Overview

- Introduction to SACS
- What do we mean by “Data Cleaning” and why do we do it?
- The SACS data cleaning procedure
  - Types of missingness
  - Why we care
- Missing data analysis procedure
- What do we do when data are missing?
- An introduction to weighting
Student-Athlete Climate Study

- National, anonymous, population survey of student-athletes' perceptions and experiences pertaining to campus climate and its impact on student-athletes:
  - Academic success
  - Athletic success
  - Athletic identity

Almost 9,000 student-athletes from all three divisions and all NCAA sports

Focuses on the voices of those who have traditionally not been heard
What do we mean by “Data Cleaning” and why do we do it?

- Find and eliminate data entry and other errors
- Examine missing data and perhaps account for it in some way (statistically)
- Prepare the data for analysis
General data cleaning process

1. Define and determine error types
2. Search for and identify error instances
3. Correct errors
4. Document error types and instances
5. Modify data entry process to reduce future errors
The SACS data cleaning procedure

1. Check for and delete duplicate data entries (use SPSS “Identify Duplicate Cases” procedure or “Data Preparation” module).

2. Perform descriptive statistics to see if the data make sense. (e.g., Do the max and min values fall within the question’s expected range? Does the mean “make sense” for that question?)
   1. Frequency tables for categorical variables
   2. Histograms for Likert/ordinal/continuous variables
   3. Scatterplots for write-in continuous variables (e.g. age) to look for outliers and nonsense values

   ✓ Use frequency tables & histograms to examine the normality of distributions, the range, and extreme skew and kurtosis.
The SACS data cleaning procedure

3. Add to the list of new variables created for analysis purposes, including dummy variables (Appendix A).
   ✓ Also consider combining variables. There may be sets of questions that could be combined into a metric beyond the factor scales already identified (Appendix B).

4. Make sure dichotomous variables are coded with 1=the category of interest and 0=not.

5. Fix variable & value labels (e.g., typos, wrong scale).

6. Make sure variables are of the appropriate type (numerical/string) and variable measures are of the right level (SPSS language: scale/ordinal/nominal).
The SACS data cleaning procedure

7. Examine write-in text for “Other” responses and recode into already-existing categories.

8. Recode “Institution” write-in values to IPEDS institutional codes (ID variable for merging institutional data).

9. Clean qualitative answers, but retain the respondent’s voice. Edit the punctuation and grammar, but don’t change the meaning of the response.

10. Code missing data properly (skipped due to skip pattern = 999999, unanswered for any other reason (system missing) = sysmis). Check any items where SRC coded 0=No Response and bring to the group for discussion.
The SACS data cleaning procedure

11. Reverse recode if necessary to ensure the conceptual consistency of items within a group of questions. Reverse value labels as well.

✓ Also run correlations on the set of questions. Negative correlations are further evidence that reverse coding may be necessary.

To reverse code, use the following equation:

$$\text{New Value} = (\text{High Value} + 1) - \text{Original Value}$$

For example, if the High Value for a metric is 5, we obtain the New (transformed) Value by subtracting each Original Value from 6 (i.e., $5 + 1$).
The SACS data cleaning procedure

12. Create placeholder variables that will help us find variables when using point and click and will make the dataset more readable—use CAPS for the placeholder variable NAMES and LABELS.

13. Perform descriptive statistics again. Make sure everything still makes sense (e.g., check to make sure the dummy variables add up to the number of cases).

14. Perform a missing data analysis to determine survey fatigue and if there is a pattern to the missing data. Follow the procedure outlined in Missing Data Analysis Procedure.doc.
The importance of addressing missing data

Why are there missing data?

- Fatigue
- Sensitivity
- Lack of knowledge
- Not applicable
- Data processing errors
- Programming errors
Types of Missingness

- **Missing Completely at Random (MCAR)**
  - There is no pattern—the cases with missing data are indistinguishable from those that are complete. Awesome!

- **Missing at Random (MAR)**
  - Only vary depending on other observable variables (e.g. income and parental education).
  - ✔ Want key IVs to be MAR with respect to the DV.

- **Missing not at Random (MNAR)**
  - Not random, but missingness cannot be predicted by variables in the model.
  - ✔ Not ignorable

Missing Data
Why do we care?

- **Independent Variables (IVs)**
  - Sample size and statistical power
  - Bias—Data no longer be representative

- **Dependent Variables (DVs)**
  - Sample size and statistical power
  - Generalizability
    - We must limit our conclusions to the cases which do not have missing data and, therefore, limit our generalization to populations which share the same qualities as those cases.

Missing Data
Missing Data Analysis Procedures

1. Check for drop out/fatigue.

2. Create dummy variables representing cases that are missing data.

3. Test to see if the missing data are biased or if they are randomly distributed along each of the other IVs and DVs of interest.
   - Chi Square test for categorical variables
   - T-test for continuous variables
   - Little’s Chi Square Test for MCAR

4. SPSS “Data Preparation” and “Missing Values Analysis” modules
What do we do when Data are Missing?

If the Data are MCAR

- Listwise (casewise) deletion: uses only complete cases
- Pairwise deletion: uses all available cases (different cases, different Ns)
  - Listwise is preferred over pairwise when sample size is large in relation to the number of cases that have missing data ($\approx \leq 5\%$).
What do we do when Data are Missing?

Methods of Missing Data Replacement

- Mean substitution: substitute mean value computed from available cases

- Regression methods: predict value based on regression equation with other variables as predictors (assumes MAR)
What do we do when Data are Missing?

- Hot Deck: identify the most similar case to the case with missing data and impute the value.

- Approximate Bayesian Bootstrap (ABB): use logistic regression to estimate the probability of response/non-response on the dependent variable.
What do we do when Data are Missing?

- Maximum Likelihood Estimation (MLE): use all available data to generate maximum likelihood-based statistics. EM in SPSS. (Assumes MAR)

- Multiple Imputation: combines methods of MLE to produce multiple data sets with imputed values for missing cases
  - Most recent method. Good because the variance/covariance structure’s preserved.
  - More computationally demanding
Weighting

- A value assigned to each case in the data file.
- Normally used to make the statistics computed from the data more representative of the population.
- The value indicates how much each case will count in a frequency distribution, cross-tab, or other descriptive statistics of a central tendency (i.e., percent, mean)
  - A weight of 2 would mean that the case would count as if it were two identical cases.
  - A weight of 1 would mean that a case only counts as one case (unweighted data)
  - Weights are often fractions, but always positive and non-zero.
- Only one weight per case can be used. If we weight for different characteristics, these weights must be combined together into one weight.
Common Types of Weighting

- Design Weights
  - Normally used to compensate for over- or under-sampling of specific cases.
  - Used to present statistics that are representative of the population.

- Post-Stratification (Non-response) Weights
  - Used to compensate for the fact that persons with certain characteristics are not as likely to respond to the survey.
Calculating Weights

Design Weights
- If we know the sampling fraction (ratio of sampling size to population size) for each case, the weight is merely the inverse of the sampling fraction.

Post-Stratification (Non-response) Weights
- Base on population estimates.
Calculating Weights

- When weighting on multiple characteristics we must use a more complicated procedure that I haven’t learned yet.
  - Weight each separately but sequentially
  - Use a huge X row by Y column table
  - Manual iteration
  - Automatic iteration (Ranking)
  - Logistic regression
Problems with Weights

- Standard errors are more difficult to evaluate.
- The process can be difficult, with little clear advice.
  - Creating practical weights requires arbitrary choices about inclusion of weighting factors and interactions, pooling of weighting cells, and truncation of weights.


