Steps to Analyze Surveys or Administrative Data

1. *Enter the data into an appropriate electronic format*.
   A. For existing administrative data or online surveys, export all necessary fields to a .csv or .xls file. For administrative data, clarify the data source and precise definition of each field. If you are working with data from more than one source, merge the files using a unique identifier that connects each case.
   B. For paper files or surveys:
      1) Number each paper form with a unique identifier. This usually just means collecting the forms into a single stack and numbering each form in sequence.
      2) Prepare a spreadsheet for data entry. Using a standard spreadsheet application, for each item on the form enter the variable name in the first row of each column. The first column should be the unique identifier written on each form. For "select all that apply" items each option should have its own field or column. Include columns for open-ended items. You can set the format appropriate for the type of data in each column. If you are using a statistical analysis package the process and end result are the same, though the individual steps to prepare for data entry will be slightly different than a spreadsheet.
      3) Enter all information from each form in a single row, or case, beginning with the unique identifier. "Yes/No" or other dichotomous items are typically entered as one and zero. Scaled items are typically entered using a number, with the scale beginning at zero or one. "Don't know" or "N/A" responses are often entered as 999. Later you may elect to designate 999 as a missing value. Skip items with ambiguous responses and do your best with handwriting that is difficult to read.
      4) After completing data entry, re-examine the confusing or difficult to read forms and make corrections.
2. *Make a copy of your data file*. Never work on the original data file. Occasionally, you make an irreparable mistake, or you may need to refer to the original data export.
3. *Look at the data*. Make sure it exported/imported correctly into the analysis package. Is there text in fields that should not have text? Numbers in fields that should not have numbers? Are the codes correct for each field? In other words, are binary fields coded 0/1, or continuous fields coded as integers or decimals? Are there the same number of records or rows as there were survey submissions? Is the missing data displayed as you anticipated? If you are not using a package that is conducive to inspecting and cleaning the data, such as R, use an ordinary spreadsheet application and save your data as .csv.

Libre or Open Office are the most reliable and least likely to embed unnecessary information that may later cause problems. Libre Office and Open Office are both free applications and may be downloaded from https://www.libreoffice.org/ or https://www.openoffice.org/

4. *Clean up the data*. As you work through the data file save the all of the syntax and create a second word processing file to make notes on everything you do to the data. This file is a good place to make notes to help you keep track of next steps and reminders about things you need to come back to as you do more advanced analysis.

   A. Make sure all the data fields are in the format you need.

   B. Check/code the variable names if they did not import automatically. As you select variable names consider the following:

      1) Some packages still limit variable names to eight characters. Realistically, meaningful variable names can usually be defined in eight characters or less. If you are working in syntax it is easier to work with shorter variable names than longer names.

      2) Use descriptive, value-free variable names. You do not know who may be looking at the data file or under what circumstances.

   C. Dummy code all nominal level data. "Select all" questions should be entered as dummy coded originally, but "select one" data might not be, especially if you used a web survey. Before creating dummy coded variables run a frequency analysis on the original data field, and again on the new dummy coded variables and compare the two. Look for discrepancies. Save the syntax file.

   D. Make sure the ordinal, interval, and ratio level data are coded in the right direction. In other words, make sure all the positive scales go up, and negative scales go down. If some questions were worded so negative scales go up you need to reverse code those fields. Before you re-code data run a frequency analysis, and then run it again on the recoded data. Look for discrepancies. Save the syntax file.

   E. Code missing data as necessary. Double check to make sure that questions only available to some respondents due to survey branching and skipping are reflected appropriately in the data file. Before you re-code data run a frequency analysis, and then run it again on the recoded data. Look for discrepancies. Save the syntax file.

   F. Assign labels to each variable name and to each scale if you will be using the output of the analysis in your report. Spelling counts.

   G. Double check the coding on scaled variables, especially if you will be doing regression analyses. Remember the intercept is defined as the value of y when x

equals zero. If you decide to re-code any scales, before you do so run a frequency analysis, and then run it again on the recoded data. Look for discrepancies. Save the syntax file.

H. If you will not be working with text fields in this portion of the analysis you can delete these fields.

5. *Make a copy of the data file*. Store the copy of the cleaned data file in a different location so that you may use it if you make an irreparable mistake or need to refer to it.

6. *Examine descriptive statistics*. Examine descriptive statistics for each variable in the data file. Descriptive statistics include frequencies, percentages, range or minimum/ maximum, mean, median, mode, standard deviation, skewness and kurtosis. For dummy coded and other categorical variables you need only run frequencies and percentages. Compute these analyses in batches of like questions. Save the syntax file.

   A. Rough guidelines

      1) Skewness should be < 2

      2) Kurtosis should be < 7

   B. Highly non-normally distributed variables may need to be transformed to be more normally distributed (e.g., square root, log).

   C. Identify important frequencies and percentages relevant to the study. Determine how to present results effectively.

7. *Examine measures of association*. These relationships may include simple group differences using means comparisons and tests of independence or homogeneity, and bivariate correlations.

   A. Crosstabs, with chi-square goodness of fit (categorical)

      1) Independence examines proportionate differences in two categorical variables from a random sample

      2) Homogeneity examines proportionate differences in response between two different populations on a categorical variable

   B. T-tests (continuous)

      1) Single sample examines whether the mean of a variable is different from a fixed value (e.g., zero, known population average)

      2) Independent samples examines the mean difference between two groups on a variable (e.g., individuals who hold four-year degrees vs. those who do not)

3) Paired samples examines the mean difference between two variables (e.g., individual participants before the program and after the program)

C. One-way ANOVA examines the mean difference between two or more groups on a single variable

1) Post-hoc tests to determine which differences between groups are statistically significant

D. Several other types of ANOVA, MANOVA, ANCOVA

E. Correlation examines the direction and magnitude of the relationship between two variables

8. *Examine causal relationships*. These tests examine the direction and magnitude of influence between two or more variables, estimating a causal relationship between a single outcome variable and the partial influence of one or more predictor variables.

A. OLS Regression - tests influence of one or more independent variables on a continuous dependent variable (e.g., quality of services predicted by age of participant, years of participation, gender identity)

B. Logistic Regression - tests influence of one or more independent variables on a categorical dependent variable (e.g., likelihood of completing program predicted by age of participant, years of participation, gender identity)

9. *Examine complex relationships*. After examining basic frequencies, percentages, descriptive statistics, simple associations and group comparisons, and simple OLS or logistic models, you are positioned to examine more complex relationships through interactions, curvilinear models, multilevel models, longitudinal/change over time, factor analysis, or other analyses.