Missing Data

Missing Data

- Missing data is inevitable
- People drop out of studies
- People skip questions
- People mess up their answers to questions
- Samples get ruined
- Lab tests have errors, are below the threshold of detection, etc.

Three “Types” of Missingness

- Missing *Completely* at Random
- Missing at Random
- Missing Not at Random

Missing Completely at Random

- The probability of a data value being missing is independent of all other other data (observed and unobserved)
- Missing data is a random sample of your data
- Common Examples:
  - Low-level data corruption (a bit of memory gets hit by stellar radiation – yes this happens rarely)
  - The power goes out while running a set of samples through a test
  - Someone spills coffee on some paper records

Missing at Random

- The probability of a data value being missing is dependent only on observed data
- *Conditional on all other variables* a missing data point is random
- Examples:
  - Healthcare workers are more likely to have detailed case histories
  - Eligible patients with advanced disease less likely to have complete data
    - May miss appointments, have medical contraindications for some tests, etc.

Missing Not at Random

- The probability of a data value being missing is dependent on some unknown factor
- Examples
  - Low values may be harder to detect by an assay
  - The probability of the result of the assay being missing is related to the value of the assay – which cannot be observed

Dealing with Missing Data

- Complete case analysis
  - Only analyze the data sets with complete data
  - This is the default in many/most software packages
- This can dramatically reduce your sample size
- This is only valid if your missing data is MCAR
- That's a really strong assumption

Indicator Variables

- Assigning missing values an indicator that they're missing, and then treat that as a value of the variable
  - i.e. Yes/No/Missing
- This has been extensively studied and is a very bad idea that will produce biased values

Imputation

- Assign values to the missing variables
- Believe it or not, this is more conservative than complete case analysis
- Mean Imputation: All missing values get the mean value of the non-missing data
- Single Imputation: Build a model to predict the value of missing variables, each missing value takes a predicted value
- Other methods to decide values

Multiple Imputation

- Create many versions of the data
- Build a model to predict the value of missing variables, each missing data is probabilistically assigned
- Example:
  - 40% chance of a binary variable being 1 results in 4 data sets with a value of 1 and 6 with a value of zero
- Run the model on each data set, pool the results
- Details in R.A. Little and D.B. Rubin. 1989. The Analysis of Social Science Data with Missing Values. *Sociological Methods and Research*.

Typical Assumptions

- Missing at Random
- Multivariate normal distribution for missing data
- More elaborate models are possible, even for MNAR
  - This involves modeling the process by which the data is generated