# Help! Statistics!

# Missing data. An introduction

Sacha la Bastide-van Gemert

Medical Statistics and Decision Making

Department of Epidemiology

UMCG

## Help! Statistics! Lunch time lectures

What? Frequently used statistical methods and questions in a manageable timeframe for all researchers at the UMCG.
No knowledge of advanced statistics is required.

When? Lectures take place every 2nd Tuesday of the month, 12.00-13.00 hrs.

Who? Unit for Medical Statistics and Decision Making

*>>> lectures will be announced on the UMGC Intranet Agenda <<<*

| When? | Where? | What? | Who? |
|---|---|---|---|
| Sep 11, 2018 | Room 16 | Binary and ordinal logistic regression | H. Burgerhof |
| Oct 9, 2018 | Rode zaal | ….. | ….. |
| Nov 13, 2018 | Room 16 | ….. | ….. |
| Dec 11, 2018 | Room 16 | ….. | ….. |

Slides can be downloaded from:
http://www.rug.nl/research/epidemiology/download-area

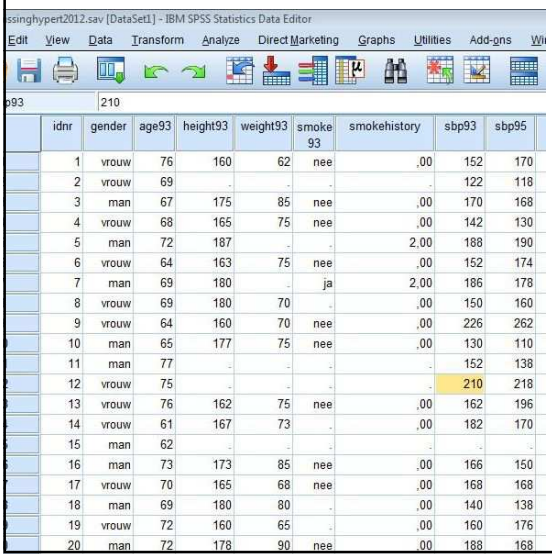
## Missing data. An introduction

- What constitute ``missing data''?
- Why are missing data a problem?
- Causes and mechanisms of missing data

  *MCAR, MAR and MNAR*
- Analysis with missing data

  *complete cases analysis, available case analysis, summary measures, single imputation, multiple imputation, likelihood based methods, …*
- Summary: general guidelines for dealing with missing data

## What constitute ``missing data''?




| | idnr | gender | age93 | height93 | weight93 | smoke 93 | smokehistory | sbp93 | sbp95 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | vrouw | 76 | 160 | 62 | nee | ,00 | 152 | 170 |
| 2 | vrouw | 69 | . | . | . | . | 122 | 118 |
| 3 | man | 67 | 175 | 85 | nee | ,00 | 170 | 168 |
| 4 | vrouw | 68 | 165 | 75 | nee | ,00 | 142 | 130 |
| 5 | man | 72 | 187 | . | . | 2,00 | 188 | 190 |
| 6 | vrouw | 64 | 163 | 75 | nee | ,00 | 152 | 174 |
| 7 | man | 69 | 180 | . | ja | 2,00 | 186 | 178 |
| 8 | vrouw | 69 | 180 | 70 | . | ,00 | 150 | 160 |
| 9 | vrouw | 64 | 160 | 70 | nee | ,00 | 226 | 262 |
| 10 | man | 65 | 177 | 75 | nee | ,00 | 130 | 110 |
| 11 | man | 77 | . | . | . | . | 152 | 138 |
| 12 | vrouw | 75 | . | . | . | . | 210 | 218 |
| 13 | vrouw | 76 | 162 | 75 | nee | ,00 | 162 | 196 |
| 14 | vrouw | 61 | 167 | 73 | . | ,00 | 182 | 170 |
| 15 | man | 62 | . | . | . | . | . | . |
| 16 | man | 73 | 173 | 85 | nee | ,00 | 166 | 150 |
| 17 | vrouw | 70 | 165 | 68 | nee | ,00 | 168 | 168 |
| 18 | man | 69 | 180 | 80 | . | ,00 | 140 | 138 |
| 19 | vrouw | 72 | 160 | 65 | . | ,00 | 160 | 176 |
| 20 | man | 72 | 178 | 90 | nee | ,00 | 188 | 168 |

Missing values (``non-response'') occur frequently, especially in observational data

Missing values can occur:
- in covariates,
- in outcome,
- per patient (dropout), ...

•Unit non-response:
subjects/patients/cases are missing

•Item non-response:
values for some variables are missing

## Reasons for missing values

What could be reasons for missingness?

- Natural processes
  - eg in longitudinal studies: dropout, death
- Non-applicability: skipping questions in a questionnaire
  - eg questions only for persons who eat meat
- Informative non-response
  - eg patients are too ill to participate in a study
- Logistic/administrative reasons
  - eg due to data management, patient recruiting

Report and document reasons for missingness when and where possible: these might be found in the study documentation, questionnaire, etc.

## Why are missing data a problem?

1. Reduction in sample size (n):
   - less precision/efficiency
   - lower power
2. Possible introduction of bias
   - differences in completers/non-completers
   - depends on type of underlying missing data process and statistical analysis used

Problem: different ways of dealing with missing data may result in different results and conclusions

Important question to answer first:
what is the missing data process/mechanism?

## Mechanisms of missing data

The missing mechanism:

How are missing data distributed? What is the probability of missingness?

- Certain groups may be more likely to have missing values
  e.g. controls more than cases
- Certain responses may be more likely to have missing values
  e.g. persons with high income may be less likely to report income

Identification of missingness is crucial:

not all statistical methods in combination with specific missing mechanisms will yield unbiased results

Rubin (1976) distinguished three mechanisms: MCAR, MAR, MNAR

## Missing Completely At Random: MCAR (1)

The mechanism for missing values is independent of the observed and unobserved values

- The observed values can be seen as a random sample of the intended data
- The data can be analyzed as if the collection of data was intended this way (although loss in power remains)

no bias

(for almost all statistical techniques)

Examples:

- Technical failure of the measurement system
- Human failure or planning issues

## Missing Completely At Random: MCAR (2)

MCAR in statistical terms:

Y: variable that has missing data

X: set of variables with no missing data

*variables included in study*

Z: causes of missingness unrelated to X and Y

R: indicator for missingness of Y:

- R=1 if Y is missing
- R=0 if Y is not missing


MCAR: the probability that Y is missing, does not depend on X or Y

$$P(R=1|X,Y) = P(R=1)$$

## Missing At Random (MAR) (1)

Missingness depends not on the variable value itself, but does depend on other (observed) variable values

bias,
but can possibly be dealt with using the correct analysis

i.e. the mechanism for missing values is

- independent of the unobserved values
- dependent on the observed values

- The observed values can no longer be seen as a random sample of the intended data
- As a consequence, the observed values are a biased sample of the intended data

Example:

- protocol could state that an extreme response outcome would require an additional measurement

## Missing At Random: MAR (2)

MAR in statistical terms:

Y: variable that has missing data

X: set of variables with no missing data

*variables included in study*

Z: causes of missingness unrelated to X and Y

R: indicator for missingness of Y:

- R=1 if Y is missing
- R=0 if Y is not missing


MAR: the probability that Y is missing, does not depend on Y, *when conditioned on X*:

$$P(R=1|X,Y) = P(R=1|X)$$

## Missing Not At Random: MNAR (1)

The mechanism for missing values is dependent of the unobserved values and referred to as **non-ignorable**:

- Analysis of the data requires a model for the missing mechanism to overcome bias by the missing data
- Models for missing mechanisms are unverifiable
  - Since the relation with the unobserved values are unknown
  - Different models could lead to different conclusions
  - This implies the need for sensitivity analysis

bias, difficult to solve

(additional information needed)

Examples:

- Subjects are too ill to attend the appointment

## Missing Not At Random: MNAR (2)

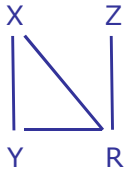
MNAR in statistical terms:

Y: variable that has missing data
X: set of variables with no missing data
} *variables included in study*

Z: causes of missingness unrelated to X and Y

R: indicator for missingness of Y:

- R=1 if Y is missing
- R=0 if Y is not missing




MNAR: the probability that Y is missing depends on Y and X:

"P(R=1|X,Y) = P(R=1|X,Y)"

## MCAR, MAR, MNAR: examples (1)

Example: studying weight (Y) while (amongst others) taking into account gender (X)

Possible mechanisms causing missing values of Y:

1) The research-assistant accidentally deleted part of the weight-data
   >>> type of missing?
2) His/her dog eats part of the paper-questionnaires
   >>> type of missing?
3) One gender-type was less likely to disclose its weight
   >>> type of missing?
4) Heavy (or light) people may be less likely to disclose their weight
   >>> type of missing?

MCAR
(no relationship to X or Y),
**no bias**

MAR
(missingness depends on X),
**no bias when corrected for X in the analysis**

MNAR
(missingness depends on Y),
**bias**
(additional information would be needed to solve this)

## MCAR, MAR, MNAR: examples (2)

Example (Schafer&Graham 2002): studying blood-pressure (Y) repeatedly over time (longitudinal data). In January: first measurements are taken.

Possible mechanisms causing missing values of Y:

1) February: part of the patients does not show up due to bad weather
   >>> type of missing?

MCAR

2) February: patients who did not have high blood-pressure in January do not show up
   >>> type of missing?

MAR
(missingness depends on variables from the past, but not on those from the present)

3) February: only those measurements were recorded, that were too high
   >>> type of missing?

MNAR
(missingness depends on values from present)

## Dealing with missing data in statistical analyses

- Complete case analysis (listwise deletion)
- Available case analysis (pairwise deletion)
- Summary measures (longitudinal data)
- Single imputation
- Multiple imputation
- Likelihood based methods
- *Selection and pattern-mixture models*

# Complete case analysis (listwise deletion)

| idnr | gender | age93 | height93 | weight93 | smoke 93 | smokehistory | sbp93 | sbp95 |
|---|---|---|---|---|---|---|---|---|
| 1 | vrouw | 76 | 160 | 62 | nee | ,00 | 152 | 170 |
| 2 | vrouw | 69 | . | . | . | . | 122 | 118 |
| 3 | man | 67 | 175 | 85 | nee | ,00 | 170 | 168 |
| 4 | vrouw | 68 | 165 | 75 | nee | ,00 | 142 | 130 |
| 5 | man | 72 | 187 | . | . | 2,00 | 188 | 190 |
| 6 | vrouw | 64 | 163 | 75 | nee | ,00 | 152 | 174 |
| 7 | man | 69 | 180 | . | ja | 2,00 | 186 | 178 |
| 8 | vrouw | 69 | 180 | 70 | . | ,00 | 150 | 160 |
| 9 | vrouw | 64 | 160 | 70 | nee | ,00 | 226 | 262 |
| 10 | man | 65 | 177 | 75 | nee | ,00 | 130 | 110 |
| 11 | man | 77 | . | . | . | . | 152 | 138 |
| 12 | vrouw | 75 | . | . | . | . | 210 | 218 |
| 13 | vrouw | 76 | 162 | 75 | nee | ,00 | 162 | 196 |
| 14 | vrouw | 61 | 167 | 73 | . | ,00 | 182 | 170 |
| 15 | man | 62 | . | . | . | . | . | . |
| 16 | man | 73 | 173 | 85 | nee | ,00 | 166 | 150 |
| 17 | vrouw | 70 | 165 | 68 | nee | ,00 | 168 | 168 |
| 18 | man | 69 | 180 | 80 | . | ,00 | 140 | 138 |
| 19 | vrouw | 72 | 160 | 65 | . | ,00 | 160 | 176 |
| 20 | man | 72 | 178 | 90 | nee | ,00 | 188 | 168 |

All respondents with missing data are left out: only subjects with complete data are part of the statistical analysis

Advantages:
- Easy to apply
- Can be used with any statistical technique/model
- Same data set for each analysis

Disadvantages:
- Inefficient due to reduction in sample size (n)
- Induces bias when missing data are MAR or MNAR
- Only recommended when percentage missingness < 5%

17

# Available cases analysis (pairwise deletion)

| idnr | gender | age93 | height93 | weight93 | smoke 93 | smokehistory | sbp93 | sbp95 |
|---|---|---|---|---|---|---|---|---|
| 1 | vrouw | 76 | 160 | 62 | nee | ,00 | 152 | 170 |
| 2 | vrouw | 69 | . | . | . | . | 122 | 118 |
| 3 | man | 67 | 175 | 85 | nee | ,00 | 170 | 168 |
| 4 | vrouw | 68 | 165 | 75 | nee | ,00 | 142 | 130 |
| 5 | man | 72 | 187 | . | . | 2,00 | 188 | 190 |
| 6 | vrouw | 64 | 163 | 75 | nee | ,00 | 152 | 174 |
| 7 | man | 69 | 180 | . | ja | 2,00 | 186 | 178 |
| 8 | vrouw | 69 | 180 | 70 | . | ,00 | 150 | 160 |
| 9 | vrouw | 64 | 160 | 70 | nee | ,00 | 226 | 262 |
| 10 | man | 65 | 177 | 75 | nee | ,00 | 130 | 110 |
| 11 | man | 77 | . | . | . | . | 152 | 138 |
| 12 | vrouw | 75 | . | . | . | . | 210 | 218 |
| 13 | vrouw | 76 | 162 | 75 | nee | ,00 | 162 | 196 |
| 14 | vrouw | 61 | 167 | 73 | . | ,00 | 182 | 170 |
| 15 | man | 62 | . | . | . | . | . | . |
| 16 | man | 73 | 173 | 85 | nee | ,00 | 166 | 150 |
| 17 | vrouw | 70 | 165 | 68 | nee | ,00 | 168 | 168 |
| 18 | man | 69 | 180 | 80 | . | ,00 | 140 | 138 |
| 19 | vrouw | 72 | 160 | 65 | . | ,00 | 160 | 176 |
| 20 | man | 72 | 178 | 90 | nee | ,00 | 188 | 168 |

= left out in univariate analysis (outcome: sbp95)

= additionally left out in analysis corrected for height93 ...

= ... and when corrected for smoke93 ...

Only complete cases <u>for the variables in each specific model</u> are analysed (or: from specific wave in longitudinal settings)

Advantages:
- Easy to apply
- Can be used with any statistical technique/model
- Uses more available information
- More efficient than complete case analysis

Disadvantages:
- Different data sets (different n!) for different analyses
- Induces bias when missing data are MAR or MNAR

## Summary measures

In a longitudinal (repeated measures) setting:

Reduce (summarize) variables with missing measurements by choosing a particular aspect of the data

e.g. maximum, minimum, median, mean,…

Not generally recommended

Advantages:

- Easy to apply
- Can be used with any statistical technique/model

Disadvantage:

- Inefficient use of data
- Induces bias when missing data are MAR or MNAR

## Single imputation

Replaces missing values with sensible estimates (``educated guess''), resulting in a data set without missing values.

Various choices:

- Last Value Carried Forward (LVCF)
- mean imputation
- imputation based on a regression model, ….

Not generally recommended

Advantages:

- Easy to apply
- Can be used with any statistical technique/model
- No information is discarded

Disadvantages:

- Missingness uncertainty is not taking into account: underestimation of standard errors (decreases variances)
- Only valid under specific MCAR and MAR mechanisms

## Multiple imputation (1)

Multiple imputation: replace missing value with ``educated guess", not once, but multiple (m) times

- Takes into account the correlation with other variables
- Provides better estimates of the standard error (takes into account variation)

Widely recommended approach; based on (assumptions on) the joint distribution of the data

Multiple imputation combines analyses-results from a specific number of (model-based imputed, hence complete) data sets, using the following three steps: (…)

## Multiple imputation (2)

Step 1: Imputation step

- First, the imputation model is defined by the researcher:
  - type of model is chosen (e.g. regression)
  - selection of predictors used in the model is chosen
- Based on the imputation model and available data, plausible values are drawn for imputation in the missing data set

  *Now: appropriate random variation is introduced!*
- This step is repeated m times
- Result: m (slightly) different, complete data sets

## Multiple imputation (3)

Step 2: Statistical Analysis

- Analyse each imputed and complete data set with standard statistical techniques
- Result: m (different) estimates q of the parameter of interest

(e.g. a mean, a regression coefficient, …):

$q_1, q_2, \ldots\ldots, q_m$ ,

each with its standard error:

$v_1, v_2, \ldots\ldots, v_m$

## Multiple imputation (4)

Step 3: Combination or estimation step

- Combines the results from the m analyses
- Usually this means a combined estimate $\overline{q}$ for the parameter of interest (using estimates $q_1, q_2, \ldots\ldots, q_m$) and an appropriate standard error (using $v_1, v_2, \ldots\ldots, v_m$) containing both between imputation variance and within imputation variance

- *In SPSS: multiple imputation module*

## Multiple Imputation (5)

How many imputations are needed?

- General advise was to take 5 or 6 imputed data sets

  (SPSS default: 5)
- Researchers investigated the size of imputed data sets for different settings, like confidence intervals and power
- These studies suggested to take at least 10 to 20 imputed or even more data sets
- Advice: start with 5 imputed data sets, perform model selection and finish with a higher number for a final analysis

## Multiple imputation (6)

Advantages:

- Results in better estimates of the standard errors than previous approaches
- Missing data process can be incorporated into the imputation model
- Results in unbiased estimates under the assumption of MCAR and MAR

Disadvantages:

- Can be a lot of work
- Results depend on (many possible) choices made by the researcher for the imputation model, this can be tricky!

## Maximum Likelihood (ML) based methods

Maximum Likelihood estimation (ML):

estimation of parameters through the maximization of the likelihood function (the theoretical distribution of the data)

i.e.: for a given data set and probability model, ML estimation finds those values of the model parameters that are most probable, given the observed data

All observed values are taken into account:
unbiased effect estimates under MCAR or MAR

Widely recommended approach; based on (assumptions on) the joint distribution of the data

ML estimation methods are particularly applied in mixed effects models (multilevel models, repeated measurement analysis)

*In SPSS Mixed Models module likelihood functions are used for estimating*

## Likelihood based methods

Advantages:

- Unbiased results under MCAR and MAR
- Where available in the software, it is relatively easy to use

Disadvantages:

- Results and reliability depend on choices made by the researcher regarding (the assumptions on) the joint distribution of the data
- In general: requires background of more advanced knowledge of statistics

## MCAR, MAR or MNAR?

As not all statistical techniques are valid under different missing data mechanisms, categorization of the missing mechanism is crucial.

No ultimate test exists, but there are ways to help diagnose missingness:

- MCAR versus MAR:
  - Little's test (but might be wrong, as all tests…)
  - use missingness-dummy variables and chi-square or t-tests to help suggest patterns (both available in SPSS missing value module)
- MAR versus MNAR:
  - use information on the missing data whenever possible (f.e. follow-up phone-calls in a survey)
  - examine model fit of different patterns of missingness
  - use commonly known information on ignorability of missing data from your research field
- Be aware: for different variables with missing data within the same study, different types of mechanisms may play a role!

## Use sensitivity analysis

- Multiple imputation and maximum likelihood based approaches assume MAR: most likely not true for most studies
- With limited amounts of missingness: assumption of MAR might be fine
- For MNAR with a substantial amount of missing data, even multiple imputation and maximum likelihood approaches result in biased estimates
- MNAR: likelihood based analyses are needed, jointly modelling both missingness and response:
  - Selection models
  - Pattern mixture models

  } *beyond scope of this introduction*
- In general: investigate robustness of choices using sensitivity analysis

## Summary
## General guidelines for missing data

- ``Prevention is better than cure'': put time and effort in preventing and retrieving missing values
- Always report details of missing data and state your assumptions explicitly
- Under the assumption of MAR: use multiple imputation or likelihood based methods (check your software for availability)
- In case of MNAR: model missing data processes

  (*pattern mixture models or model selection models*)
- Use sensitivity analyses

## A selection of literature


Graham, JW, Missing data analysis: making it work in the real world. Annu. Rev. Psychol. 2009 (60) 549-576

Schafer, JL, Graham JW. Missing data: our view of the state of the art. Psychological Methods 2002

Pigott, TD. A review of methods for missing data. Educational Research and Evaluation 2001 7 (4) 353-83

Donders AR et al. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006 59 (10) 1087-91

JAC Sterne et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009 338

Moons, KGM, Donders, RART, Stijnen, T, Harrel, FE. Using the outcome for imputation of missing predictor values was preferred. Journal of Clinical Epidemiology. 2006 (59) 1092-1101

PD Allison. Handling missing data by Maximum Likelihood. Statistics and Data Analysis SAS 2012 Paper 312

Next Help! Statistics! lecture:

September 11, 2018

Room 16

**Binary and Ordinal Logistic Regression**

Hans Burgerhof