Statistical Methods for Analysis with Missing Data

Lecture 1: syllabus, motivating examples

Mauricio Sadinle

Department of Biostatistics

UNIVERSITY of WASHINGTON

Your Instructor: Mauricio Sadinle

- Research focus: methodologies for record linkage and missing data
- Assistant Professor, UW Biostat
  - Postdoc in Statistics, Duke
  - PhD in Statistics, Carnegie Mellon
  - BSc from National University of Colombia (Bogota)

The Syllabus

Let's carefully go through the syllabus, so you can decide whether to stay in the course!

Today's Lecture

- Introduction
- Syllabus
- Motivating examples (pages 1 – 13, Chapter 1, of Davidian and Tsiatis – read!)

Disclaimer

My work in this area has been more theoretical/conceptual than applied

Expect focus on:

- Big ideas
- Assumptions
- Criticism
- Technical details

The Goal of the Course

To provide a comprehensive overview of modern methods for handling missing data in statistical analyses

Incomplete Multivariate Data

Gender	Age	Income	...
F	25	60,000	...
M	?	?	...
?	51	?	...
F	?	150,300	
...	...	...	...

Why Do We Get Missing Data?

Data collection is not always ideal

- Sample surveys: a random sample of individuals are to be contacted with the intention of asking them a set of questions
  - Individuals may not answer the door/phone/email, or may respond only to certain questions
- Clinical trials: a study is conducted to compare the effectiveness of a number of treatments in a target population
  - Study participants may fail to show up to some check-ups, some may drop out of the study. Participants may also fail to report some of their baseline characteristics, recorded at the beginning of the study
- Administrative registries: data were being collected for administrative purposes, but later we realize that they can be exploited for statistical analyses
  - Certain variables might have missingness or even only be sporadically observed if their collection was not enforced

Why Do We Get Missing Data?

Data collection is not always ideal

- Sample surveys: a random sample of individuals are to be contacted with the intention of asking them a set of questions
  - Individuals may not answer the door/phone/email, or may respond only to certain questions
- Clinical trials: a study is conducted to compare the effectiveness of a number of treatments in a target population
  - Study participants may fail to show up to some check-ups, some may drop out of the study. Participants may also fail to report some of their baseline characteristics, recorded at the beginning of the study
- Administrative registries: data were being collected for administrative purposes, but later we realize that they can be exploited for statistical analyses
  - Certain variables might have missingness or even only be sporadically observed if their collection was not enforced

Why Do We Get Missing Data?

Data collection is not always ideal

- Sample surveys: a random sample of individuals are to be contacted with the intention of asking them a set of questions
  - Individuals may not answer the door/phone/email, or may respond only to certain questions
- Clinical trials: a study is conducted to compare the effectiveness of a number of treatments in a target population
  - Study participants may fail to show up to some check-ups, some may drop out of the study. Participants may also fail to report some of their baseline characteristics, recorded at the beginning of the study
- Administrative registries: data were being collected for administrative purposes, but later we realize that they can be exploited for statistical analyses
  - Certain variables might have missingness or even only be sporadically observed if their collection was not enforced

Why Do We Get Missing Data?

Missing data can also occur by design

- Two-phase epidemiologic studies: cheap measurements are collected on all study individuals, expensive measurements are collected only on a subset of individuals
- Survey sampling: we do not observe the characteristics for individuals who were not selected to be in the sample
- Split-questionnaires: to reduce respondent burden, only subsets of questions are asked to individuals

Why Do We Get Missing Data?

Missing data can also occur by design

- Two-phase epidemiologic studies: cheap measurements are collected on all study individuals, expensive measurements are collected only on a subset of individuals
- Survey sampling: we do not observe the characteristics for individuals who were not selected to be in the sample
- Split-questionnaires: to reduce respondent burden, only subsets of questions are asked to individuals

Why Do We Get Missing Data?

Missing data can also occur by design

- Two-phase epidemiologic studies: cheap measurements are collected on all study individuals, expensive measurements are collected only on a subset of individuals
- Survey sampling: we do not observe the characteristics for individuals who were not selected to be in the sample
- Split-questionnaires: to reduce respondent burden, only subsets of questions are asked to individuals

Nontraditional Missing Data

Several problems can be framed as missing data problems

- Record linkage: individuals' information may appear scattered across data sources, but no unique identifier available
  - Data: hospital data containing treatment information, mortality registry that measures survival. Missing data: "links" connecting records that refer to the same individuals

- Measurement error: we can only measure a noisy or surrogate version of what we want
  - Data: 24-hour recall, self-reported measurement of daily fat intake. Missing data: true fat intake

Nontraditional Missing Data

Several problems can be framed as missing data problems

- Record linkage: individuals' information may appear scattered across data sources, but no unique identifier available
  - Data: hospital data containing treatment information, mortality registry that measures survival. Missing data: "links" connecting records that refer to the same individuals
- Measurement error: we can only measure a noisy or surrogate version of what we want
  - Data: 24-hour recall, self-reported measurement of daily fat intake. Missing data: true fat intake

Sometimes We Make Up The "Missing Data"

Techniques for handling missing data can be useful for other problems

- Latent-variable modeling: the data might be well modeled hypothesizing the existence of a latent (fully unobserved) variable
  - Data: in-favor/opposed to a number of social/political issues. Latent variable: "political spectrum"
  - Data: friendship connections between people. Latent variable: "community membership" or "social space"
  - Data: responses to test questions. Latent variable: "ability"

- Causal inference: we only observe the outcome under the assigned treatment – what would the outcome be had the subject been assigned to another treatment?
  - One can argue that "potential outcomes" under other treatments are made up missing data as their values never existed, although they *could* have existed

Sometimes We Make Up The "Missing Data"

Techniques for handling missing data can be useful for other problems

- Latent-variable modeling: the data might be well modeled hypothesizing the existence of a latent (fully unobserved) variable
  - Data: in-favor/opposed to a number of social/political issues. Latent variable: "political spectrum"
  - Data: friendship connections between people. Latent variable: "community membership" or "social space"
  - Data: responses to test questions. Latent variable: "ability"

- Causal inference: we only observe the outcome under the assigned treatment – what would the outcome be had the subject been assigned to another treatment?
  - One can argue that "potential outcomes" under other treatments are made up missing data as their values never existed, although they *could* have existed

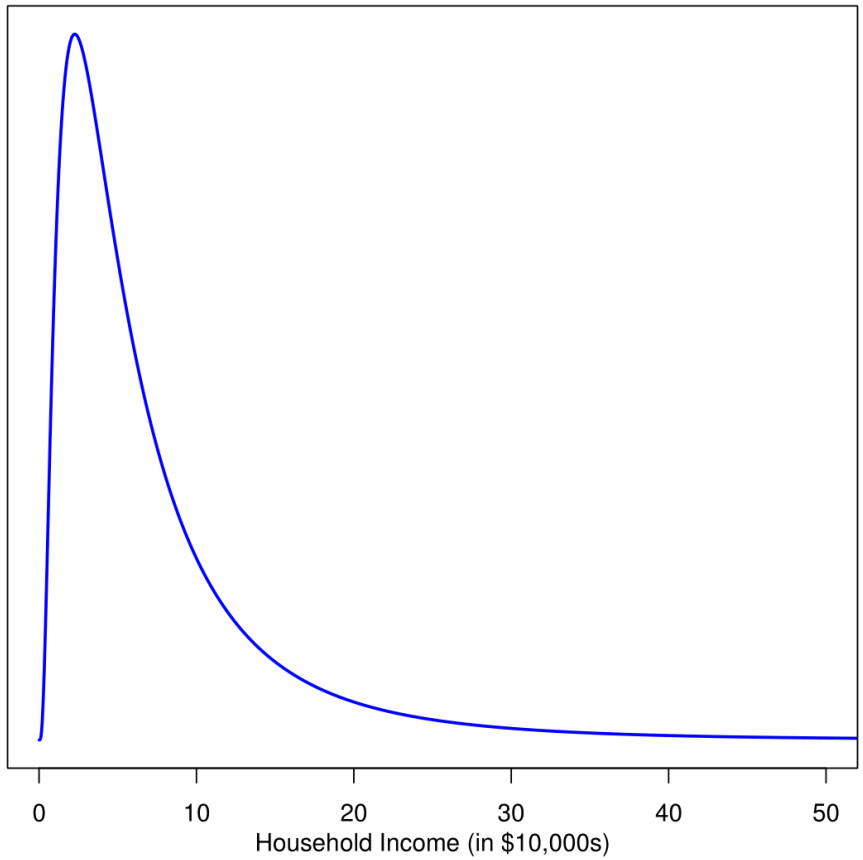
Example 1 from Davidian and Tsiatis' Chapter 1

Nonresponse in a sample survey

- A survey is conducted to study a population's income per household
- Questionnaire sent to sample of individuals randomly selected from this population
- Many participants fail to report their income

Example 1 from Davidian and Tsiatis' Chapter 1

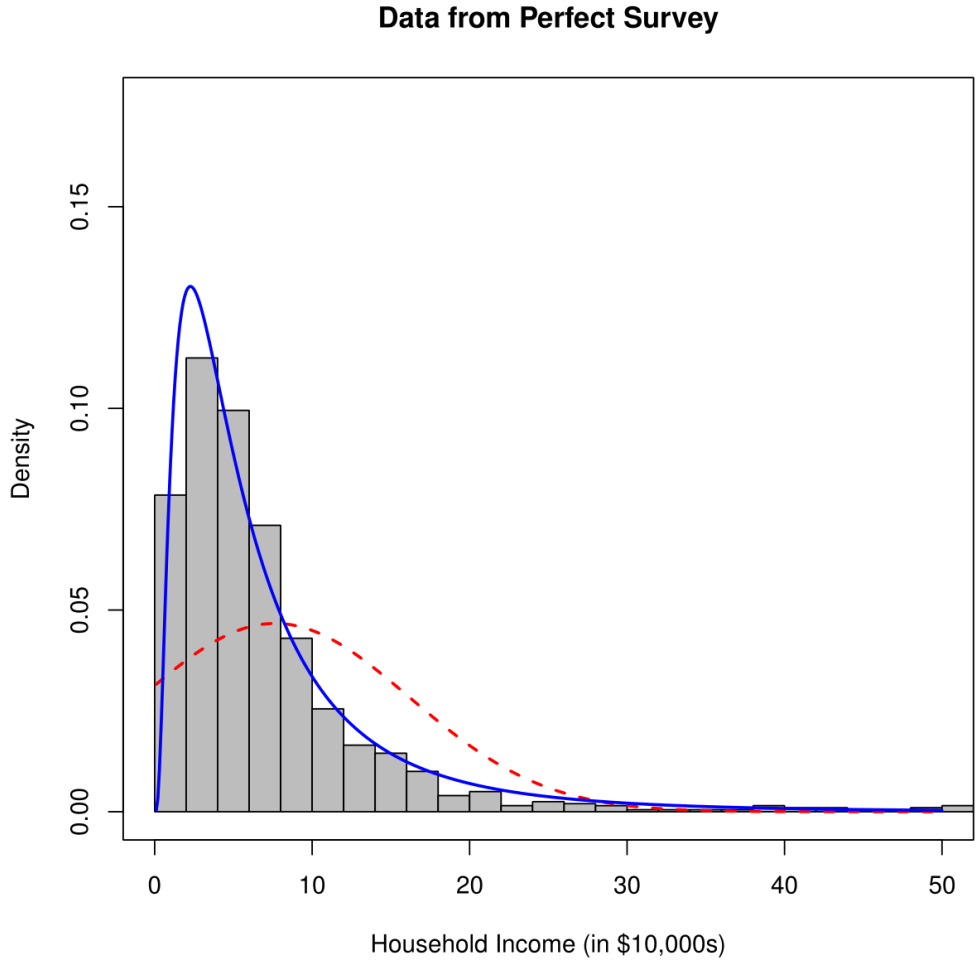
Based on data from the US Census Bureau for 2012, this could be a reasonable guess of the probability density of US household incomes, with median \$51,000 and Gini coefficient 0.48:



Of course, if we knew this distribution for sure, there wouldn't be a reason to conduct the survey!

Example 1 from Davidian and Tsiatis' Chapter 1

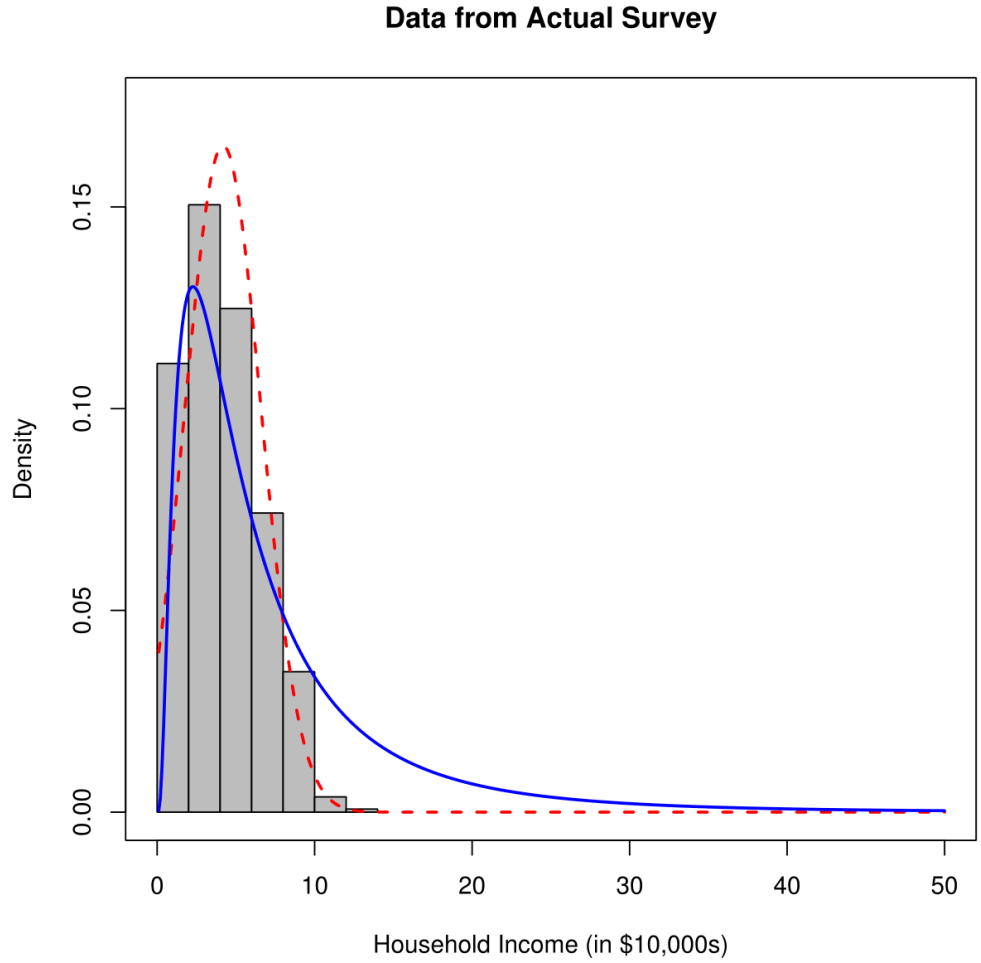
In a perfect world where all survey participants report their income, the histogram of the sample data would look like:



The red dashed line shows the best normal fit – not great

Example 1 from Davidian and Tsiatis' Chapter 1

In a more realistic survey, some participants do not report their income, and the observed data histogram might look very differently from the underlying distribution (of course, we wouldn't know this)



Red dashed line: best normal fit to the observed data – not too bad

Example 1 from Davidian and Tsiatis' Chapter 1

Based on the actual survey results, what can we conclude about the true distribution of household incomes?

First, some notation:

- Y: random variable measuring the income for a generic household
- R: response indicator, 1 if Y is observed, 0 otherwise

We might think that whether Y gets reported is a random event (perhaps it depends on the respondent's current level of paranoia?)

Example 1 from Davidian and Tsiatis' Chapter 1

- $p(y)$: density of income in the population
- $p(y \mid R = r)$: density of income given $R = r$, $r = 0, 1$

$$\underbrace{p(y)}_{\text{what we want}} = p(y \mid R = 0)p(R = 0) + \underbrace{p(y \mid R = 1)}_{\text{what we can get}}p(R = 1)$$

We cannot recover $p(y \mid R = 0)$ nor $p(y)$ from observed data alone

The fundamental problem of inference with missing data: it is impossible without extra, usually untestable, assumptions on how missingness arises

Example 1 from Davidian and Tsiatis' Chapter 1

- $p(y)$: density of income in the population
- $p(y \mid R = r)$: density of income given $R = r$, $r = 0, 1$

$$\underbrace{p(y)}_{\text{what we want}} = p(y \mid R = 0)p(R = 0) + \underbrace{p(y \mid R = 1)}_{\text{what we can get}}p(R = 1)$$

We cannot recover $p(y \mid R = 0)$ nor $p(y)$ from observed data alone

The fundamental problem of inference with missing data: it is impossible without extra, usually untestable, assumptions on how missingness arises

Example 1 from Davidian and Tsiatis' Chapter 1

What could we assume?

- No difference in the distribution of income between respondents and nonrespondents, that is,

$$p(y \mid R = 0) = p(y \mid R = 1).$$

This implies (HW1)

$$p(y \mid R = 1) = p(y),$$

and therefore we can recover $p(y)$ from the distribution of income among respondents

Note: assuming

$$p(y \mid R = 0) = p(y \mid R = 1)$$

is equivalent to assuming

$$p(R = r \mid y) = p(R = r),$$

for all y, $r = 0, 1$ (HW1)

Example 1 from Davidian and Tsiatis' Chapter 1

What could we assume?

- No difference in the distribution of income between respondents and nonrespondents, that is,

$$p(y \mid R = 0) = p(y \mid R = 1).$$

This implies (HW1)

$$p(y \mid R = 1) = p(y),$$

and therefore we can recover $p(y)$ from the distribution of income among respondents

Note: assuming

$$p(y \mid R = 0) = p(y \mid R = 1)$$

is equivalent to assuming

$$p(R = r \mid y) = p(R = r),$$

for all y, $r = 0, 1$ (HW1)

Example 1 from Davidian and Tsiatis' Chapter 1

What could we assume?

- The probability of nonresponse increases with income, that is, $p(R = 0 \mid y)$ increases as a function of y.

If this is the case, we need to specify a functional form for $p(R = 0 \mid y)$ based on *external knowledge* – $p(R = 0 \mid y)$ cannot be estimated from the observed data (HW1)

In this case

$$p(y \mid R = 1) \neq p(y),$$

and therefore we *cannot* recover $p(y)$ directly from $p(y \mid R = 1)$

Example 1 from Davidian and Tsiatis' Chapter 1

The moral of the story

- Different missing-data assumptions will lead to different inferences
- Missing-data assumptions are unverifiable from the observed data

Example 2 from Davidian and Tsiatis' Chapter 1

Randomized clinical trial

- Two treatments coded as $A = 0, 1$
- Participants are randomized at time t_1
- Participants are to be followed up at times $t_2, \ldots, t_T$
- Y_j: outcome of interest at time t_j
- We might be interested in the trajectories $E(Y_j \mid A = a)$, $a = 0, 1$, as a function of time
- Depending on the context, it could be reasonable to hypothesize a linear relationship

$$E(Y_j \mid A = a) = \beta_{0,a} + \beta_{1,a} t_j$$

- Without missing data, estimating $\beta_{0,a}$ and $\beta_{1,a}$ can be done using "standard" techniques

Example 2 from Davidian and Tsiatis' Chapter 1

Randomized clinical trial

- Two treatments coded as $A = 0, 1$
- Participants are randomized at time t_1
- Participants are to be followed up at times $t_2, \ldots, t_T$
- Y_j: outcome of interest at time t_j
- We might be interested in the trajectories $E(Y_j \mid A = a)$, $a = 0, 1$, as a function of time
- Depending on the context, it could be reasonable to hypothesize a linear relationship

$$E(Y_j \mid A = a) = \beta_{0,a} + \beta_{1,a} t_j$$

- Without missing data, estimating $\beta_{0,a}$ and $\beta_{1,a}$ can be done using "standard" techniques

Example 2 from Davidian and Tsiatis' Chapter 1

Randomized clinical trial

- Two treatments coded as $A = 0, 1$
- Participants are randomized at time t_1
- Participants are to be followed up at times $t_2, \ldots, t_T$
- Y_j: outcome of interest at time t_j
- We might be interested in the trajectories $E(Y_j \mid A = a)$, $a = 0, 1$, as a function of time
- Depending on the context, it could be reasonable to hypothesize a linear relationship

$$E(Y_j \mid A = a) = \beta_{0,a} + \beta_{1,a} t_j$$

- Without missing data, estimating $\beta_{0,a}$ and $\beta_{1,a}$ can be done using "standard" techniques

Example 2 from Davidian and Tsiatis' Chapter 1

Randomized clinical trial

- Two treatments coded as $A = 0, 1$
- Participants are randomized at time t_1
- Participants are to be followed up at times $t_2, \ldots, t_T$
- Y_j: outcome of interest at time t_j
- We might be interested in the trajectories $E(Y_j \mid A = a)$, $a = 0, 1$, as a function of time
- Depending on the context, it could be reasonable to hypothesize a linear relationship

$$E(Y_j \mid A = a) = \beta_{0,a} + \beta_{1,a} t_j$$

- Without missing data, estimating $\beta_{0,a}$ and $\beta_{1,a}$ can be done using "standard" techniques

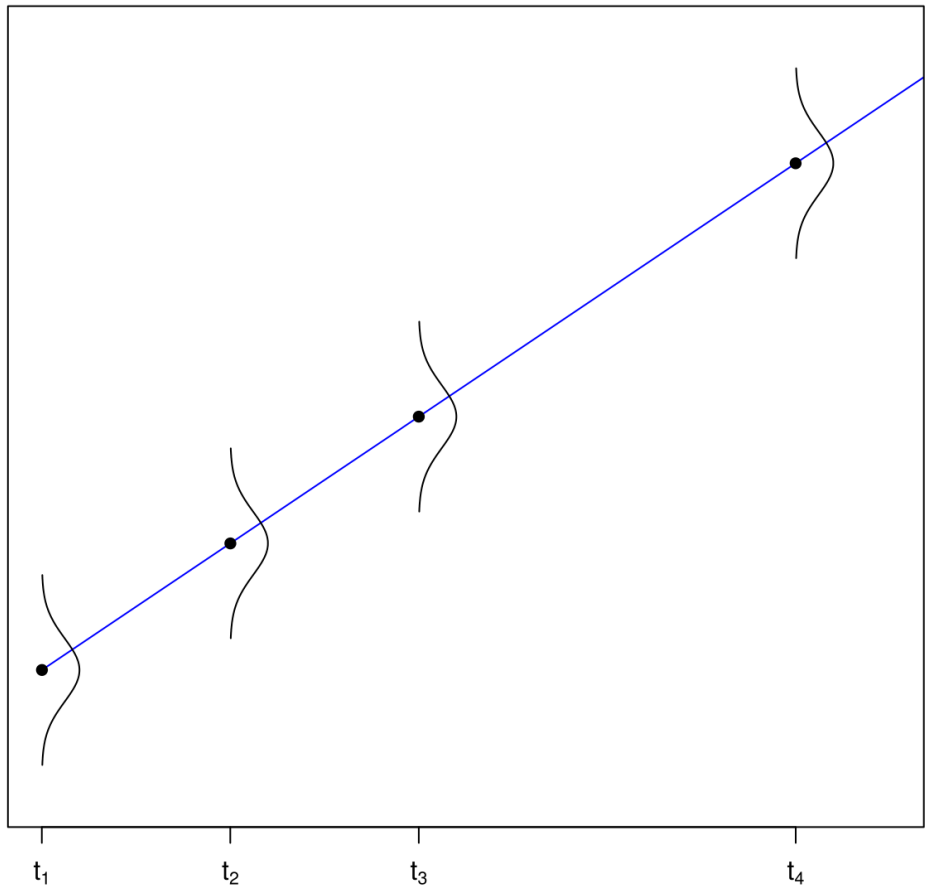
Example 2 from Davidian and Tsiatis' Chapter 1

In such studies it is common for subjects to *drop out*

- A patient drops out at time t_j if they appear up until time t_{j-1}
- This means that we observe $Y_1, \ldots, Y_{j-1}$ but do not observe $Y_j, \ldots, Y_T$

Example 2 from Davidian and Tsiatis' Chapter 1

Suppose that the true distribution of $Y_j \mid A = 1$, for $j = 1, \ldots, 4$ is given by:



Can we recover this when the study is subject to dropout?: we need to make assumptions!

Example 2 from Davidian and Tsiatis' Chapter 1

What could we assume?

- Dropout is unrelated to outcome: the fact that participants drop out has nothing to do with the variables being measured

If this is the case, at each time, the distribution of the outcomes is the same for those who drop out and for those who stay – we can estimate each $E(Y_j \mid A = a)$ from the observed data

Example 2 from Davidian and Tsiatis' Chapter 1

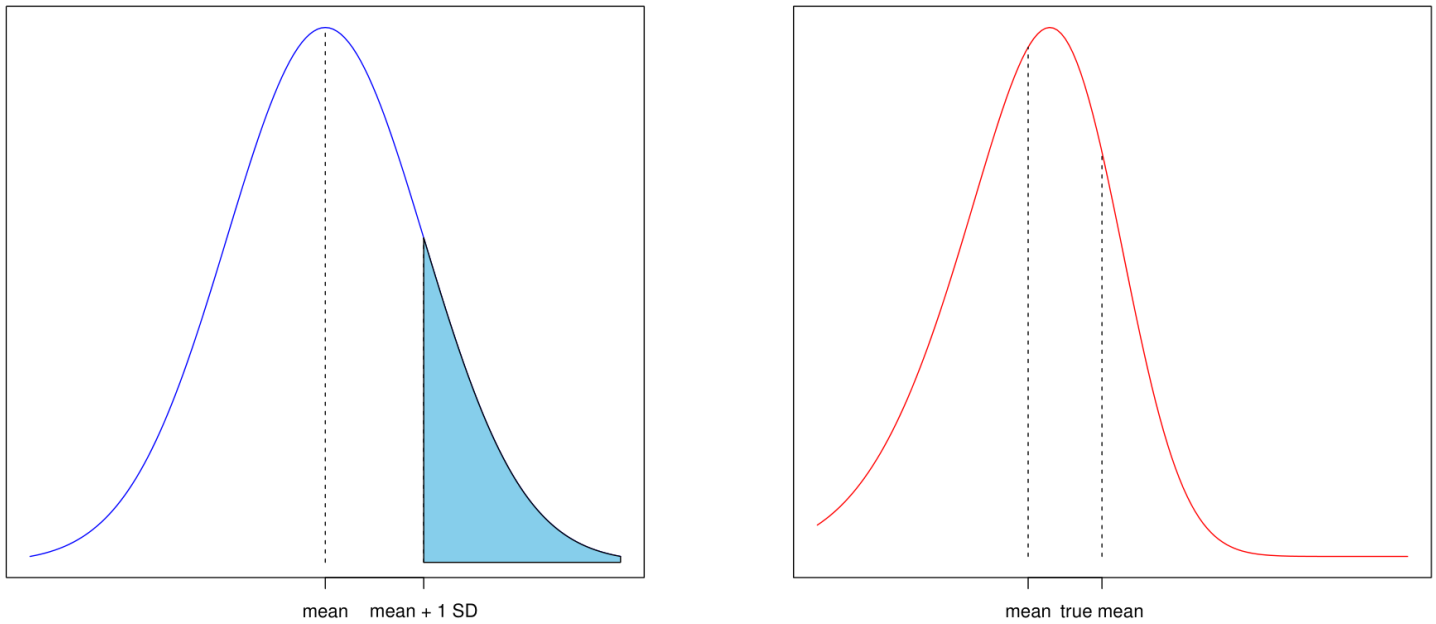
What could we assume?

- Dropout is related only to information that is observed: the fact that participants drop out can be explained by characteristics that have been measured

For example, the probability of dropping out at time t_j could increase as a function of Y_{j-1}

Example 2 from Davidian and Tsiatis' Chapter 1

Left: observed distribution of Y_1. Right: observed distribution of Y_2.

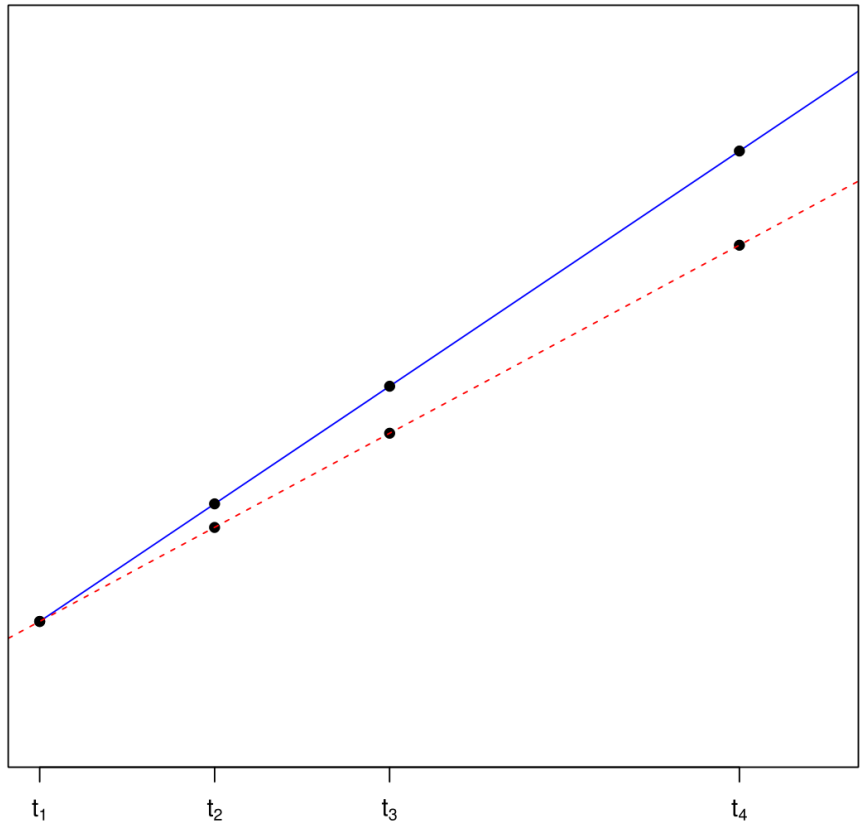


This is assuming that the probability of not observing Y_2 increases with the value of Y_1, and that Y_1 and Y_2 are positively correlated.

The mean of Y_2 among those who do not drop out is biased downwards.

Example 2 from Davidian and Tsiatis' Chapter 1

If the probability of dropping out at time t_j increases as a function of Y_{j-1}, we'd obtain something like:



The observed mean outcome at each time (red) will be smaller compared to the actual mean (blue). As we'll see later, this scenario is not too bad, as the "reasons" for dropping out are observed – corrections can be made.

Example 2 from Davidian and Tsiatis' Chapter 1

What could we assume?

- Dropout is related to information that is *not* observed: the fact that participants drop out is explained by the variables that we want to measure

For example, the probability of dropping out at time t_j could increase as a function of Y_j – in this case the observed mean outcome at each time will also be smaller compared to the actual mean. However, this scenario seems hopeless – the "reasons" for dropping out are not observed.

Again, the functional form for the probability of dropping out would have to be specified based on *external knowledge*.

Summary

Main take-aways from today's lecture:

- Inference with missing data requires making unverifiable assumptions
- Different missing-data assumptions will typically lead to different inferences

The fundamental problem of inference with missing data: it is impossible without extra, usually untestable, assumptions on how missingness arises

Next lecture:

- Notation
- General setup
- Missing-data mechanisms
- Contents based on pages 14 – 22, Chapter 1, of Davidian and Tsiatis (read!)

Summary

Main take-aways from today's lecture:

- Inference with missing data requires making unverifiable assumptions
- Different missing-data assumptions will typically lead to different inferences

The fundamental problem of inference with missing data: it is impossible without extra, usually untestable, assumptions on how missingness arises

Next lecture:

- Notation
- General setup
- Missing-data mechanisms
- Contents based on pages 14 – 22, Chapter 1, of Davidian and Tsiatis (read!)

Summary

Main take-aways from today's lecture:

- Inference with missing data requires making unverifiable assumptions
- Different missing-data assumptions will typically lead to different inferences

The fundamental problem of inference with missing data: it is impossible without extra, usually untestable, assumptions on how missingness arises

Next lecture:

- Notation
- General setup
- Missing-data mechanisms
- Contents based on pages 14 – 22, Chapter 1, of Davidian and Tsiatis (read!)