Missing data: conventions and some recent ideas

at JHU-CFAR BEM Core Symposium

by Trang Quynh Nguyen (trang.nguyen@jhu.edu)
Department of Mental Health, JHSPH

2024-02-15

Missing data is ubiquitous!

- non-response/attrition
- data entry errors, lost survey forms
- individuals not wanting to disclose/not knowing certain info

How do we deal with it?

This talk is mostly conceptual.

- some long-standing ideas and conventional wisdom
- some recent ideas and new clarity

Missing data is ubiquitous!

- non-response/attrition
- data entry errors, lost survey forms
- individuals not wanting to disclose/not knowing certain info

How do we deal with it?

This talk is mostly conceptual.

- some long-standing ideas and conventional wisdom
- some recent ideas and new clarity

Acknowledgements

- Elizabeth Stuart
- NIMH grant R03MH128634

Rubin's (1976) taxonomy of missingness

one of the most influential ideas

3 types:

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing not at random (MNAR or NMAR)

Missing completely at random (MCAR)

Missingness is totally random, does not depend on anything.

More precisely, missingness is independent on our analysis variables.

Missing completely at random (MCAR)

Missingness is totally random, does not depend on anything.

More precisely, missingness is independent on our analysis variables.

Say, our analysis involves three variables V_1, V_2, V_3 with some missingness.
R_1, R_2, R_3 indicate whether these are observed ($R_j = 1$) or missing ($R_j = 0$).

MCAR means: $R_j \perp\!\!\!\perp (V_1, V_2, V_3)$ for $j = 1, 2, 3$

Missing completely at random (MCAR)

Missingness is totally random, does not depend on anything.

More precisely, missingness is independent on our analysis variables.

Say, our analysis involves three variables V_1, V_2, V_3 with some missingness.
R_1, R_2, R_3 indicate whether these are observed ($R_j = 1$) or missing ($R_j = 0$).

MCAR means: $R_j \perp\!\!\!\perp (V_1, V_2, V_3)$ for $j = 1, 2, 3$

- Cases with missing values are a random sample of the original sample; no systematic differences b/w those with missing and observed values
- Generally unrealistic, although might be reasonable for things like data entry errors or lost questionnaires
- Testable: can rule out but can't confirm

Missing at random (MAR)

Missingness depends on data that are observed.

Conventional presentation:

data D partitioned into $(D_{\text{ob}}, D_{\text{mis}})$; R being missing indicator

MAR says: $R \perp\!\!\!\perp D_{\text{mis}} \mid D_{\text{obs}}$

Missing at random (MAR)

Missingness depends on data that are observed.

I find it easier to think about it this way:

Suppose V_2, V_3 are partially missing, and V_1 is fully observed

MAR says: $R_j \perp\!\!\!\perp (V_2, V_3) \mid V_1$ for $j = 2, 3$

Missing at random (MAR)

Missingness depends on data that are observed.

I find it easier to think about it this way:

Suppose V_2, V_3 are partially missing, and V_1 is fully observed

MAR says: $R_j \perp\!\!\!\perp (V_2, V_3) \mid V_1$ for $j = 2, 3$

- Missingness doesn't depend on variables with missing data.
- Missingness in a variable doesn't depend on the value of that variable.
- Conditional on the fully observed variables, we again have exchangeability.

- Probably the assumption made most frequently
- Conventionally thought of as untestable, but there is some testability (Potthoff et al. 2006)

Missing not at random (MNAR)

Missingness depends on unobserved data.

Basically, cases not MCAR or MAR are MNAR.

Examples

- People with high income more likely not to disclose their income
- Probability of reporting medication adherence depends on whether the person adheres
- Probability of responding to a survey about treatment interruption depends on having a working cell phone, which also affects treatment interruption/continuity

This means

- Even among people with same values on the fully observed variables, the missing values on the partially observed variable have a different distribution than the observed variables, ie no exchangeability!

Not testable, ie can't rule it out

Conventional wisdom

- Under MCAR, complete case analysis is valid
- MCAR is rare, so it's safer to assume MAR
- Under MAR, complete case analysis is biased, should instead use MAR-based methods
  - maximum likelihood (FIML, EM) or Bayesian analysis – complicated
  - multiple imputation (MI) – general method and popular
  - inverse probability of response weighting – most often used for outcome missingness
- Use auxiliary variables to get close to MAR (this came up in MI context)
- If worry about MNAR, should do sensitivity analysis (but this is hard)

Inverse probability weighting

Well established method under assumed exchangeability, ie MAR

- IPW estimation of average treatment effect – under unconfoundedness
- survey weighting – with known sampling probabilities
- weighting to correct for survey non-response
- inverse probability of censoring weighting – assuming ignorable censoring
- etc.

Not suited for the usual situation of missingness in many covariates

Modeling methods

For modeling-centric methods (ML, Bayesian and also MI), there have been a lot of development with different types of models for different types of data, eg

- dependent data eg multi-level
- categorical variables
- items on a scale
- a large number of auxiliary variables
- etc.

see Enders (2023) for an update.

Multiple imputation (Rubin 1987)

Intuitive idea: conditional on observed variables, missing values have the same distribution as observed values,
so can impute plausible values from that distribution

(other methods can be thought of as implicit imputation)

For how to do MI, there are many resources. Not our focus today.

Multiple imputation (Rubin 1987)

Intuitive idea: conditional on observed variables, missing values have the same distribution as observed values,
so can impute plausible values from that distribution

Popular approach (now sort of a default), perhaps because

- handles missingness in multiple variables
- can use standard analysis method
- lots of MI software available

Criticisms of MI (or the practice of MI)

Criticism – statistical

- sensitivity to imputation model (misspecification, model incompatibility)
- variance and variance estimation

Criticism – causal

- limited consideration (and transparency) of missing data mechanisms in practice – researchers often default to MAR
- lack of consideration of what specifically is being estimated
- the usual guidance on auxiliary variables (predictive of missingness and missing value) may be simplistic

A quick visit with the causal criticisms

(Daniel et al. 2011, Thoemmes & Mohan 2015, Carpenter & Smuk 2020, Mohan & Pearl 2021)

Common estimands:

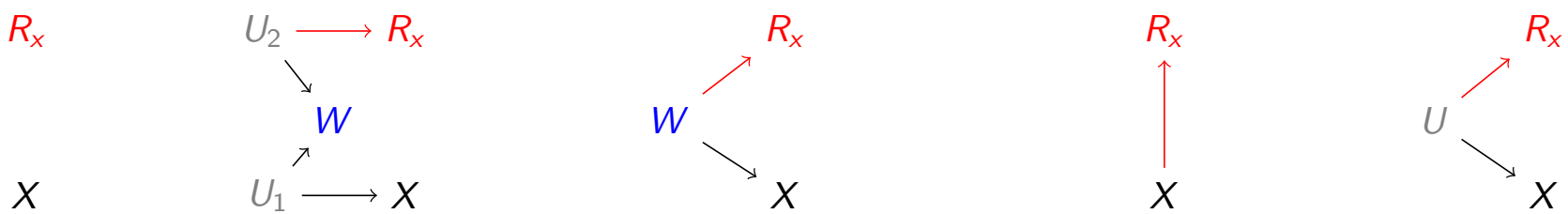
1. distribution or mean of a variable: $\mathrm{P}(Y)$ or $\mathrm{E}[Y]$
2. conditional distribution $\mathrm{P}(Y \mid X)$ or $\mathrm{P}(Y \mid A, X)$
   (where A, Y, X are exposure, outcome and covariates)
   or corresponding regression coefficients
3. average treatment effect (ATE) under conditional unconfoundedness:

$$\mathrm{E}[Y_1 - Y_0] = \mathrm{E}\{\mathrm{E}[Y \mid A = 1, X] - \mathrm{E}[Y \mid A = 0, X]\},$$

   which is a function of $\mathrm{P}(X)$ and $\mathrm{P}(Y \mid A, X)$

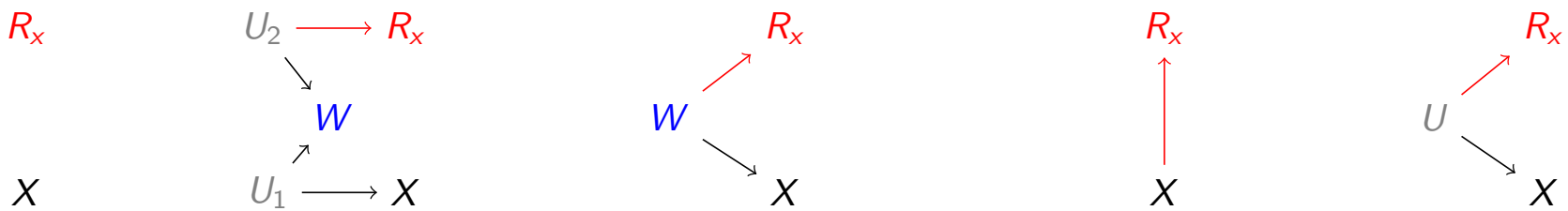
1. distribution or mean of a variable: $\mathrm{P}(X)$ or $\mathrm{E}[X]$

- if MCAR, complete-case analysis is valid, ie $\mathrm{P}(X) = \mathrm{P}(X \mid R_x = 1)$; it is also efficient
- if MAR (given W), is identified ($\mathrm{P}(X) = \mathrm{E}_W[\mathrm{P}(X \mid W, R = 1)]$), requires MAR-based methods
- if MNAR, unidentified



1. distribution or mean of a variable: $P(X)$ or $E[X]$

- if MCAR, complete-case analysis is valid, ie $P(X) = P(X \mid R_x = 1)$; it is also efficient
- if MAR (given W), is identified ($P(X) = E_W[P(X \mid W, R = 1)]$), requires MAR-based methods
- if MNAR, unidentified



In the case of the second graph (a MCAR case), if mistakenly assume MAR given W, the estimate is biased. Due to W being a collider of causes of X and R_x.

(This is related to the criticism of lack of specificity in the selection of auxiliary variables in pursuit of MAR.)

2. conditional distribution $\mathrm{P}(Y \mid A, X)$ (or corresponding reg coefs)

- Complete case analysis is valid if $Y \perp\!\!\!\perp (R_y, R_a, R_x) \mid X, A$.

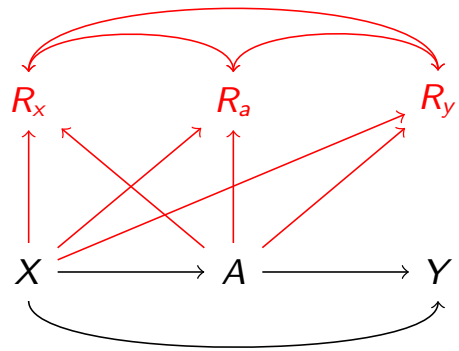
  This is a condition for "recovery" (Mohan and Pearl) of conditional distribution.

2. conditional distribution $P(Y \mid A, X)$ (or corresponding reg coefs)

- Complete case analysis is valid if $Y \perp\!\!\!\perp (R_y, R_a, R_x) \mid X, A$.

  This is a condition for "recovery" (Mohan and Pearl) of conditional distribution.

- This includes many cases
  - MCAR
  - Y MAR given (X, A) and X, A fully observed
  - A, Y, X all can have missing data, but the missingness is not caused by Y

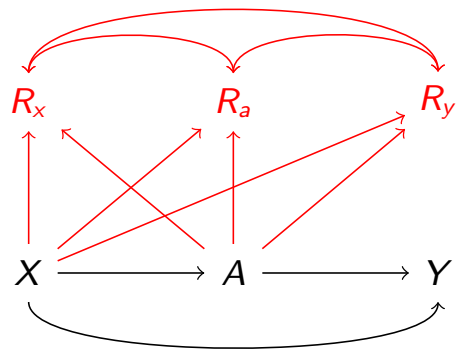


2. conditional distribution $P(Y \mid A, X)$ (or corresponding reg coefs)

- Complete case analysis is valid if $Y \perp\!\!\!\perp (R_y, R_a, R_x) \mid X, A$.

  This is a condition for "recovery" (Mohan and Pearl) of conditional distribution.

- This includes many cases
  - MCAR
  - Y MAR given (X, A) and X, A fully observed
  - A, Y, X all can have missing data, but the missingness is not caused by Y



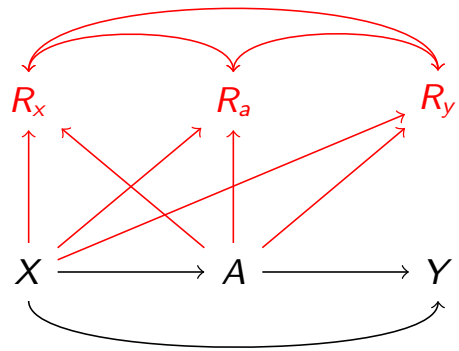
- This means $P(Y \mid A, X)$ can be recoverable by complete case analysis when X, A are MNAR
  (Note though that $P(X)$ and $P(X, A)$ are not identified in this case.)

2. conditional distribution $P(Y \mid A, X)$ (or corresponding reg coefs)

- Complete case analysis is valid if $Y \perp\!\!\!\perp (R_y, R_a, R_x) \mid X, A$.

  This is a condition for "recovery" (Mohan and Pearl) of conditional distribution.

- This includes many cases
  - MCAR
  - Y MAR given (X, A) and X, A fully observed
  - A, Y, X all can have missing data, but the missingness is not caused by Y



- This means $P(Y \mid A, X)$ can be recoverable by complete case analysis when X, A are MNAR
  - Consider the graph above, except Y is fully observed. If wrongly assume MAR and impute X, A, the estimate of $P(Y \mid A, X)$ is biased.
  - This emphasizes the importance of thinking carefully about what might be the mising data mechanisms

2. conditional distribution $P(Y \mid A, X)$ (or corresponding reg coefs)

- Complete case analysis is valid if $Y \perp\!\!\!\perp (R_y, R_a, R_x) \mid X, A$.

  This is a condition for "recovery" (Mohan and Pearl) of conditional distribution.

- This includes many cases
  - MCAR
  - Y MAR given (X, A) and X, A fully observed
  - A, Y, X all can have missing data, but the missingness is not caused by Y

- This means $P(Y \mid A, X)$ can be recoverable by complete case analysis when X, A are MNAR
  - Consider the graph above, except Y is fully observed. If wrongly assume MAR and impute X, A, the estimate of $P(Y \mid A, X)$ is biased.
  - This emphasizes the importance of thinking carefully about what might be the mising data mechanisms

- However, a challenge with complete case analysis is that a lot of data may be discarded due to covariate missingness
  - investigation needed to understand bias-variance trade-off

3. the ATE

This is a function of both $\mathrm{P}(X)$ and $\mathrm{P}(Y \mid A, X)$, specifically

- the conditional average treatment effect (CATE) is a function of $\mathrm{P}(Y \mid A, X)$
- averaging the CATE to identify the ATE requires $\mathrm{P}(X)$

In the situation on the last slide, the CATE is recoverable but the ATE is not. In this case

- one thought is to consider standardizing to a different covariate distribution $\mathrm{P}^*(X)$, eg that of a relevant target population
- but need to be careful in choosing estimation strategy to avoid running into unidentified elements of the underlying data distribution

This generally means that there is work to be done on both estimands and estimation.

That was just a glimpse

There are many other estimands, data settings and forms of missingness

- mediation
- survival analysis
- multi-level setting
- repeated outcomes and treatments
- settings with non-compliance or truncation due to death
- etc.

- monotone and non-monotone missingness
- intermittent missingness
- right-, left-, interval-censoring
- etc.

The missingness of each form in each setting may be caused by different mechanisms, with different implications for each estimand.

There are now theories about conditions for recoverability of parameters based on missingness graphs. They need to be applied to produce specific results and guidance for these different situations.

Small note on testability of MAR

(Mohan & Pearl 2021; also Potthoff et al. 2006)

Need more than one variable with missingness

V_1, V_2, V_3 with V_2, V_3 partially missing, V_1 fully observed

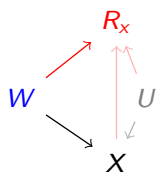
MAR implies

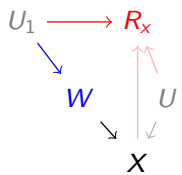
$$V_2 \perp\!\!\!\perp R_3 \mid V_1, R_2 = 1,$$
$$V_3 \perp\!\!\!\perp R_2 \mid V_1, R_3 = 1,$$

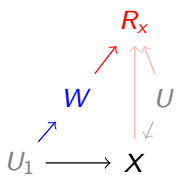
which are testable.

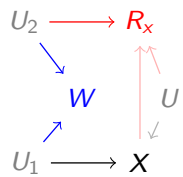
A bit more on auxiliary variables (W) to get close to MAR

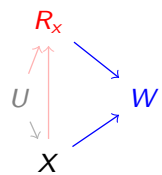
Simple case with one analysis variable X:





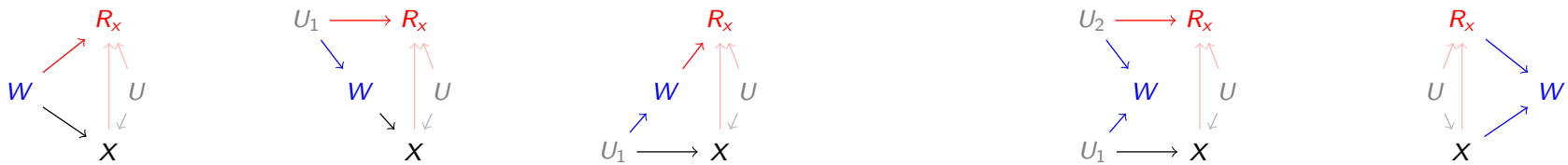






A bit more on auxiliary variables (W) to get close to MAR

Simple case with one analysis variable X:



In typical case with more analysis variables, more likely to run into collider bias.

Also, things get more complicated when auxiliary variables have missing values.

To sum up

We have done

- a brief review of conventional thinking about missingness mechanisms and strategies for dealing with missing data
- a quick visit with some recent ideas using causal graphs that provide ways to check the recoverability of target parameters under different missingness mechanisms and the testability of some assumptions

THANK YOU!

and I'd like to hear your thoughts on all this!

Some references

- Carpenter JR, Smuk M. (2020). Missing data: A statistical framework for practice. *Biometrical Journal*. 63:915-947. doi:10.1002/bimj.202000196.
- Daniel RM, Kenward MG, Cousens SN, De Stavola BL. (2011). Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research*. 21(3):243-256. doi:10.1177/0962280210394469.
- Enders CK. (2023). Missing Data: An Update on the State of the Art. *Psychological Methods*. doi:10.1037/met0000563.
- Li P, Stuart EA. (2019). Best (but oft-forgotten) practices: Missing data methods in randomized controlled nutrition trials. *The American Journal of Clinical Nutrition*.
- Mohan K, Pearl J. (2021). Graphical Models for Processing Missing Data. *Journal of the American Statistical Association*. 116(534):1023-1037. doi:10.1080/01621459.2021.1874961.
- Potthoff RF, Tudor GE, Pieper KS, Hasselblad V. (2006). Can one assess whether missing data are missing at random in medical studies?. *Statistical Methods in Medical Research*. 15:213-234.
- Rubin DB. (1976). Inference and missing data. *Biometrika*. 63(3):581-92.
- Rubin DB. (1987) *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.
- Thoemmes F, Mohan K. (2015). Graphical Representation of Missing Data Problems. *Structural Equation Modeling: A Multidisciplinary Journal*. 22(4):631-642, doi:10.1080/10705511.2014.937378.
- White IR, Royston P, Wood AM. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*. 30:377-399.