INTRODUCTION TO MISSING DATA - CONCEPTS AND PERSPECTIVES

Juha Karvanen and Kari Auranen

Tilastopäivät, Turku, May 18, 2017

Missing data or missing information

- *Data* (Latin), plural of datum = something given (!)
- Missing data or missing information?

- The science of statistics is all about missing information!
  - Nevertheless, we often need to be more systematic about what "missingness" means
  - In particular, we need to delineate the assumptions that allow principled inference under missing data or missing information

"Just ignoring missing data is not an acceptable option when planning, conducting or interpreting the analysis of a confirmatory clinical trial" (Guideline on Missing Data in Confirmatory Clinical Trials, EMA, 2010)

Concepts vs. practice

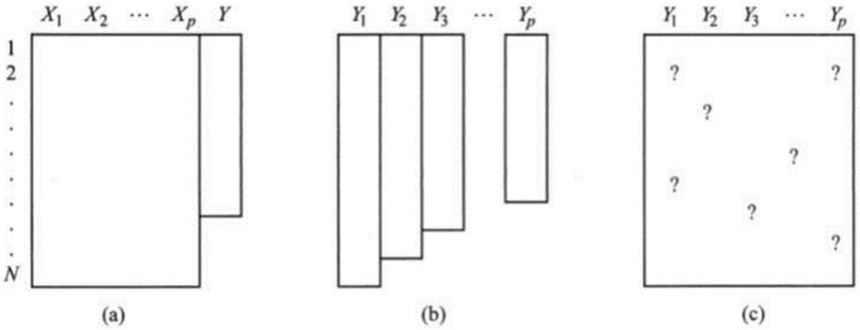
- Concepts are important because they guide the art of dealing with missing data/information
  - Intentional vs. unintentional missingness
  - Missing data mechanisms: MCAR, MAR, MNAR
  - Inferential framework and the ignorability of missingness
- In practice, dealing with missing data calls for probability modelling of the "data generating process"
  - including the observation process
- Untestable assumptions often need to be made about the reasons of why and how data might be missing

Intentional vs. unintentional missingness

- Intentional/planned missingness, i.e. missing information by design, arises in many settings, e.g.
  - Surveys, in which only a portion of the population is included in the sample
  - Randomised experiments, in which the contrafactual outcomes are not observed
  - Observational studies with systematic sampling
    - e.g.many models with latent variables
- But also unintentional missingness occurs very often
  - Non-response, attrition/drop-out, informative censoring, post-randomisation events, ...
  - Causes typically more severe questions about how to deal with what has not been observed

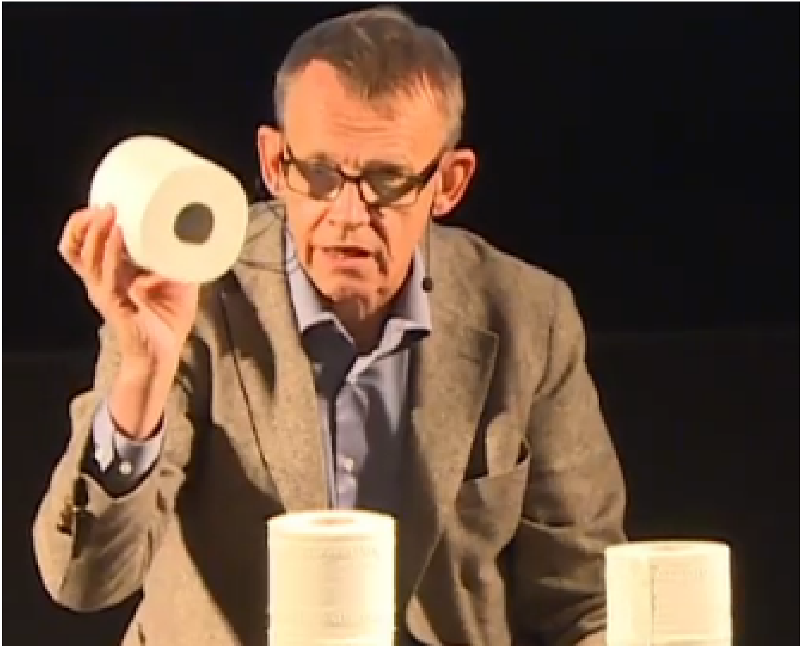
Patterns of non-response

- univariate
- monotone
- arbitrary



MISSING DATA MECHANISMS

Tribute to Hans Rosling 1948–2017



Observed data, missing data and complete data

- Complete data Y split into two parts: $Y = (Y_{obs}, Y_{mis})$
- Define an observable (vector) R as an indicator of missingness (or, rather, indicator of response) for each unit of Y
  - $R_i = 1$ if Y_i is observed and 0 otherwise
- Assume a parametric joint model $p(Y, R|\theta, \phi)$, where parameters θ characterise the complete data and parameters ϕ characterise the missing data mechanism
- We may then write

$$\begin{aligned} p(Y_{obs}, R|\theta, \phi) &= \int p(Y_{obs}, Y_{mis}, R|\theta, \phi) dY_{mis} \\ &= \int p(R|Y_{obs}, Y_{mis}, \phi) p(Y_{obs}, Y_{mis}|\theta) dY_{mis} \end{aligned}$$

- A big question: What can be assumed about the missing data mechanism $p(R|Y_{obs}, Y_{mis}, \phi)$?

Types of missing data mechanisms

- MCAR (missing completely at random)

$$p(R|Y_{obs}, Y_{mis}, \phi) = p(R|\phi) \text{ for all } Y_{obs}, Y_{mis}, \phi$$

- MAR (missing at random)

$$p(R|Y_{obs}, Y_{mis}, \phi) = p(R|Y_{obs}, \phi) \text{ for all } Y_{mis}, \phi$$

- Otherwise MNAR (missing not at random)

These are typical textbook definitions but there is **a problem**: Missingess indicator R depends on Y_{obs} but Y_{obs} itself is a function of R.

Realised MAR vs. Everywhere MAR

- Seaman et al. (2013) clarify the definition of MAR
- Random variables: complete data $\mathbf{Y}$, missingness indicators $\mathbf{R}$, observed data $o(\mathbf{Y}, \mathbf{R})$

  Realised values: complete data $\tilde{\mathbf{y}}$, missingness indicators $\tilde{\mathbf{r}}$, observed data $o(\tilde{\mathbf{y}}, \tilde{\mathbf{r}})$

  Missing data mechanism: Assumed model $g_\psi(\mathbf{R} = \mathbf{r} \mid \mathbf{Y} = \mathbf{y})$, where ψ is an unknown parameter
- The data are **realised MAR** if for all ψ

$$g_\psi(\tilde{\mathbf{r}} \mid \mathbf{y}) = g_\psi(\tilde{\mathbf{r}} \mid \tilde{\mathbf{y}}) \text{ for all } \mathbf{y} \text{ such that } o(\mathbf{y}, \tilde{\mathbf{r}}) = o(\tilde{\mathbf{y}}, \tilde{\mathbf{r}})$$

- The data are **everywhere MAR** if for all ψ

$$g_\psi(\mathbf{r} \mid \mathbf{y}) = g_\psi(\mathbf{r} \mid \mathbf{y}^*) \text{ for all } \mathbf{r}, \mathbf{y}, \mathbf{y}^* \text{ such that } o(\mathbf{y}, \mathbf{r}) = o(\mathbf{y}^*, \mathbf{r})$$

Realised MAR vs. Everywhere MAR: an example

- Realised data:

  Realised complete data: $\tilde{\mathbf{y}} = (10, 3, 4, 2)^T$

  Realised missingness indicators: $\tilde{m} = (1, 0, 1, 1)^T$

  Realised observed data: $o(\tilde{\mathbf{y}}, \tilde{\mathbf{r}}) = (10, 4, 2)^T$

- The data are realised MAR if, for any ψ

  $$g_\psi\left((1,0,1,1)^T|(10,a,4,2)^T\right) = g_\psi\left((1,0,1,1)^T|(10,b,4,2)^T\right)$$

  for values a and b in the sample space of the second element of $\mathbf{Y}$.

- The data are everywhere MAR if a similar equality holds for all possible values $\mathbf{Y}$ and $\mathbf{R}$. Everywhere MAR implies realised MAR.

IGNORABILITY

MAR revisited

- The realised MAR assumption entails the following subtleties
  - Missingness is only considered conditionally on its *realised* pattern, i.e. the information about which observations were missing and which were not
  - Given that realised pattern, the missingness model consideres observations that could potentially have been observed
  - For whichever values of potential observations, their probability of becoming missing is the same (but may depend on Y_{obs})
- If one needs to consider *all possible* patterns of missingess (any components of Y missing), one talks about *everywhere MAR* (Seaman et al., 2013)

When is the observed-data likelihood ok?

- MAR: $p(R|Y_{obs}, Y_{mis}, \phi) = p(R|Y_{obs}, \phi)$ for all Y_{mis} and ϕ
- Under MAR, the observed-data likelihood is proportional to the likelihood based on the observed data *and* the missingness

$$
\begin{aligned}
p(Y_{obs}, R|\theta, \phi) &= \int p(R|Y_{obs}, Y_{mis}, \phi) p(Y_{obs}, Y_{mis}|\theta) dY_{mis} \\
&= p(R|Y_{obs}, \phi) \int p(Y_{obs}, Y_{mis}|\theta) dY_{mis} \\
&= constant \times \overbrace{p(Y_{obs}|\theta)}^{\text{observed-data likelihood}}
\end{aligned}
$$

- N.B. The above holds under realised MAR if Y_{obs} is fixed to its observed value

Ignorability and statistical paradigms

- Ignorability = inferences based on a parametric model of the observed data alone are the same as inferences from the joint model of complete data and missingness
- Treatment of missing values cannot be separated from the choice of the inferential framework
  - In particular, there is a line between likelihood-based inference vs. sampling distribution-based inference
  - The difference between whether only the realised values of the observed data or all possible observable data need to be considered

Ignorability and statistical paradigms cont.

- Ignorability applies under
  - realised MAR for direct likelihood or Bayesian inference
  - everywhere MAR for likelihood-based frequentist inference (Seaman et al., 2013)
  - MCAR for sampling-distribution based inference
- In addition, disctinct parameters (θ and ϕ) need to be assumed

Ignorability vs. intention of missingess

- Different study designs can be cross-tabulated according to whether missing data are ignorable (yes/no) and whether missingness is intentional by design (yes/no)

By design	Ignorable	
	Yes	No
Yes	• simple random/stratified sampling • randomised experiments	• type I censoring
No	• drop-out dependent on past history • treatment dependent on covariates	• all messy stuff

Dealing with the problem (even under MAR)

- There remains the problem of integrating over/imputing missing values
- Two major conceptual approaches of choice
  - ML estimation (EM algorithm, Dempster, 1977)
  - Bayesian modelling (data augmentation, MCMC)
- In practice, multiple imputation (Rubin, 1987) is often used as an approximate Bayesian approach
  - Use of a separate imputation model(s) to missing data items repeatedly, conditionally on each unit's observed data *and* population-level parameters
  - Convenient choice when the data are large, missing patterns unbalanced and/or entail covariates, or the imputed data need to be stored or shared
- Weighted complete case analysis is useful in some situations

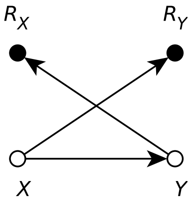
FURTHER TOPICS

Beyond MCAR, MAR and MNAR

- The missing data mechanism may vary variable by variable
- Separate missingness indicators are then needed for each variable
- Complicated missing data mechanism can be described by graphical models
- MNAR can be divided into subclasses
  - There exists situations where the data are MNAR but the joint distribution can be still estimated in a non-parametric form

Beyond MCAR, MAR and MNAR: an example

Mohan, Pearl & Tian (2013) provide the following example:



The causal structure implies $R_X \perp\!\!\!\perp (X, R_Y)|Y$ and $R_Y \perp\!\!\!\perp (Y, R_X)|X$. These independencies allow us to write

$$P(X,Y) = P(X,Y)\frac{P(R_X, R_Y|X,Y)}{P(R_X, R_Y|X,Y)} = \frac{P(R_X, R_Y)P(X,Y|R_X,R_Y)}{P(R_X|Y,R_Y)P(R_Y|X,R_X)}.$$

The resulting expression can be estimated by replacing probabilities by empirical probabilities calculated in subsets where X or Y or both of them are observed. The estimator is consistent but not fully efficient.

Topics not covered

- Analysis in practice
- Implementation and software
- Weighting methods, generalised estimation equations, etc.
- Consequences of incorrect assumptions
- Sensitivity analysis
- Missing data in causal inference

How to talk about missing data in Finnish?

Finnish	English
sivuutettavuus	ignorability
sivuutettava puuttuvuus	ignorable missingness
sivuuttamaton puuttuvuus	non-ignorable missingness
täysin satunnainen puuttuvuus	MCAR
satunnainen puuttuvuus	MAR
ei-satunnainen puuttuvuus	MNAR
realisoitunut täysin satunnainen puuttuvuus	realised MCAR
yleispätevästi täysin satunnainen puuttuvuus	everywhere MCAR
realisoitunut satunnainen puuttuvuus	realised MAR
yleispätevästi satunnainen puuttuvuus	everywhere MAR
suunnitellusti puuttuva	missing by design

Difficult to translate: proper/improper imputation, propensity score, propensity weighting

References

- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (methodological), 1–38.
- Little, R. J., & Rubin, D. B. (2002). Statistical analysis with missing data. 2nd edition, John Wiley & Sons.
- Mohan K., Pearl J., & Tian J. (2013). Graphical Models for Inference with Missing Data, In Proceedings of Advances in Neural Information Processing System 26 (NIPS-2013), 1277–1285.
- J. Pearl J. & Mohan K. (2104). Recoverability and testability of missing data: Introduction and summary of results, UCLA Cognitive Systems Laboratory, Technical Report (R-417).
- Rubin, D. B. (1976). Inference and missing data. Biometrika, 581–592.
- Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. John Wiley & Sons.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. Psychological methods, 7(2), 147.
- Seaman, S., Galati, J., Jackson, D., & Carlin, J. (2013). What Is Meant by" Missing at Random"?. Statistical Science, 257–268.
- Van Buuren, S. (2012). Flexible imputation of missing data. CRC Press.