Pre-Processing & Item Analysis
DeShon - 2007

Pre-Processing
- Method of Pre-processing depends on the type of measurement instrument used
- General Issues
  - Responses within range?
  - Missing data
  - Item directionality
- Scoring
  - Transforming responses into numbers that are useful for the desired inference

General Issues
- Responses within range?
- Missing data
- Item directionality
- Scoring
  - Transforming responses into numbers that are useful for the desired inference

Responses within range?
- Responses should be within the range of your measure.

Checking response range
- First step...
- Make sure there are no observations outside the range of your measure.
- If you use a 1-5 response measure, you can't have a response of 6.
- Histograms and summary statistics (min, max)

Reverse Scoring
- Used when combining multiple measures (e.g., items) into a composite
- All items should refer to the target trait in the same direction
- Alg: (high scale score +1) – score

Missing Data
- Huge issue in most behavioral research!
- Key issues:
  - Why is the data missing?
    - Random, missing randomly, response bias?
  - What's the best analytic strategy with missing data?
  - Statistical Power
  - Biased results

Causes of Missing Data
- Common in social research
  - Nonresponse, loss to followup
  - Lack of overlap between linked data sets
  - Social processes
  - Dropping out of school, graduation, etc.
  - Survey design
    - "skip patterns" between respondents

Huge issue in most behavioral research!
- Key issues:
  - Why is the data missing?
    - Random, missing randomly, response bias?
  - What's the best analytic strategy with missing data?
  - Statistical Power
  - Biased results
Missing Data

- Step 1: Do everything ethically feasible to avoid missing data during data collection
- Step 2: Do everything ethically possible to recover missing data
- Step 3: Examine amount and patterns of missing data
- Step 4: Use statistical models and methods that replace missing data or are unaffected by missing data

Missing Data Mechanisms

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Not Missing at Random (NMAR)

Example:
- X \(\rightarrow\) not subject to nonresponse (age)
- Y \(\rightarrow\) subject to nonresponse (income)

Missing Completely at Random

- MCAR
- Probability of response is independent of X \& Y
  - Ex: Probability that income is recorded is the same for all individuals regardless of age or income

Missing at Random

- MAR
- Probability of response is dependent on X but not Y
- Probability of missingness does not depend on unobserved information
  - Ex: Probability that income is recorded varies according to age but it is not related to income within a particular age group

Not Missing at Random

- NMAR
- Probability of missingness does depend on unobserved information
  - Ex: Probability that income is recorded varies according to income and possibly age

How can you tell?

- Look for patterns
- Run a logistic regression with your IV's predicting a dichotomous variable (1=missing; 0=nonmissing)
Missing Data Mechanisms

- If MAR or MCAR, the missing data mechanism is ignorable for full information likelihood-based inferences.
- If MCAR, the mechanism is also ignorable for sampling-based inferences (OILS regression).
- If NMAR, the mechanism is nonignorable – thus any statistic could be biased.

Missing Data Methods

- **Always Bad Methods**
  - Listwise deletion
  - Pairwise deletion (a.k.a. available case analysis)
  - Person or item mean replacement

- **Often Good Methods**
  - Regression replacement
  - Full-Information Maximum Likelihood (FIML)
  - SEM – must have full dataset
  - Multiple Imputation

Listwise Deletion

- Assumes that the data are MCAR.
- Only appropriate for small amounts of missing data.
- Can lower power substantially.
- Now very rare.
- Don’t do it!

Pairwise Deletion

- Use all available information.
- Can result in impossible results that violate the triangle inequality for correlations.
- More likely to occur for analyses with larger numbers of variables.
- Non-positive definite errors.

Mean Substitution/imputation

- Technique:
  - Calculate mean over cases that have values for $Y$.
  - Impute this mean where $Y$ is missing.
  - Data for $X$, $Y$, etc.

- Implicit models:
  - $Y = \mu_Y$, $X_1 = \mu_{X_1}$, $X_2 = \mu_{X_2}$

- Problems:
  - Ignores relationships among $X$ and $Y$.
  - Underestimates variances.
  - Artificially reduces variance and standard errors.

FIML - AMOS
Imputation-based Procedures

- Missing values are filled-in and the resulting "Completed" data are analyzed
- Hot deck
- Mean imputation
- Regression imputation
- Some imputation procedures (e.g., Rubin's multiple imputation) are really model-based procedures.

Regression Imputation

- Technique & implicit models
- If Y is missing
  \[ \hat{Y} = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} \]
- Likewise, if \( X_{11} \) is missing
  \[ \hat{X}_{11} = \gamma_0 + \gamma_1 Y + \gamma_2 Y \]
- If both Y and \( X_{11} \) are missing
  \[ \hat{Y} = \delta_0 + \delta_1 X_{11} \]

Little and Rubin's Principles

- Imputations should be
  - Conditioned on observed variables
  - Multivariate
  - Draws from a predictive distribution

Multiple Imputation

- Context: Multiple regression (in general)
- Missing values are replaced with "plausible" substitutes based on distributions or model
- Construct m simulated versions
- Analyze each of the m simulated complete datasets by standard methods
- Combine the m estimates
- get confidence intervals using Rubin's rules

Multiple Imputation (Rubin, 1987, 1996)

- Point estimate
- Variance
- Variance within = Variance Imputation
- Variance between = 

Another View

- IMPUTATION ANALYSIS POOLING
- IMPUTED DATA
- ANALYSIS RESULTS
- FINAL RESULTS

- IMPUTATION: Impute the missing entries of the incomplete data sets M times, resulting in M complete data sets.
- ANALYSIS: Analyze each of the M completed data sets using weighted least squares.
- POOLING: Integrate the M analysis results into a final result. Simple rules exist for combining the M analyses.
How many Imputations?

- 5
- Efficiency of an estimate: \((1+\gamma/m)^2\)
  - \(\gamma\) = percentage of missing info
  - \(m\) = number of imputations
- If 30% missing, 3 imputations \(\rightarrow 91\%\)
- 5 imputations \(\rightarrow 94\%\)
- 10 imputations \(\rightarrow 97\%\)

Imputation in SAS

- By default generates 5 imputation values for each missing value
- Imputation method: MCMC (Markov Chain Monte Carlo)
- EM algorithm determines initial values
- MCMC repeatedly simulates the distribution of interest from which the imputed values are drawn
- Assumption: Data follows multivariate normal distribution

Example

- Case 1 is missing weight
  - Given 1’s sex and age
  - generate a plausible distribution for 1’s weight
  - At random, sample 5 (or more) plausible weights for case 1
  - Impute Y!
  - For case 6, sample from conditional distribution of age.
  - Use Y to impute X!
  - For case 7, sample from conditional bivariate distribution of age & weight

Example – Standard Errors

- Total variance in \(b\)
  - Variation due to sampling + variation due to imputation
  - \(\text{Mean}(\text{Var}(b)) = \text{Var}(b) + \text{Var}(\text{Imputation})\)
  - Actually, there’s a correction factor of \((1+1/M)\)
  - For the number of imputations \(M\) (Here \(M=5\))
  - So total variance in estimating \(b\)
    - \(\text{Mean}(\text{Var}(b)) = \text{Var}(b) + (1+1/5)\text{Var}(\text{Imputation})\)
    - \(= 179.53 + (1.2) 511.59 = 793.44\)
    - Standard error is \(28.17\)

Example

- PROC MI:
  - DATA=missing_weight_age
  - OUT=weight_age_mi
  - VAR years_over_20 weight maleness;
- run;

Other Software

- www.stat.psu.edu/~jls/misoftwa.html
Item Analysis

- Relevant for tests with a right / wrong answer
  - Score the item so that 1=right and 0=wrong
- Where do the answers come from?
  - Rational analysis
  - Empirical keying

Goal of Item Analysis

- Determine the extent to which the item is useful in differentiating individuals with respect to the focal construct
- Improve the measure for future administrations

Typology of Item Analysis

- Classical Item Response theory
  - Rasch
  - IRT2
  - IRT3

Classical Item Analysis

- Classical analysis is the easiest and most widely used form of analysis
- The statistics can be computed by generic statistical packages (or by hand) and need no specialist software
- The item statistics apply only to that group of testees on that collection of items
- Sample Dependent!

- Item Difficulty
  - Proportion Correct (1=correct; 0=wrong)
  - The higher the proportion the easier the item
  - In general, need a wide range of item difficulties to cover the range of the trait being assessed
  - If mastery test, need item difficulties to cluster around the cut score
  - Very easy (0.0) or very hard items (1.0) are useless
  - Most variance at p=.5
Classical Item Analysis

- **Item Discrimination – 2 methods**
  - Difference in proportion correct between high and low test score groups (27%)
  - Item-total correlation (output in Cronbach’s alpha routines)
    - No negative discriminators
    - Check key or drop item
    - Zero discriminators are not useful
    - Item difficulty and discrimination are interdependent!

<table>
<thead>
<tr>
<th>Item</th>
<th>Correct</th>
<th>Item Diff</th>
<th>Item-total r</th>
<th>Item-Disc</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1</td>
<td>4</td>
<td>0.85</td>
<td>0.30</td>
<td>0.23</td>
</tr>
<tr>
<td>I2</td>
<td>2</td>
<td>0.96</td>
<td>0.28</td>
<td>0.07</td>
</tr>
<tr>
<td>I3</td>
<td>1</td>
<td>0.98</td>
<td>0.28</td>
<td>0.03</td>
</tr>
<tr>
<td>I4</td>
<td>3</td>
<td>0.78</td>
<td>0.45</td>
<td>0.30</td>
</tr>
<tr>
<td>I5</td>
<td>4</td>
<td>0.88</td>
<td>0.17</td>
<td>0.26</td>
</tr>
<tr>
<td>I6</td>
<td>2</td>
<td>0.81</td>
<td>0.30</td>
<td>0.32</td>
</tr>
<tr>
<td>I7</td>
<td>1</td>
<td>0.92</td>
<td>0.46</td>
<td>0.41</td>
</tr>
<tr>
<td>I8</td>
<td>2</td>
<td>0.87</td>
<td>0.44</td>
<td>0.29</td>
</tr>
<tr>
<td>I9</td>
<td>3</td>
<td>0.91</td>
<td>0.19</td>
<td>0.22</td>
</tr>
<tr>
<td>I10</td>
<td>3</td>
<td>0.89</td>
<td>0.18</td>
<td>0.21</td>
</tr>
<tr>
<td>I11</td>
<td>3</td>
<td>0.90</td>
<td>0.18</td>
<td>0.11</td>
</tr>
<tr>
<td>I12</td>
<td>2</td>
<td>0.96</td>
<td>0.34</td>
<td>0.30</td>
</tr>
<tr>
<td>I13</td>
<td>2</td>
<td>0.96</td>
<td>0.34</td>
<td>0.24</td>
</tr>
<tr>
<td>I14</td>
<td>2</td>
<td>0.94</td>
<td>0.32</td>
<td>0.34</td>
</tr>
<tr>
<td>I15</td>
<td>3</td>
<td>0.92</td>
<td>0.32</td>
<td>0.32</td>
</tr>
<tr>
<td>I16</td>
<td>2</td>
<td>0.95</td>
<td>0.24</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Classical Item Analysis