Missing Data: II

Why is data missing?

Not at random (NMAR)

- Whether or not missing correlates to an unknown value.

At random (MAR)

- Missing or not correlates to a known variable.

Completely at random (MCAR)

- No reason for missing data; any entry is equally likely to be missing.

Detecting MAR

- Quick check: split data into two groups: one where a given field is missing some data, one where it is not.
  - Are any of the summary statistics significantly different?
- We can replace a field with missing data with a “dummy” indicator variable (1 if the data is present; 0 if not).
  - Are the other columns predictive of the value of the indicator variable (by any statistical method)?

Little's Test for MCAR

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association*, 83(404), 1198-1202.

Defines a test statistic (asymptotically chi-squared) one can use to assess probability data is missing completely at random.

No widely used Python package, but available in R and SPSS.

Missing data

- Details may vary, but generally if fields are missing in a row of data we can:
  - Ignore it,
  - Drop the row,
  - Drop the field, or
  - “Guess” a value (imputation).
- It may also be the case that a row exists but has no data.

Dong and Peng *SpringerPlus* 2013, **2**:222
http://www.springerplus.com/content/2/1/222

SpringerPlus
a SpringerOpen Journal

METHODOLOGY Open Access

Principled missing data methods for researchers

Yiran Dong and Chao-Ying Joanne Peng*

Abstract

The impact of missing data on quantitative research can be serious, leading to biased estimates of parameters, loss of information, decreased statistical power, increased standard errors, and weakened generalizability of findings. In this paper, we discussed and demonstrated three principled missing data methods: multiple imputation, full information maximum likelihood, and expectation-maximization algorithm, applied to a real-world data set. Results were contrasted with those obtained from the complete data set and from the listwise deletion method. The relative merits of each method are noted, along with common features they share. The paper concludes with an emphasis on the importance of statistical assumptions, and recommendations for researchers. Quality of research will be enhanced if (a) researchers explicitly acknowledge missing data problems and the conditions under which they occurred, (b) principled methods are employed to handle missing data, and (c) the appropriate treatment of missing data is incorporated into review standards of manuscripts submitted for publication.

Keywords: Missing data, Listwise deletion, MI, FIML, EM, MAR, MCAR, MNAR

Dropping row

- **Listwise deletion** removes every case with any missing values from all analyses.
- **Pairwise deletion** removes cases only when the missing data is needed for a particular analysis.
- Decreases statistical power (since decreases effective sample size).
- If the data is anything other than missing completely at random, this biases your data set.

Dropping fields

- If a field has a high percentage of missing data, it is probably best to just drop it.
  - e.g. `diabetes.drop('SkinThickness')`
- This is mostly harmless, although it reduces the amount of data available to base a prediction on.

Deductive imputation

- Sometimes the value of a missing field is uniquely determined by the previous fields.
- For example, if a patient's age is less than 5, then the patient is not currently pregnant.
- This is always safe, as long as your rules are correct.

- **Don’t** replace missing data with mean / median / mode / etc.

- Why not?
  - Mean value may be quite rare or non-existent.
  - Distorts histogram.
  - Reduces variance.
  - Does not incorporate correlations in the data.
  - These alternatives can give different results.

Bad plan: replacing missing values with summary statistic

- **Don't** replace missing data with regression line predictions.
  - $estimate = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$

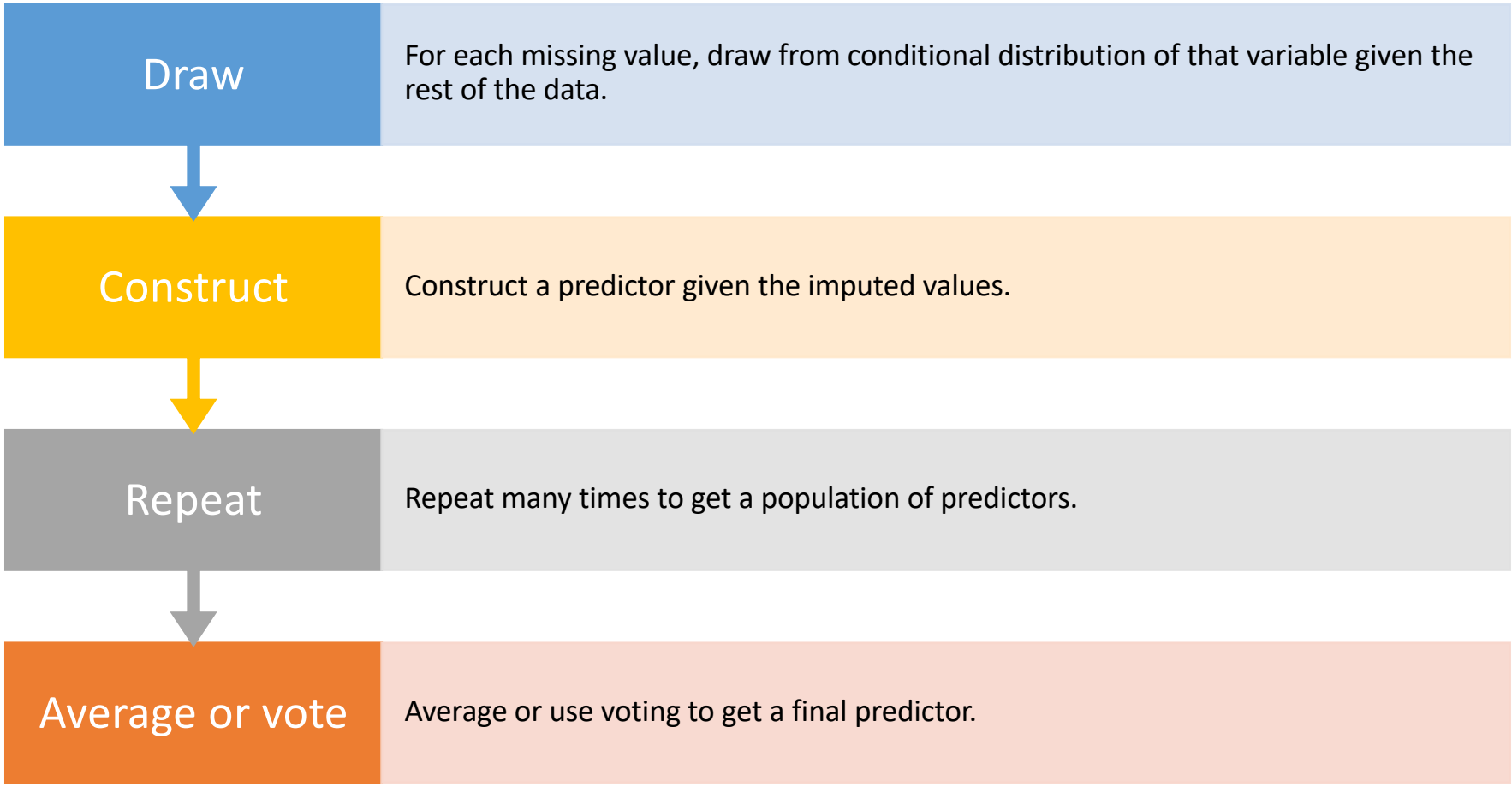
- Why not?
  - Distorts histogram.
  - Reduces variance.

Bad plan: using deterministic regression for missing values

- **Don't** replace missing data with single stochastic regression predictions.
  - $estimate = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon_i$
  - $\varepsilon_i \sim N(0, s)$

- Why not?
  - Reduces variance.

Bad plan: using stochastic regression for missing values



Multiple imputation

General strategy developed by Rubin, 1988.

FIML and EM

- Full information maximum likelihood (FIML)
- Expectation maximization (EM)

- Avoid imputing values.
- Find measures that are most consistent with the observations.