Missing Data Analysis

Habtamu Bizuayehu

Session Objectives

- ❖ Types of missing data
- ❖ Benefits of missing data imputation
- ❖ Techniques of imputation

➢ **Session delivery**:

Presentation, discussion, and software practice

Causes Of Missing Data

- **Not responding:** One or many variables
- **Drops out:** death, relocation, movement
- **Data entry errors**
- Questionnaire damaged

Missing Data Mechanism

1. Missing Completely at Random (MCAR): The missing values have no correlation with other values in the dataset observed or missing.

eg (like coin flip)

2. Missing at Random (MAR)

- Missingness may be related to measured variables.
- But no residual relationship with **unmeasured** variables
- No bias estimate/model if you control for measured variables

eg: age at menarche with country of birth,

Missing Data Mechanism …

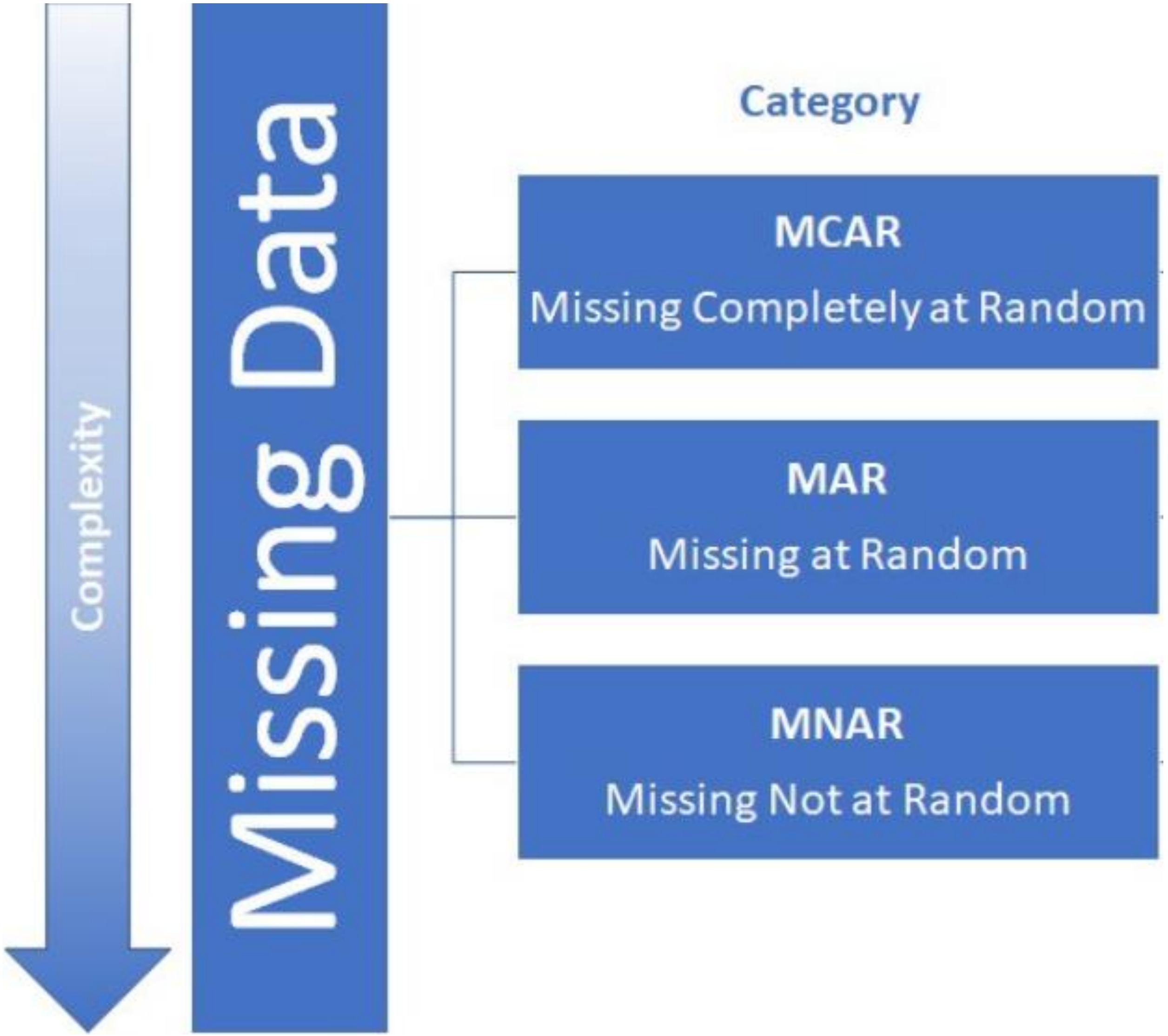
3. Missing Not at Random (MNAR):

- Probability of missing varies by reasons that are unknown/unmeasured.
- Incidentally not responding for some reason
- Residual relationship with unmeasured variables
- Bias estimation/model after imputation or controlling for measured variables

eg: drug use reason for absence, rich participants not reporting income,

weighing scale wear out over time and producing more missing data, especially for **heavier objects,** IQ missing with only the people with low IQ

Missing Data Mechanism …



- P (missing | compete data) = P (Missing)
- P (missing | compete data) =
  P (Missing | observed data)
- P (missing | compete data) ≠
  P (Missing | observed data)

Identifying Missing Data Mechanism

1. Common sense

2. Statistical

- **Univariate *t*-Test Comparisons**
- **Little's MCAR Test**

Example for *t*-Test

- Job performance scores (missing and complete cases)
- Psychological well-being score (missing and complete cases): 9.13 and 11.44, respectively
- *t* test for mean difference, p=0.19

- IQ score(missing and complete cases): 88.50 and 111.50, respectively
- *t* test for mean difference, $p < .001$

Missing Data Pattern Identification

- **One variable:** Sorting, Listing, Frequency
  - **eg** BMI (STATA code, ta BMI, m)

- **Many variables:** misschk BMI Exercise HTN smoke Edu Phi ARIA+, gen(miss)

# Variable	Freq.	%
BMI	668	11.8
Exercise	683	12.1
HTN	102	1.8
Smoke	51	0.9
Edu	76	1.3
Phi	30	0.5
ARIA+	92	1.6

#Factors	Freq.	%	Cum.
0	3,929	69.53	69.53
1	1,464	25.91	95.43
2	229	4.05	99.49
3	30	0.51	100.0
Total	5,651	100.0	100.0

Missing Data Management

- **Importance:** Missing ≥ 5%-10% (rule of thumb)

(Dong et al, 2013, Bennett et al 2001)

- Complete case analysis
- Last Observation Carried Forward (LOCF) & Next Observation Carried Backward (NOCB)
- Imputation
- Sensitivity Analysis

Imputation

Good estimate of variability

Standard errors

Fooled by Randomness

Best statistical power

Predictive Accuracy

b-weight coefficients

Preserves Structure of Data

Keep important data patterns

Analysis possible in many software

Techniques of imputation

- **Logic: Last Observation Carried Forward (LOCF) & Next Observation Carried Backward (NOCB)**
  - **Survey:** If responded other confirmatory

ID	HTN Dx	Antihypertensive Rx	Ever smoking	# Cigarettes per pay
105	. Yes	Yes	. Yes	5

  - **Cohort and trials:**

Other similar response, enduring conditions or history

eg HTN, DM, ever violated, ever smoking

ID=100	Survey 2000	Survey 2003	Survey 2006	Survey 2009
HTN	Yes	Yes	. Yes	. Yes
Ever violated	Yes	. Yes	Yes	. Yes

Imputation

- **Mean imputation**
  - **Survey**

ID	SBp	dsBp
50	123	85
51	108	78
52	86	. (75)
53	. (97)	. (75)
54…	119	96
100	96	100
Mean	97	75

No
much
supp
ort

  - **Cohort and trials**

ID	Var.	2000	2003	2006	2009	Mean
100	SBp	122	98	. (108)	105	108
100	dsBp	73	.(79)	68	96	79
101	SBp	.(110)	98	.(110)	123	110
101	dsBp	88	.(95)	98	101	95

Highly
recomm
ended

Multiple Imputation by Chained Equation (MCIM)

Mostly recommended methods White et al, 2011

Imputation number: 20 (mostly used) (options 5-60)

- **General steps**
- Develop model that associated with missingness (both univariate and multivariate analysis)

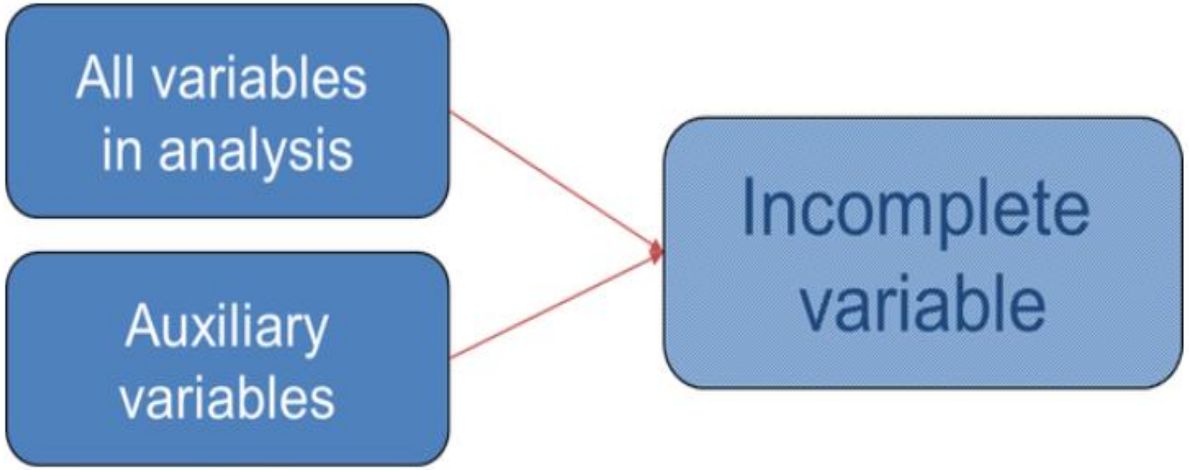
recode afmpg 0/1=1 .=0, gen (afmpgm) ; label define afmpgm 0"miss" 1"not miss"

mi set wide

mi register imputed afmpg

mi impute chained (logit) **afmpg**

= *ageg3* ***cob*** *lbw bo3 dmg htng*, add(20) force



Content of an imputation model with auxiliary variables

Multiple Imputation by Chained Equation (MCIM)

Diagnosis Test

- If observed, imputed and completed data are comparable // good imputation

- **midiagplots** afmpg // show proportions
- **midiagplots** afmpg, tabfreq // show frequencies

age at first menustrual period grouped	Observed	Imputed	Completed
0. >=12 years	0.880	0.887	0.880
1. <12 years	0.120	0.113	0.120

Multiple Imputation by Chained Equation (MCIM)

- Estimation of **descriptive** results: use **M1 imputations**

- Final **Regression** analysis by including imputed data
  - ***mi estimate***: *xtlogit lbw ib1.bo3 i.dmg i.htng i.afmpg i.bmirmg4, re*

	_1_afmpg	_2_afmpg	_19_afmpg	_20_afmpg
5	0. >=12 years	0. >=12 years	0. >=12 years	0. >=12 years
6	0. >=12 years	0. >=12 years	0. >=12 years	1. <12 years
7	1. <12 years	0. >=12 years	0. >=12 years	0. >=12 years
8	0. >=12 years	0. >=12 years	0. >=12 years	0. >=12 years
9	0. >=12 years	0. >=12 years	0. >=12 years	0. >=12 years

Sensitivity analysis

- Including and excluding the factor with high missing value in the model
- Would be option for:
- For some analysis types ***mi estimate****: may not work*

eg Path analysis

- Missing Not at Random (MNAR): tried under various scenarios but good to try to identify the causes for the missingness first

Summary

- Causes, mechanisms, and handling of missing data

- There is no a one way of managing the missing data
- Reading the literature to specific data type

- **Demonstration:** software: STATA
- use "C:\Users\c3271807\OneDrive - The University Of Newcastle\data\Analysis @ LBW\all\MI LBW singleton Hx Only.dta
- C:\Users\c3271807\OneDrive - The University Of Newcastle\data\Analysis @ LBW\stata code\all birth @ LBW SH mothers Mod 6.do

References

A. Book and Articles

1. Dong Y, Peng C-YJ. Principled missing data methods for researchers. SPRINGERPLUS. 2013;2(1):222.
2. Bennett DA. How can I deal with missing data in my study? AUSTRALIAN AND NEW ZEALAND JOURNAL OF PUBLIC HEALTH. 2001;25(5):464-9.
3. StataCorp L. Stata statistical software: Release 13.(2013). COLLEGE STATION, TX: STATACORP LP. 2013.
4. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. STAT MED. 2011;30(4):377-99.

https://www.sagepub.com/sites/default/files/upm-binaries/45664_6.pdf

B. Useful websites

https://missingdata.org/

https://www.missingdata.nl/missing-data/missing-data-methods/

C. Video summaries

https://www.youtube.com/watch?v=sUAMiAIUhcI


QUESTIONS?
COMMENTS?
CONCERNS?