Predicting Missing Data

Course developed by
Deborah H. Glueck and Keith E. Muller

Slides developed by Jessica R. Shaw, Keith E. Muller, Albert D. Ritzhaupt and Deborah H. Glueck

 1

Learning objectives

Define missing data.

Describe the types of missing data.

Describe the sources for predicting missing data in a design.

2

INTRODUCTION TO MISSING DATA

3

Missing data occurs when one or more outcome measurements fail to be recorded

Multilevel and longitudinal studies frequently produce missing data.

4

Missing data can occur for many reasons

Examples:

- Inconsistent participation
- Study drop-out
- Machine failure
- Data entry errors

5

Missing data complicates estimation and inference in three key ways

1. May bias estimates
2. May affect hypothesis test accuracy
3. Typically reduces power

6

Missing completely at random

Simplest Case: The chance of any observation being missing is independent of both observable variables and unobservable values.

7

Most power analyses assume that data is missing at random

For data missing at random, the probability of an observation being missing is not influenced by unobserved data.

A weaker assumption than completely at random.

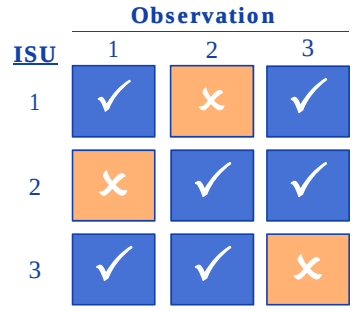
Example: In a study comparing two cancer treatments, whether survival time is known or not depends only on observed values.

8

TYPES OF MISSING DATA

9

We usually assume data is missing at random

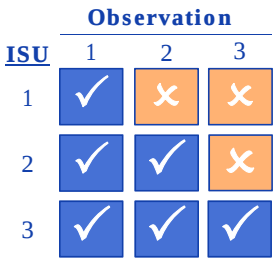


10

Sometimes missing data may be a result of dropout

Example:

Missing data may be a result of dropout if all other observations after the first missing data point are also unobserved.

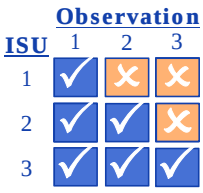


11

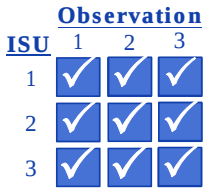
Data may be missing at different rates for different levels of a predictor

Example: Missing data may be a result of treatment-related dropout

Treatment group



Control group



12

Investigators must evaluate missing data for patterns of loss

Person	Observation 1	2	3
1	✓	✗	✗
2	✓	✓	✗
3	✓	✓	✓

Present: **2/3**

Missing: **1/3**

13

In particular, be aware of differentially missing data, wherein one group or time displays higher dropout

Treatment group

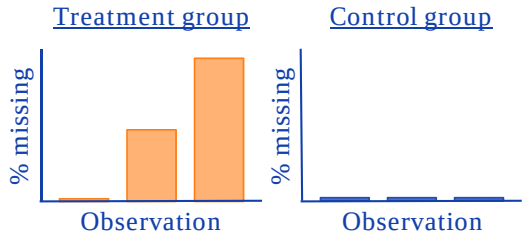
ISU	Observation 1	2	3
1	✓	✗	✗
2	✓	✓	✗
3	✓	✓	✓

Control group

ISU	Observation 1	2	3
1	✓	✓	✓
2	✓	✓	✓
3	✓	✓	✓

14

Treated participants may drop out at a higher rate if the treatment is time consuming or has adverse side effects



15

IMPLICATIONS FOR STUDY PLANNING

16

The reasons for missing data may affect power analysis

Differential dropout can affect design choices.

If more women miss visits, one may need to recruit more women. Does this bias results?

If some treatment group members develop treatment related toxicity and leave the study, the effect has to be taken into account.

17

Missing data percentage and pattern can be predicted through several methods

Methods include:

Literature review

Internal pilot studies

Planned pilot studies

Published or unpublished studies from your laboratory

Clinical knowledge of study population

18

As an example, consider the excerpt below from a published study

Out of a total of 68 patients, 60 (88.2%) completed the study treatment without serious adverse events. Treatment in two (2.9%) patients was discontinued due to elevated AST or ALT levels to more than three times the upper limit of normal, and noncompliance or loss to follow-up in six (8.8%) patients. Of the 60 patients who completed the study treatment, mean fasting plasma glucose, A1C, fasting plasma insulin, mean ALT and homeostasis model assessment for insulin resistance were all significantly reduced. Normal AST and ALT levels were achieved and maintained for at least three consecutive measurements and through to the end of the study period in 20 (33.3%) Weight increased by a mean of 2.6 +/-2.4 kg ($p < 0.001$).

19

(continued) As an example, consider the excerpt below from a published study

Treatment in two (2.9%) patients was discontinued due to elevated AST or ALT levels to more than three times the upper limit of normal, and noncompliance or loss to follow-up in six (8.8%) patients.

20

The example excerpt helps plan a new study

Based on the study, an investigator could predict the percentage missing due to adverse side effects, non-compliance, and loss to follow-up.

21

The key statement reports drop-out, noncompliance, and loss to follow-up

RESULTS: Treatment in two (2.9%) patients discontinued due to elevated AST or ALT levels to more than three times the upper limit of normal, and noncompliance or loss of follow-up in six (8.8%) patients.

22

Many data analysis approaches do not allow any missing data

Recall that the general linear multivariate model does not allow any missing outcomes.

23

Researchers should design studies to minimize risk of missing data

Best practices include using reliable technology, establishing data validation processes, minimizing the time and discomfort required of participants, refining retention protocols in pilot studies.

24

One approach for design is to first calculate sample size assuming no missing data and then adjust

After initial sample size calculation, design adjustments can be made to account for the predicted amount and pattern of missing data.

25

Review Summary

- Missing data occurs when one or more outcome measurements fail to be recorded
- Be aware of differentially missing data, wherein one group or time displays higher dropout
- Missing data percentage and pattern can be predicted through several methods like literature reviews or pilot studies

26