Dealing with non-response: Introduction to imputation

Aksel Thomsen & Peter Tibert Stoltze
Statistics Denmark

Dealing with non-response: Overview

- Defining non-response
- Assessment of non-response
- The problem with non-response
- Imputation as a way to deal with non-response
- Types of imputation
- Added uncertainty by applying imputation

A simple question ☺

- We have a sample of three units with observed values

$$y = \{1, 2, 6\}$$

- What is the sample mean?

- What if one of the values were missing?

$$y = \{1, 2, \mathrm{NA}\}$$

Definition of non-response

- Item non-response
  - Subject A has not responded to the survey
- Partial non-response
  - Subject B has responded to questions 1, 2, and 4 but not 3 and 5
- Central and non-central variables in a multivariate survey
  - All questions are important, but some are more important than other
- When does partial non-response become item non-response?
  - Perhaps the central questions are 1 and 2 and we are relatively happy – but it could also be that the real core information is question 3?

Assessment of non-response

- When given a data set from a survey, we need to have good understanding of *routing*
  - If subject answers XXX to question A, skip question B, and go directly to question C
- We also need to know how to interpret values in the data file, i.e. what value represents a missing value
  - We have seen many data files coded with either . or NA or even -9999

- Assume a data file containing the results of a questionnaire with *p* questions (partially) answered by *n* subjects
- The data are organized in a $n \times p$ data frame*

$$Y = \begin{bmatrix} Y_{11} & \cdots & Y_{1p} \\ \vdots & \ddots & \vdots \\ Y_{n1} & \cdots & Y_{np} \end{bmatrix}$$

- This means that the answer from subject *i* to question *j* is Y_{ij}
- We call this a *data frame* because the columns need not be of the same type – if all columns are numeric we have a matrix

- We now define a matrix R of same dimensions as Y to indicate non-response:

$$R_{ij} = \begin{cases} 1 & if\ Y_{ij} \in V_j \\ 0 & else \end{cases}$$

- Here we define the case of response to mean that a specific answer X_{ij} belongs to the V_j (the set of admissible values for question j)
- Hence, the matrix R becomes filled with 0's and 1's allowing us to start assessing the amount of non-response
- Note: This seems like a simple task, but it is not

Non-response in a sampling perspective

- When we do probabilistic sampling, we try to describe a population with *N* units through a sample with *n* units
- The mechanism leading from *N* to *n* is the sampling design, and we know the inclusion probability π
- For simple random sampling we have

$$\pi = \frac{n}{N}$$

  leading to the design weight

$$w = \frac{1}{\pi} = \frac{N}{n}$$

- With non-response we move from *n* sampled units to *m* responding units
- If the mechanism from sample to response is truly random, then we can estimate the response propensity as

$$\theta = \frac{m}{n}$$

- The combined probability of both being sampled and responding (assuming these are actually independent) is simply

$$p = \pi\theta = \frac{n}{N}\frac{m}{n} = \frac{m}{N}$$

Why is non-response a potential problem?

- However, θ is almost never a constant value, since response propensity almost always varies with our variables of interest
- In the 1970's Rubin suggested a more formal approach to this distinguishing between
  - Missing complete at random (MCAR) where in fact θ is the same for all cases
  - Missing (conditionally) at random (MAR) where θ can be constant within groups of observed data
  - Missing not at random (MNAR) where θ is individual
- Simple methods like Mean Imputation only works for MCAR, and they will give biased results for MAR and MNAR

Dealing with partial non-response

- Simplest solution: Dismiss all rows with partial non-response, i.e. reduce data set to *complete cases*
  - Simple but can potentially mean that too much information is lost
- A better solution is often to apply *imputation* to suitable variables
- Probably not all variables should be subject to imputation

Dealing with item non-response

- If a subject has not answered at all, we will only in rare circumstances do imputation
- Weighting is normally the best solution when no individual response is particularly important
- In the case of business statistics and large enterprises, it can make sense to do imputation
  - In this case we will resort to “expert imputation” which is a very manually driven process based on a combination of whatever knowledge is available – this process should only be applied to very few cases!

Types of imputation

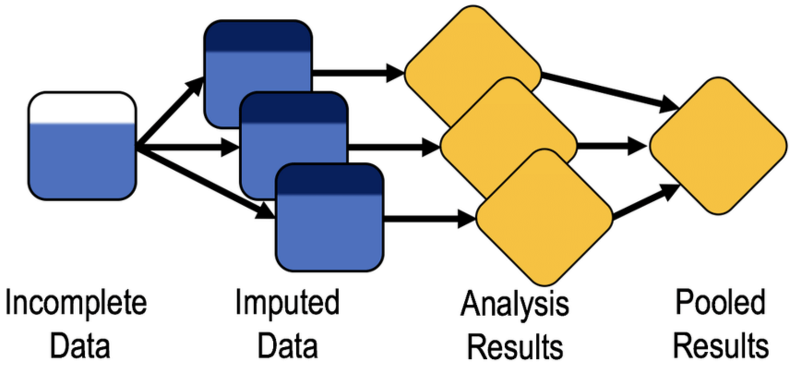
- Donor based or model based
  - Duplicate values from a neighbor or construct from statistical/ML model
- Stochastic or deterministic
  - Get exact same result each time or introduce a bit of variation
- Hot-deck or cold-deck
  - Using information from current round of survey or from previous ones

Examples of imputation methods

- Mean imputation
- Last Value Carried Forward
- Multiple Linear regression
- Multinomial logistic regression
- Random forest

The idea of multiple imputation

- When imputing we add a new source of uncertainty
- If we have used a stochastic method, we can make several versions of imputed data sets and measure the variation between these



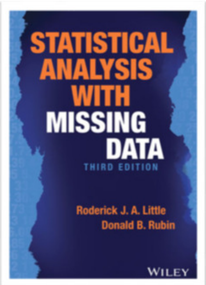
(Nissen, Donatello & Van Dusen, 2019)

Evaluating deterministic imputation

- If we have used a deterministic method it is possible to apply resampling techniques to evaluate the result
  - Jackknife: Basically leave out items one-by-one and evaluate
  - Bootstrap: Sampling with replacement from the incomplete data

Resources

- The standard text is Donald Rubin and Roderick Little: “Statistical Analysis with Missing Data” (3rd edition, Wiley 2019)



- Freely available is ”Flexible Imputation of Missing Data” by Stef van Buuren (also available as hardcopy at Chapman & Hall)
- https://stefvanbuuren.name/fimd/