-consuming.
- ↪ A simpler measure is to record all the food subjects ate in the last 24 hours.
- ↪ It is not hard to imagine that this 24 hour 'diary' may be correlated with the long term fat intake. Obviously, it is not a perfect measure of it, 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 practically universally accepted that all 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 require information.
- ↪ As a consequence, incomplete/missing data are a routine challenge to the data analyst in human studies.
- ↪ **Problem**: As mentioned at the beginning, interest usually lies on 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 ‘hot topic’ in modern 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 paper 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 (yesterday!!) this paper has 7867 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.

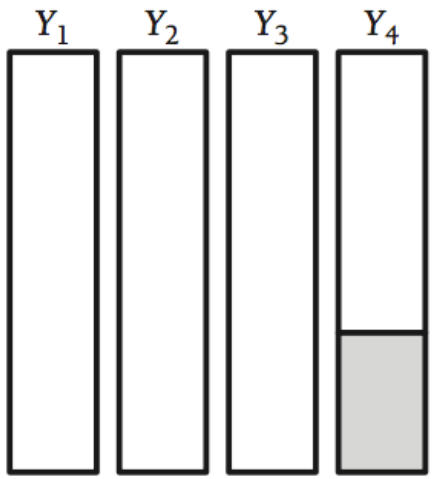
Some key words: Bayesian inference; Incomplete data; Likelihood inference; Missing at random; Missing data; Missing values; Observed at random; Sampling distribution inference.

Missing data patterns

- A pattern of missing data describes the location of the missing values in a dataset.
- The pattern describes the location of the ‘holes’ in the data but says nothing about why the data are missing.
- For simplicity, consider a rectangular data matrix with rows representing subjects and columns representing variables.
- The rows and columns in the data matrix can be sorted or rearranged to get special patterns.

Missing data patterns

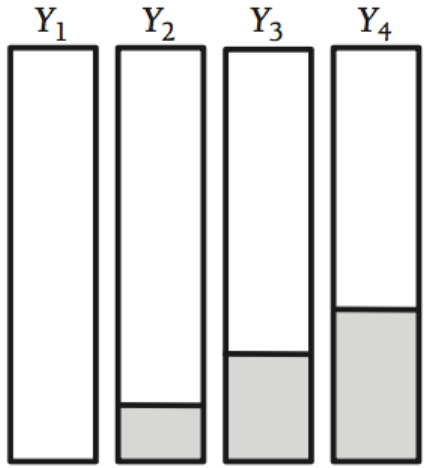
↪ In a **univariate pattern**, a common pattern, data are missing only on one variable in the analysis.



The shaded areas represent the location of the missing values in the data set. Figure from Enders, 2010, p. 4.

Missing data patterns

↪ The figure below shows a **monotone pattern** of missing data where the variable $j = 2, 3, \ldots, p$ is observed on a subset of subjects with variable $j - 1$ observed.



The shaded areas represent the location of the missing values in the data set. Figure from Enders, 2010, p. 4.

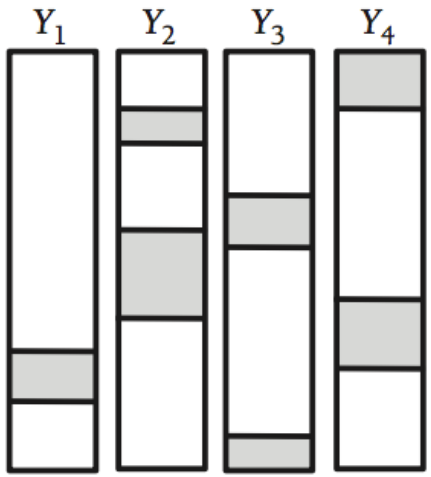
↪ A monotone missing data pattern is typically associated with a longitudinal study where participants drop out and never return.

↪ For example, consider a clinical trial for a new medication in which participants quit the study because they are having adverse reactions to the drug. Visually, the monotone pattern resembles a staircase, such that the cases with missing data on a particular assessment are always missing subsequent measurements.

Missing data patterns

- An arbitrary/general pattern in which any set of variables may be missing for any subject is shown on the figure below.

- This pattern corresponds to the most common configuration of missing data.



The shaded areas represent the location of the missing values in the data set. Figure from Enders, 2010, p. 4.

- The methods we focus on can handle general patterns.

Missing data mechanisms

- To decide how to handle missing data, it is helpful to know why the data are missing.
- The missing data mechanism can be thought of as a model that describes the probability or chance that a variable is observed or missing.
- We consider three general missing mechanisms:
  1. Missing completely at random (MCAR).
  2. Missing at random (MAR).
  3. Missing not at random (MNAR).

Missing data mechanisms

- To make the discussion more concrete, we consider the following example is from Fitzmaurice (2008).
- Let us consider the setting where it is of interest to relate an outcome variable Y_1, say blood glucose level, to another variable Y_2, say body mass index (BMI).
- Suppose that values of Y_2 are not always measured. That is, for some individuals we obtain the blood glucose level, but do not obtain their BMI. The missing data pattern can be thought of as a univariate pattern.
- Then, the missing data mechanism can be thought of a statistical model for the probability that Y_2 is missing.

Missing data mechanisms

MCAR

- In this example, there are missing data on Y_2–BMI only; however this does not always need to be the case.
- The data on Y_2 are said to be MCAR if the probability that Y_2 is missing is unrelated to the specific values of Y_2 that, in principle, should have been obtained or to the observed values of Y_1.
- Specifically, in the context of this example, MCAR implies that those subjects with missing values for BMI are no more likely to be obese or underweight or to have extreme values for blood glucose than those subjects with observed values for BMI.
- In a certain sense, with an MCAR mechanism, missingness on Y_2 can be thought of as being the result of a chance mechanism that does not depend on what was observed or on what happens to be missing.
- The essential feature of MCAR is that the observed data can be thought of as a purely random sample of the complete data (i.e., the data that would have been obtained if there were no missing data).

Missing data mechanisms

MCAR

- ↪ It must be emphasised that MCAR is a very strong assumption and should be made only in cases where there is a strong rationale for it being tenable.
- ↪ Violations of the assumption of MCAR are actually testable from the data at hand.
- ↪ For example, if the sample is stratified on the basis of missingness in Y_2, the two groups should not differ in terms of their values for Y_1.
- ↪ Concretising, if we divide the glucose levels in two groups, one for those with observed BMI and one for those whose BMI measurement is missing, the two groups should not differ. It is then possible to compare the groups for systematic differences (e.g., using a t-test).
- ↪ However, we must be aware, that there could still be some relationship between missingness on Y_2–BMI and the values of Y_2 themselves and the aforementioned procedure does not take that into account.

Missing data mechanisms

MAR

- Most missingness is not completely at random.
- A more general assumption, MAR, is that the probability of a variable is missing depends only on available/observed information.
- Specifically, in the context of the current example, MAR implies that those subjects with missing values for BMI may be more likely to have extreme values for blood glucose.
- Under the MAR assumption, the probability of missing data on BMI depends on the individual glucose level, but within glucose level group (e.g., strata with similar glucose levels) the probability of a subject having a missing BMI value is the same as for any other subject, i.e., within glucose level strata missing is MCAR.
- That is, overall BMI is MAR but within glucose level strata is MCAR.

Missing data mechanisms

MAR

- ↪ Note that, unlike MCAR mechanism, it is not possible to verify the MAR assumption from the data at hand.

- ↪ This is because we do not know the vales of missing data, thus, we cannot compare the values of those with and without missing data to see if they systematically differ on that variable.

Missing data mechanisms

MNAR

- ↪ Data are missing not at random if it depends on information that has not been recorded and this information also predicts the missing values.
- ↪ In the context of the blood glucose/BMI example, the data would be MNAR if those subjects with missing values for BMI were more likely to be obese (or underweight). That is, missing in BMI is related to unobserved obesity.
- ↪ A familiar example from medical studies is that if a particular treatment causes discomfort, a patient is more likely to drop out of the study. This missingness is not at random (unless "discomfort" is measured and observed for all patients).
- ↪ When data are suspected to be MNAR, it is important to carefully assess the sensitivity of results to a variety of plausible assumptions concerning the missingness process.

Missing data mechanisms

Exercise (from Raghunathan, 2016, p. 23)

↪ A survey is being conducted based on a random sample of firms from the population list which has the name and size of the firm (number of employees). A key survey variable of interest is whether or not the firm offers health insurance to its employees and the number of health plans offered. Consider the following missing data mechanisms:

(a) All firms exceeding some certain firm size refuse to participate.

(b) Firms that do not offer health plans and/or very limited number of plans are more likely to be nonrespondents.

State whether the plans are MCAR, MAR, or MNAR.

Missing data mechanisms

Another exercise (from last year's exam)

↪ For the situations below, decide, justifying, whether the missing data are MCAR, MAR, or MNAR.

(a) In a sample survey on income, people with higher earnings are less likely to reveal them.

(b) An investigator is studying the effects of Ghrelin, a hormone that stimulates appetite, on eating. A sample sent to the laboratory from subject y is contaminated during analysis and therefore no data are recorded.

(c) A sample survey on employment is conducted and the variables age, gender, race, and education are fully recorded. It is found that a high number of Hispanics refuse to answer questions concerning their employment status.

Summing up...what did I learn today?

- Missing values are more than the exception in data analyses.
- In order to obtain valid inferences the issue of missing data should be properly addressed.
- By now, you should be able to distinguish between **missing data pattern** (where the missing values occur in a dataset) and **missing data mechanism** (which can be thought of as a model for the probability that a variable is missing).
- For the missing data mechanisms, Rubin distinguished between
  - **MCAR**: Probability of missing values has nothing to do with what is observed or missing.
  - **MAR**: Probability of missing values depends only on the observed values.
  - **MNAR**: Probability of missing values depends on the missing values themselves or on unmeasured variables, and in addition it can depend on the observed values as well.

Summing up...what did I learn today?

↪ Quoting Fitzmaurice (2008):

> "*In closing, it should be emphasized that assumptions about missing data are inherently difficult, if not possible, to verify from the data at hand. Consequently, whenever possible, researchers should go to great lengths to minimize the amount of missing data in their studies. In general, the potential for bias is somewhat greater when the proportion of missing data is relatively large.*"