UMC Utrecht

Types of missing data

K.G.M. (Carl) Moons, PhD
k.g.m.moons@umcutrecht.nl

Introduction

Study is completed:

- All patients measured
- Determinants/predictors/covariates/independent variables/X'n AND the outcome/dependent variable/Y
- Data in computer
- Descriptive/frequency tables - MISSINGS

Introduction

- MISSINGS = problems → precision (loss of statistical power) + validity (bias)

- TO PREVENT IS BETTER THAN TO CURE

Introduction

- Missing data always occur (any study):
  - retrospective and prospective
  - existing (routine care) databases
  - large scale population-based
  - (even!) well organised randomised trials

- Challenge: how proper analyses with missing values?
  - unbiased effect estimates (validity)
  - precise effect estimates (precision)

Introduction

- Epidemiological analyses:
  - Association determinants (predictors/ covariates/confounders) with outcome
  - Multivariable (regression) analyses
  - What happens with participant record if one variable (X or Y) is missing?

    - = complete case (CC) analyses
    - = most common
    - = always affects precision of effect estimates (not all data used)
    - = commonly causes invalidity (bias)

Introduction

- Most (epidemiological) studies use complete case analysis → Ignoring (=one method)
- There are other methods to handle missings – may also cause bias.
- Type and severity bias depend on method used and type of missing data.

Type of missing values

- 3 types: MCAR, MAR, MNAR

- Missing Completely At Random (MCAR)
- Missing At Random (MAR)
- Missing Not At Random (MNAR)

Type of missing values

1. **Missing Completely At Random (MCAR)**
   - The probability that an observation is missing does not depend on 'anything' except chance
   - Examples?

   - The probability that the observation of a given variable for a certain subject is missing is constant for all subjects
     - Missingness NOT related to any other patient characteristics -> including the outcome status

Type of missing values

- If MCAR holds, almost all analytical methods (see later) for handling missing data give unbiased results, although less precise

- Realistic?

  - Reason: missing related to other patient characteristics, including outcome (!!!)
  - = MAR = missing at random

Type of missing values

2. Missing At Random (MAR)

most advanced methods work very well

- Probability that an observation is missing depends only on other observed values (patient characteristics, including the outcome)

Most advanced methods to handle missing values under MAR yield in principle unbiased + more precise study results → see later

Type of missing values

3. Missing Not At Random (MNAR)

- The probability that an observation is missing depends also on unobserved values.

- E.g. probability of missing on a variable depends on the true (but unknown) value of that variable itself

- Examples
  - sexual preference
  - income level (to estimate SES)
    - higher levels values have larger probability of being missed than lower values

homosexual/heterosexual (homosexual are less likely to put down their sexual preference in a questionnaire) or higher incomes don't want to write down their income

Type of missing values

- Missing data seldom (if ever) MCAR

  rare

- MNAR = problems → no general methods for properly dealing with MNAR data!

How to check likelihood of missing data being MCAR or rather MAR?

usually it's MAR - how can we find out?

Next slide = very important table in empirical research!

"most"
starts all analysis

Table. Distribution of co-variates among subjects without and with missing values (total n=398).

one hopes that those are a random subcategory

Variables	No missings n=246 (62%)		≥ 1 missing n=152 (38%)		p-value
Pulmonary embolism (outcome variable)	**47**	complete CCA	**36**	at least one NA	**0.02**
Dyspnoea	**80**		**66**		**<0.01**
Malignancy	**28**		**16**		**<0.01**
Surgery in previous 3 months	**24**		**16**		**0.04**
Prior deep venous thrombosis	**6**		**10**		**0.17**
Wheezing	**18**		**11**		**0.09**
Previous pulmonary embolism	**5**		**12**		**0.02**
Collapse with or without loss of consciousness	**10**		**5**		**0.06**
Signs of deep venous thrombosis	**11**		**7**		**0.15**
Age (years)*	**57**	**(17)**	**54**	**(18)**	**0.19**
Positive Chest x-ray	**43**		**36**		**0.17**
Respiratory rate (breaths/min)*	**22**	**(7)**	**18**	**(6)**	**<0.01**

index tests

values are associated with missingness - showed that NA were not MCAR

* Mean (sd)

they are different however!

Testing for MCAR/MAR

- Missing data CLEARLY not MCAR
  because otherwise the subsets would have been equal in observed characteristics
  - Analyzed subset of 246 subjects is not random subset of the original study sample (N=398) –> SELECTION bias due to missings
  - Missing related to other observed characteristics (incl. outcome) = MAR
    use observed characteristics to predict the missing values
  - If missingness related to observed characteristics – these can be used to estimate/predict the missing values!
  - Missing could still partly MNAR – but also MAR. Cannot test for MNAR – only reduce MNAR-part as much as possible by including many observed chars (increasing MAR)
    - Compare: adjustment for known confounders (MAR) versus residual confounders (MNAR)

    document known confounders as much as possible and asjust for known confounders - the more conf. the less residual confounders

    the more observed characteristics the less likely MNAR

- Exception: missing outcomes in RCTs - previous table not enough
  - See later (Groenwold RH et al: CMAJ 2014 + AM J EPI 2012)

Thank you for your attention