# Missing Data Mechanisms — Taxonomy, patterns, and implications

Henrik Støvring
(stovring@biostat.au.dk)

Department of Biostatistics

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

November 27, 2013

# Outline



SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

# Overview

- Taxonomy of missing data mechanisms
  - Definitions of MCAR, MAR, MNAR
  - Their implications for complete case analysis
- Determining the missing data mechanism
  - External information
  - Information within data
- Relation between mechanism and the missing values

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

# Missing Completely At Random (MCAR)

- Missingness unrelated to
  - observed values
  - unobserved values
- Equivalent to flipping a coin to determine whether an observation goes missing

# Missing Completely At Random (MCAR)

- Probability of a birthweight being missing does not depend on
  - Observed data (child gender, smoking status, etc.)
  - Un-observed data (low birthweight does not influence risk of observation being missing)
- In math:

$$R_i \text{ is independent of } Z_k \text{ for all } i, k$$

# Example: Birthweight and sex

- Data on 10,000 births
- Consider birth weight (grams) and sex of child (F/M)
- For 989 ($\approx$ 10%) birth weight was not recorded
- Missingness determined by chance alone:
  - Transfer to central server was interrupted
- I.e.: Unrelated to any characteristic of child

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

# Example: Birthweight and sex

- Effect on complete case analysis?
- Reduction in sample size
  - Does not bias estimates
  - Reduces precision (larger standard errors)
- Complete case analysis is valid, but possibly inefficient

# Missing At Random (MAR)

- Missingness unrelated to
  - unobserved values

  given whatever else we know
- Allowed to depend on observed values
- Interpretation:
  Outcomes for subjects with similar observed characteristics have the same distribution, whether observed or not

# Example: Birthweight and sex

- Probability of a birthweight being missing does not depend on
  - unobserved data value, i.e. the actual birthweight

  IF we know the other covariates
- Missingness may depend on observed characteristics (child gender, smoking status, etc.):
  - Suppose we know smoking status, $X_i$
  - Let $R_i$ denote whether birthweight $Y_i$ is recorded ($R_i = 1$) or missing ($R_i = 0$)
  - Under MAR, $P(R_i = 1)$ may depend on $X_i$
  - But once we know $X_i$ it must not depend on $Y_i$

# Missing Not At Random (MNAR)

- Probability of an observation being missing depends on
  - Unobserved data values

  even after adjustment for other covariates
- Means that data are neither MCAR or MAR
- Implication: A missing observation has a different distribution than observed values of other individuals, even when they otherwise have the same characteristics

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

# MNAR example

- Study of diabetes mortality in relation to HbA1c (average blood sugar level)
- Should be measured at least twice a year
- Some patients do not have a measurement in the register in the year of inclusion into the study
- Many reasonable explanations for this:
  - deteriorating health - high HbA1c values?
  - very well controlled diabetes - low HbA1c values?
  - patient is not interested in disease control - high HbA1c values?
  - ...

# In-class exercise: Example battle

Think of your own (PhD) project. Give your opponent two examples of a variable with missing data in your study. Explain why one variable reflects the MAR situation, and why the other is compatible with MNAR. What is it that makes one MAR and the other MNAR?
Pick for each category, MAR and MNAR, the more convincing example. You and your opponent now teams up and battles the neighboring pair: Retell the examples and pick a MAR winner and a MNAR winner.

# Can we detect MNAR?

- A catch-22 situation:
  - The observation is missing, in part because of the unobserved value
- Relations to observed values are insufficient
- Relations to observed values may even be irrelevant
- What we don't have is what we would need to establish or rule out MNAR

# What can we do about MNAR?

- If data are MNAR, our observed data is not representative of general study population
- Any adjustment would hinge on
  - Direction of the discrepancy between missing and observed data
  - Magnitude of the discrepancy between missing and observed data
- Cannot be known - since we don't have this info in our data
- Alternative strategies
  - Sub-sample of previous non-responders
  - Sensitivity analysis
  - Model likely values from external sources

# Observed causes of missing information

- Hardware failure
- Human error (dropped blood sample)
- Migration, did not want to answer, relapse $\leftarrow$ often recorded in data
- Note: Info must not depend on observed values
  Do not look at data and then re-consider external explanations

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

# Interpretations

- If unrelated to responses
  $\rightarrow$ Missing Completely At Random (MCAR)
- Be cautious:
  - Temporal trends
  - Association with test site
  - Association with interviewer
  - etc...
- May all cause Missing At Random (MAR)
- Remember: External info may also indicate Missing Not At Random (MNAR)

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

# Missing as outcome

- Interest is on missing status itself
- For each variable with missing values
  - Define a binary variable: 1 if missing, 0 otherwise
  - Use a regression model to study association with other variables
- Insight gained depends on what we find:

  Association found: Reject MCAR

  No association found: Cannot reject MCAR
- Remember: Have never ruled out MNAR

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

# No assocation – what does it mean

- Equivalent to an ordinary finding of no association
- Does not mean:
  - No association at all
  - No important association
- But only: we cannot in the present data rule out MCAR
- May possibly be due to
  - sample size
  - too few missing values
- Look for trends – reconsider a priori knowledge on mechanism

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

# Example I: One parameter for missningness, one sample

- Assume data are on body temperature:
  - Independent observations
  - Normally distributed
- Missing data mechanism
  - Electronic thermometer
  - Fails $100p\%$ of the time (loose connection)

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

# Example I (cont'd)

- Measured temperature of healthy subjects
- Model for temperature

$$Y_i \sim N(\mu, \sigma^2)$$

- Model for missingness

$$P(R_i = 0) = p$$

- No association between $Y_i$ and $R_i$:

$$P(Y_i < y, R_i = r) = P(Y_i < y)P(R_i = r), \text{ for all } y, r_i$$

- Actual number of observed values will vary across samples:
  $\rightarrow$ on average 100% of observations missing

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

# Example I (cont'd)

- Observe pairs $(y_i^*, r_i)$ where

$$y_i^* = \begin{cases} y_i & \text{if } r_i = 1 \\ ? & \text{if } r_i = 0 \end{cases}$$

- Likelihood for inference on $(\mu, \sigma, p)$

$$\begin{aligned} l(\mu, \sigma, p; y_i^*, r_i) &= \prod_{r_i=1}(1-p)f(y_i;\mu,\sigma)\prod_{r_i=0} p \int_{-\infty}^{\infty} f(s;\mu,\sigma)\,ds \\ &= \prod_{r_i=1}(1-p)f(y_i;\mu,\sigma)\prod_{r_i=0} p \end{aligned}$$

  where $f(y;\mu,\sigma)$ is density of $Y$
- Not interested in $p$, only in $(\mu, \sigma)$:

$$l(\mu, \sigma; y_i^*, r_i) = \prod_{r_i=1} f(y_i;\mu,\sigma)$$

- The same likelihood as in the complete case

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

# Example II: Two parameter model for missingness, two samples problem

- Measured temperature of women and men
- Model for temperature

$$Y_{ij} \sim N(\mu_j, \sigma^2)$$

where $j$ denotes gender

- Model for missingness (two different, randomly failing thermometers)

$$P(R_{ij} = 0) = p_j$$

- Within gender groups no association between $Y_{ij}$ and $R_{ij}$:

$$\begin{aligned} P(Y_{iF} < y, R_{iF} = r) &= P(Y_{iF} < y)P(R_{iF} = r) \\ P(Y_{iM} < yR_{iM} = r) &= P(Y_{iM} < y)P(R_{iM} = r) \end{aligned}$$

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

# Example II (cont'd)

- Population distribution (random individual)

$$P(Y_i < y) = P(j = F)P(Y_{iF} < y) + P(j = M)P(Y_{iM} < y)$$

- Distribution among non-missing

$$P(Y_i < y | R_i = 1) = \frac{P(Y_i < y, R_i = 1)}{P(R_i = 1)}$$

- Numerator

$$\begin{aligned} P(Y_i < y, R_i = 1) &= P(Y_{iF} < y, R_{iF} = 1)P(j = F) \\ &\quad + P(Y_{iM} < y, R_{iM} = 1)P(j = M) \\ &= P(Y_{iF} < y)p_F P(j = F) \\ &\quad + P(Y_{iM} < y)p_M P(j = M) \end{aligned}$$

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

# Example II (cont'd)

- Denominator

$$P(R_i = 1) = p_F P(j = F) + p_M P(j = M)$$

- Assume equally many men and women in population
- Population distribution

$$P(Y_i < y) = \frac{1}{2}(P(Y_{iF} < y) + P(Y_{iM} < y))$$

- Distribution among non-missing

$$P(Y_i < y | R_i = 1) = \frac{p_F P(Y_{iF} < y) + p_M P(Y_{iM} < y)}{p_F + p_M}$$

# Example II (cont'd)

- The mean in the population:

$$\mu = \frac{\mu_F + \mu_M}{2}$$

- Average mean in samples with missing data

$$\mu_{obs} = \frac{p_F \mu_F + p_M \mu_M}{p_F + p_M}$$

- If $p_F \neq p_M$ and $\mu_F \neq \mu_M$ then in general $\mu \neq \mu_{obs}$
- But note: Estimate of $\mu_F$ and $\mu_M$ OK
- I.e.: Estimate of $\mu_F - \mu_M$ also OK

# Example III: Two sample problem, missingness depends on mean

- Measured temperature of women and men
- Model for temperature

$$Y_{ij} \sim N(\mu_j, \sigma^2)$$

where $j$ denotes gender

- Thermometer fails more often when temperature is high:

$$\begin{aligned} P(R_{ij} = 0 | Y_{ij} \leq 38^\circ\text{C}) &= p \\ P(R_{ij} = 0 | Y_{ij} > 38^\circ\text{C}) &= \pi \end{aligned}$$

where $\pi > p$

- For example: $p = 25\%$ and $\pi = 50\%$

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

# Example III (cont'd)

- Consequences:
  - Mean underestimated in each group if based on direct estimate from observed data
  - Overall mean underestimated
  - Difference of means underestimated (big - small)
  - If for example 20% of all females' temperatures are above 38°C in population
    $\rightarrow$ 14.2% of observed females are above
  - Fraction of missing data differ between females and males, if mean temperature depends on gender

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY

# Important points

- Complete case analysis works for MAR data
- Any other situation requires more careful analysis
- For any given situation, one purpose may be OK with standard analyses, while another purpose with the same data may require a dedicated handling of the missing data

SCHOOL OF PUBLIC HEALTH
FACULTY OF HEALTH SCIENCES
AARHUS UNIVERSITY