

What should we do about missing data?

Application to NTDB Version 6.1
(and future NTDB Versions)

Outline of presentation

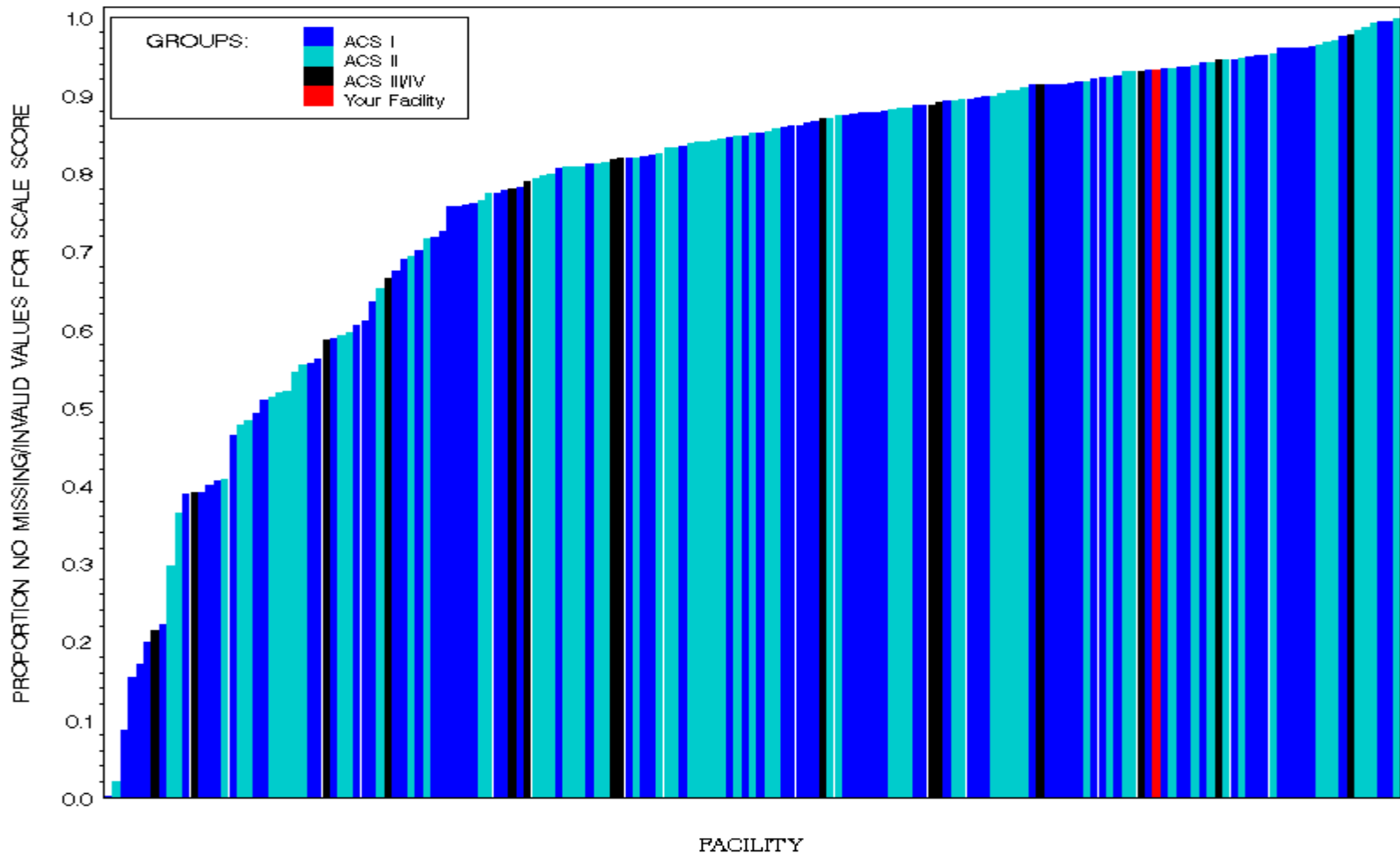
- Examples of causes of missing data
- Missing data in NTDB 6.1
- Problems resulting from missing data
- Methods of handling missing data
 - Complete Case and Available Case analyses
 - Rational Substitution
 - Single Imputation
 - Hot Deck Imputation
 - Multiple Imputation
- Examples of imputation used by other injury or surgical databases

Examples of Causes of Missing data

- Failure of measuring instrument
- Denial of access
 - Age > 89 (HIPAA)
- Data lost
- Data loading issues
- Out-of-range data (e.g., Age > 200)
- Data points physically impossible to obtain, (e.g. verbal GCS when they have an endotracheal tube and cannot talk.)

Missing data in NTDB

Proportion of no missing on 26 key fields per facility



Nature of missing data

- Missing Completely at Random (MCAR)
 - the probability that an observation (X_i) is missing is unrelated to the value of X_i or to the value of any other variables
- Missing at Random (MAR)
 - random if the data meet the requirement that missingness does not depend on the value of X_i *after controlling for another variable.*
- Not missing at Random (NMAR)

Missing Completely at Random (MCAR)

- Probability of an observation being missing does not depend on observed or unobserved measurements.
- $\Pr(\mathbf{r} \mid \mathbf{y}_o, \mathbf{y}_m) = \Pr(\mathbf{r})$
- Analysis of only those units with complete data gives valid inferences (unbiased).
- Example: Instrument malfunction, source data lost.

Missing at Random (MAR)

- Probability of an observation being missing does not depend on the unobserved data.
- $\Pr(\mathbf{r} \mid \mathbf{y}_o, \mathbf{y}_m) = \Pr(\mathbf{r} \mid \mathbf{y}_o)$.
- Example: Women less likely to reveal weight. That is, the probability of missing depends on gender and does not depend of weight itself.
- More realistic assumption, estimations can be biased

Not Missing at Random (NMAR)

- Probability of an observation being missing *depends on the unseen observations themselves.*
- Pattern is non-random, non-ignorable, and arises due to the variable in which data is missing
- To obtain valid inference, a *joint model* of the data and the missingness mechanism is required.
- Example: overweight will not give their weight.. So the missingness of weight depends on weight.

Determining the Nature of Missing Data

- Quite difficult when data is missing on a number of variables.
- Determine if probability of an observation is missing is associated with values of the variables that are missing the data.
- Sensitivity analysis are important!!

Problems Resulting From Missing Data

- Loss of information
- Bias
- Loss of power

Methods of handling missing data

1. Rational Substitution
2. Complete Case and Available Case
3. Single Impute
4. Hot Deck
5. Multiple Imputation

Rational Substitution

- Impute “known” values:
 - GCS is 3, so the Motor GCS must be 1
 - Patient is DOA, so the BP must be 0
- + Simple
- + Easy to understand
- Requires some Assumptions

Complete Case and Available Case

CC: Delete observations with any missing data.

AC: Delete observations with missing data for specific variables of interest.

- + Simple, usually the default in stat. software
- Potential loss of information and precision
- Bias introduced when observation is not MCAR
- Analysis Sample changes from variables of interest when doing AC

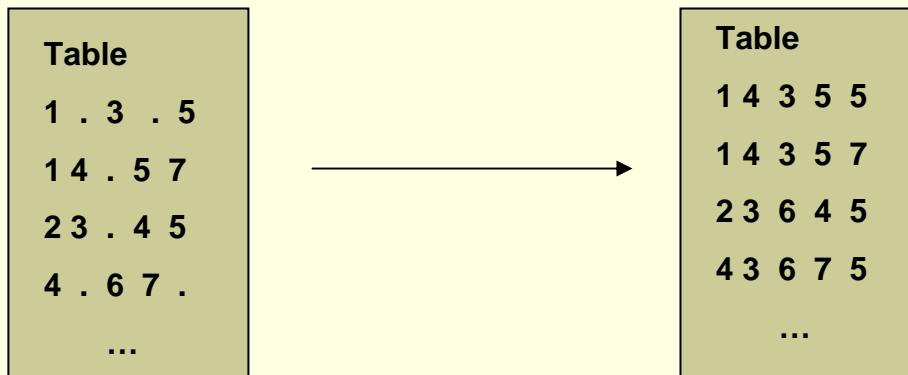
Single Imputation

Impute a single value, i.e. mean, baseline value...

- + Simple
- + Reduces Bias and improves precision.
- Underestimate the standard error and incorrect p-values
- Only provides unbiased estimates for means and totals if the missing values are MCAR.

Hot-Deck Imputation

Variables for good records in the current (hot) survey file are used to impute for blank values of incomplete records.

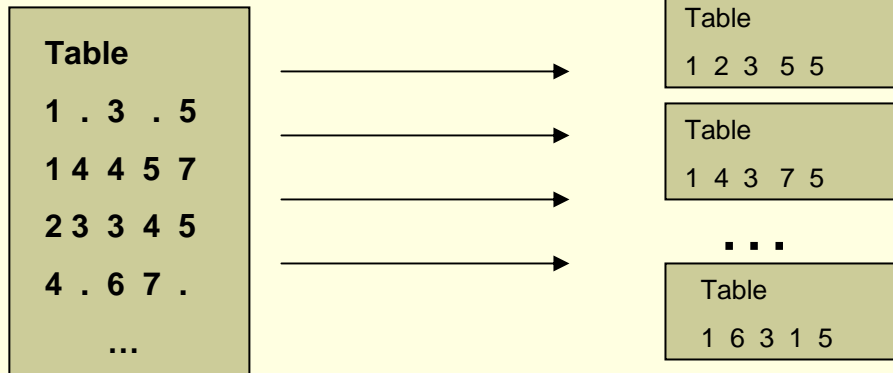


Backward-Forward approach “good” value previously not used

Hot-Deck Imputation

- + Simple theory
- + Preserves the distribution of the estimates, and increases the variance relative to the mean imputation method. Thus, the underestimation of the variance of the estimate is decreased.
- Time consuming when many variables are missing data.
- Assumes that missing cases are representative of the population that shares key predictor characteristics.

Multiple Imputation



1) $m > 1$ data sets with plausible* data replace the missing values

2) Each of the m data sets is analyzed

3) The m estimates are then combined for inference

* Regression Method, Propensity score method, MCMC method

Multiple Imputation

- + single set of imputed data sets can be used for a variety of analyses
- + accounts for missing data uncertainty
- cumbersome to use because of the need to analyze multiple data sets and combine the results to make one overall inference.
- Assumption: missing values are MAR.