Creating Scale Scores, Dealing with Missing Data, and Identifying Outliers

Scale Scores

- Researchers will often use scale (or composite) scores
  - Example: Rosenberg Self-Esteem Scale (Rosenberg, 1965)

1……2……3……4……5
Strongly Disagree Strongly Agree

1. I feel that I'm a person of worth, at least on an equal plane with others.
2. I feel that I have a number of good qualities.
3. All in all, I am inclined to feel that I am a failure. (R)
4. I am able to do things as well as most other people.
5. I feel I do not have much to be proud of. (R)
6. I take a positive attitude toward myself.
7. On the whole, I am satisfied with myself.
8. I wish I could have more respect for myself. (R)
9. I certainly feel useless at times. (R)
10. At times, I think I am no good at all. (R)

Scale Scores

- The goal of a scale score is to create a single numeric value that captures an underlying construct
- In the case of the Rosenberg Self-Esteem Scale, researchers want to create a single score that reflects global self-esteem
  - More specifically, researchers want higher scores on the Rosenberg Self-Esteem Scale to reflect higher levels of global self-esteem
- One problem is that half of the items on the Rosenberg Self-Esteem Scale are written such that higher scores are indicative of lower self-esteem

Reverse-Scoring Items

- The goal of reverse-scoring is to make it so that all of the items constituting a scale are scored in the same direction
  - For example, high scores for each item on the Rosenberg Self-Esteem Scale should indicate high levels of self-esteem
- You should ALWAYS create a new variable when reverse-scoring an item
  - Do not simply replace the old score with the new reverse-score because this can lead to confusion (e.g., "Have I already reverse-scored item 8?")
- You should ALWAYS create scale scores using a syntax file and save that file along with your data file

Syntax for Reverse-Scoring Items

- Option 1 (recode)

```
recode rses3 rses5 rses8 rses9 rses10 (1=5) (2=4) (3=3) (4=2) (5=1) INTO
rses3r rses5r rses8r rses9r rses10r.
execute.
```

- Option 2 (subtract value from highest possible value +1)

```
compute rses3r = 6-rses3.
compute rses5r = 6-rses5.
compute rses8r = 6-rses8.
compute rses9r = 6-rses9.
compute rses10r = 6-rses10.
execute.
```

Calculating Scale Score

- The most common scale score is the average of the constituent items
  - A simple sum of the item scores is also relatively common
  - You should follow what is typically used for that particular type of score (e.g., the score for the Rosenberg Self-Esteem Scale is almost always presented as an average of the items whereas the Beck Depression Inventory is almost always presented as the sum of the items)
- The average is a simple linear transformation of the sum (i.e., it is the sum divided by the number of items) so these two scores will behave in similar ways
  - Correlation of +1.0
  - Shape of score distribution will be identical (i.e., same skewness and kurtosis)
  - BUT they will have different means and standard deviations
    - Mean and standard deviation for sum score will be a multiple of these values for the average score (if 10 items, then the descriptive statistics for the sum score will be 10 times the value of the average score)

Syntax for Calculating Scale Score

- Computing an average score

```
compute rses_tot=mean(rses1,rses2,rses3r,rses4,rses5r,rses6,rses7,
rses8r,rses9r,rses10r).
execute.
```

--One advantage of an average score is that it places this score in the original units of measurement which makes it easier to interpret

- Computing a sum score

```
compute rses_tot=sum(rses1,rses2,rses3r,rses4,rses5r,rses6,rses7,
rses8r,rses9r,rses10r).
execute.
```

- Both of these approaches will calculate scores even if there is missing data
  - "mean" command will adjust the denominator to correct for missing data

Missing Data

- This is extremely common in research
  - Examples:
    - Participants intentionally not completing some invasive items on self-report measures
    - Participant accidentally skipping an item
    - Equipment malfunctioning
    - Data entry errors
- Three patterns of missing data
  - Missing Completely at Random (MCAR)
  - Missing at Random (MAR)
  - Missing Not at Random (MNAR)

Missing Data ("?" denotes missing value)

	Variables							
Unit	1	2	3	4	5	6	7	.....
1	1	4	1	3.4	5.67	A	8.251	.....
2	1	3	?	?	5.67	B	9.253	.....
3	1	2	1	2.7	5.72	B	12.812	.....
4	1	1	1	3.6	5.13	?	13.614	.....
5	2	?	1	?	?	A	11.4422	.....
6	2	2	1	3.4	5.61	A	9.241	.....
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

Missing Data ("0" denotes missing value)

	Variables						
Unit	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	1	0	0	1	1	1
3	1	1	1	1	1	1	1
4	1	1	1	1	1	0	1
5	1	0	1	0	0	1	1
6	1	1	1	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Missing Completely at Random (MCAR)

- Missing Completely at Random (MCAR): the probability that an observation is missing is unrelated to a particular response to that item or any other variables
  - Any piece of data is just as likely to be missing as any other piece of data
  - Data on family income would be MCAR if missing values had nothing to do with the family incomes of those individuals or any other features possessed by those individuals (e.g., gender, self-esteem)
  - It is relatively easy to deal with this sort of missing data!

Missing at Random (MAR)

- Missing at Random (MAR): the probability that an observation is missing does not depend on its value after controlling for another variable
  - Individuals with low self-esteem may be less likely to report their income. This means that the missing observations are due to another variable (self-esteem) rather than the value that is missing (family income)
  - "Missing at Random" is a bad name for this sort of pattern because the missing data is NOT random because it is conditional on another variable
  - There are ways to deal with this sort of missing data

Missing Not at Random (MNAR)

- Missing Not at Random (MNAR): the probability that an observation is missing depends on its value
  - Data on family income would be MNAR if those with lower family incomes were less likely to report their incomes than those with higher incomes
  - This is the most difficult sort of missing data to manage

Approaches to Dealing with Missing Data

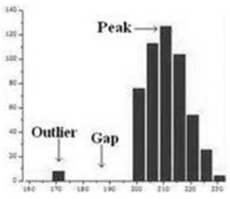
- Listwise deletion: omit cases with missing data and run analyses on remaining cases
  - Most common approach
  - If MCAR, then this approach does not lead to bias
  - If not MCAR, then it will lead to bias
  - Leads to a loss in power
- Pairwise deletion: use all available data which may lead to different numbers of participants for variables
  - Leads to problems if the data is not MCAR
- Single imputation: substitute missing values
  - A common approach
  - Use overall mean, scale mean, subgroup mean, or regression estimate for non-MD cases
- Multiple imputation: using regression to produce multiple estimates for the missing data that are averaged together

Identifying Outliers

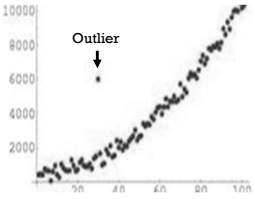
- Real data is often messy
  - Textbook examples generally deal with normally distributed data
- Outliers are a relatively common problem
- Univariate outlier: an observation that is very different (i.e., much smaller or much larger) than the other observations
- Multivariate outlier: an unusual combination of values for two or more variables

Types of Outliers

Univariate Outlier



Bivariate Outlier



Sources of Outliers

- Data entry errors
  - Can sometimes be identified because the values are implausible (e.g., a 725°F day in July)
- Unusual events
  - Correct values that are simply rare (e.g., a 45°F day in July)
- Participants provide incorrect or random data
  - Participant claims to have an extremely high score for depressive symptoms because they were just giving the highest score for each item without actually reading the items

Outliers and Statistics

- Influence of outliers on statistics
  - Bias or distort estimates
  - Distort statistical significance
  - Lead to faulty conclusions
- Identifying outliers
  - Visually inspect histograms (univariate outliers) and scatterplots (bivariate outliers)
  - Box-and-whisker plots can be very helpful
    - Uses interquartile range (IQR)
  - More than three standard deviations away from the mean