

Should We Fill the Holes?

N. Clay Mann, PhD, MS



Why is Imputation Important?

- Missing data are common.
- Usually inadequately handled in both epidemiological and experimental research.
- For example, **Wood *et al.* (2004)** reviewed 71 BMJ, JAMA, Lancet and NEJM papers.
 - 89% had partly missing outcome data.
 - In 37 trials, 46% performed complete case analysis.
 - Only 21% reported sensitivity analysis.



Just Observations

- NTDB – Two Sources of “missingness”
 - Missing cases (incomplete sample)
 - Missing data
- Mixed reviews by reviewers
- Estimations (mean) vs inferences (CI)
- Outcomes or predictors



Just Opinion

- Would a priori imputation be helpful?
 - Not to a grand degree
 - Simple stochastic imputation – perhaps some
 - Multiple imputation not really workable without model building
 - Education (the “black box”) is the bigger issue....but almost impossible to resolve



Types of Missingness: (MCAR)

- Missing completely at random (MCAR)
 - The probability that a value is missing is unrelated to any value, missing or observed
- NTDB Example
 - A hospital abstractor fails to find lab results for patients admitted last Monday because a tech dropped a tray of test tubes.
- Resulting pattern known to NTDB? - NO
- Least Plausible – yet often assumed
 - Casewise deletion
 - Pairwise deletion – NEVER!



Types of Missingness: (MAR)

- Missing at random (MAR)
 - The probability that a value is missing is related only to observed values
- NTDB Example
 - RR more likely missing among the severely injured
- Resulting pattern known to NTDB? – assumed discoverable
- Required for single and multiple imputation



Types of Missingness: NMAR

- Not missing at random (NMAR)
 - The probability that a value is missing depends on missing values
- NTDB Example
 - Ethnicity less likely reported among Whites
- Perhaps not common in NTDB: big trouble
 - I have never seen the Heckman correction applied to health-related data



Common Types of Imputation

- Assumes (MAR)
 - Single imputation
 - Mean imputation
 - Substitution
 - Regression imputation
 - Hot deck imputation
 - Last observation carry-over (longitudinal studies)
 - Worst case analysis
 - Missing indicator method – NEVER!



Single Imputation Approaches

- All approaches:
 - Underestimate standard error
 - Increase likelihood of a type-I error
 - Likely biased results



Just Opinion

- Single Imputation

- Proportion of missing $< 5\%$ - casewise deletion
- Proportion of missing 5-10% - single imputation
- Proportion of missing 15-20% - multiple imputation....but how for NTDB?

- **HOWEVER**

- Reason for “missingness” and purpose of the study is much more important than number of missing values....neither is easily known to NTDB



Just Opinion

- If we Provide Imputed Sample:
 - Increasing likelihood of a type-I error (and increased R^2)in the presence of large samples
 - Ensuring agreement among peers on method.
 - Not understanding the purpose of the study.



Multiple Imputation

- Multiple imputation strategies involve three distinct phases:
 1. The missing data are filled in m times to generate m complete data sets.
 2. The m complete data sets are analyzed using standard statistical analyses.
 3. The results from the m complete data sets are combined to produce inferential results (ie, Rubin)

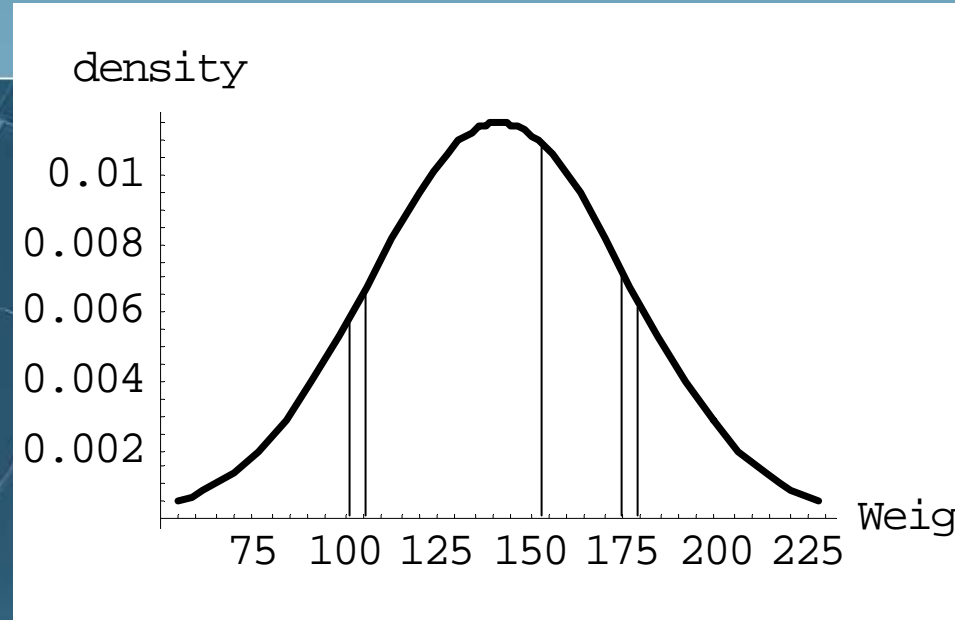


Multiple Imputation

- Methods estimate standard errors and CI that incorporate the variance of imputed values
- Not intended to impute individual values
- Draw inferences (ie, point estimates, CI) for the aggregate sample that approximates true values

Alternatives to MI: Full Information Maximum Likelihood

- If X is missing
 - MI draws a few random values from distribution
- FIML
 - Integrates across the full distribution of possible values
 - Like MI with an infinite number of imputations





Just Opinion

- If we use Multiple Imputation
 - It's a team sport
 - Imputer (study question, sample stratification)
 - Analyst (interaction terms, variable selection)
 - Doesn't provide a static sample
 - Provides parameters for a specific sample and model
 - What to code missing
 - What about non-sensible values?



Questions

