Getting to know your variables

Objectives:

- To be familiar with the unit of analysis for your data.
- To identify the roles of each major variable in your analysis, with reference to the specific research question under study.
- To know the type of each variable in your analysis and its units or categories.
- To become familiar with the valid values of those variables and their substantive interpretation.
- To ensure that your data are correctly labeled and ready for statistical analysis of your research question.
- To identify pertinent restrictions on which cases should be included in your analysis.

All of this information should be included in a data and methods section, so this exercise is the first step in helping you find the information for your data set and research question.

1. Specify the unit of analysis for your data set. Common examples include individual, family, hospital, or county, etc. Consult the documentation for your data set. If there are multiple observations for each respondent, note that as well.

2. Write your research question or hypothesis, including:
   a. The dependent variable(s)
   b. The key independent variable(s)
   c. Any major hypothesized potential confounders, mediators, or control variables

3. For ***each*** of the variables to be used in your analysis, fill information from the codebook and questionnaire for your data set into an electronic copy of table A. Add rows as necessary to accommodate additional variables. See illustrative examples below.

4. Read the questionnaire to identify skip patterns (see column marked "*" in table A).
   a. Make notes about the number and percentage of cases that were omitted due to ***valid skips*** (e.g., if birth weight is your dependent variable, you must omit from your analytic sample any cases for which the question was not asked).
   b. Differentiate the number and percentage of cases that were missing information due to ***nonresponse*** to questions that were asked of them (e.g., children for whom birth weight was requested but not provided).
   c. In your methods section, include information on valid skips and other missing values.

5. If you created new variables for your analysis,
   a. Fill in the requested information into the column marked † in table A.
   b. Add rows to table A to show units, coding, etc., of original (source) variables.
   c. Save syntax for all created variables so you can check for errors and modify syntax if necessary.

6. Prepare your analytic data set:
   a. Impose any needed restrictions (e.g., age group, gender, race, geographic area) that ***pertain to your research questions***. For example, if you are studying only males, exclude females from the data set before running these statistics.
   b. Recode or create new variables that replace missing value codes (e.g., 999) with "system missing" values.

7. Run unweighted descriptive statistics on each of the variables listed above for your data set
   a. Fill them into an electronic copy of table B.
   b. Save the output for future reference.

8. Create a frequency distribution chart of each variable in Table B. Examine it for outliers, unusual values, or unexpected distributions. If you find any, check the documentation and instructions for variable creation.

9. Check the distribution of values for ***each*** variable against the codebook for the data set (see columns marked † in table B**).** If they are inconsistent read through the codebook and literature to identify possible reasons for discrepancies, such as:
   a. Units of measurement
   b. Scale (e.g., grams instead of kilograms)
   c. Transformations (e.g., logged values, percentiles, multiples of standard deviations rather than original units)
   d. Skip patterns
   e. Other missing values

10. Track down information from the published literature (see columns marked * in table B) on ***each*** of your variables ***for a similar population***.
   a. Fill information into table B on values against which to check plausibility of range and central tendency, or percentage distribution.
   b. Check the distribution of values for ***each*** variable against the values from the external source of information about that variable. If they are inconsistent, read through the codebook and literature to identify possible reasons for discrepancies, such as:
      i. Unit of observation
      ii. Units of measurement or observation
      iii. Scale (e.g., grams instead of kilograms)
      iv. Transformations (e.g., logged values, percentiles, multiples of standard deviations rather than original units)

Until you have resolved any discrepancies between your data set and the codebook and outside source(s), do NOT use your data for analysis.

<u>Suggested readings:</u>

J. E. Miller. 2013. *The Chicago Guide to Writing about Multivariate Analysis, 2nd Edition*

- Chapter 4, sections on
  - Types of variables
  - Units
  - Distributions
- Chapter 10, section on resolving the Goldilocks problem: variables and measurement
- Chapter 13, sections related to
  - How to describe the nature and size of the analytic sample
  - Missing by design
  - Variables
  - Non-response

J. E. Miller. 2013. "Getting to know your variables: The foundation for a good working relationship with your data." In *Proceedings of the 2013 Joint Statistical Meetings*, pp. 2016-27.

<u>Suggested podcasts:</u>

- Resolving the Goldilocks problem: Measurement and variables

Table A. Labeling, coding, and missing value information for your variables

Variable name (e.g. acronym on your data set)	**Variable label** (descriptive phrase)	**Type of variable** (nominal, ordinal, interval, or ratio)	**Coding** (for categorical variables) **OR Units** (for continuous variables)	**Plausible range of values** (excluding missing values)	**Missing value codes** (if any)	**Skip pattern?*** (e.g., conditions under which variable not collected)	**Original or created variable?†**
DEPENDENT VARIABLES							
INDEPENDENT VARIABLES							
Key predictor(s)							
Other independent variables							

Illustrative examples: Labeling, coding, and missing value information

Variable name (e.g. acronym on your data set)	**Variable label** (descriptive phrase)	**Type of variable** (nominal, ordinal, interval, or ratio)	**Coding** (for categorical variables) **OR Units** (for continuous variables)	**Plausible range of values** (excluding missing values)	**Missing value codes** (if any)	**Skip pattern?*** (e.g., conditions under which variable not collected)	**Variable from source data or created new?†**
DOCLY	Saw doctor last year	Nominal	1 = yes 2 = no	1, 2	7 = refused 8 = don't know 9 = missing	None for this variable	From source data
BWGRMS	Birth weight	Ratio	Grams	0–6000	9999 = missing	Asked only about children under age 5 years at time of survey.	From source data
CESDSCRE	Depression scale score	Ratio	Continuous	0–60	99 = missing	Asked only of adults	Created from items ##-##.

Table B. **Univariate statistics for each variable from data, codebook, and external reference source**

		Observed values from dataset[1]						**Reference values from external source**					
		For continuous variables				**For categorical variables**		**For continuous variables**				**For categorical variables**	
Variable name (e.g. acronym on your dataset)	**# valid cases for variable (excl. missing values)**	Min.	Max.	Mean	SD[2]	Frequency distribution (%)	**Values & range consistent w/ codebook?**	Min.	Max.	Mean	SD[b]	Frequency distribution (%)	**Values & range consistent w/ external source?**
Dependent Variables													
Independent Variables													
Key predictor(s)													
Potential confounders or mediators													
Control variables													

Illustrative Examples: Univariate Statistics

DOCLY	1,000	NA	NA		NA	68% yes 22% no	Yes	NA	NA		NA	71% yes 29% no	Yes
BWGRMS	989	677	4,432	3,371	59	NA	Yes	338	5,102	3,400	48	NA	Yes
CESDSCRE	970	0	25	8.1	3.0	NA	Yes	0	28	7.6	2.8	NA	Yes

[1] If the data were collected using a complex design that involved stratification or disproportionate sampling, the statistics on the students' data should be weighted before they can be compared with comparison data from the overall population from which the sample was drawn or a similar reference population (Miller 2013a, chapter 13). See Table 3 for an example.

[2] SD: standard deviation