**Getting to know your variables**

**Objectives:**

* To identify the roles of each major variable in your analysis, with reference to the specific research question under study.
* To become acquainted with the context (W's) from which the data were collected.
* To be familiar with the unit of analysis for your data.
* To know the type of each variable in your analysis and its units or categories.
* To learn about skip patterns or other aspects of study design that affects availability of the variables of interest.
* 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 of a research paper, so this exercise is the first step in helping you learn about the pertinent attributes of your data set and research question.

1. Write your research question or hypothesis, including:
   1. The dependent variable(s)
   2. The key independent variable(s)
   3. Any major hypothesized potential confounders, mediators, or control variables
2. Write down the W's for your data set: when, where, and to whom they pertain.
3. 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 (e.g., time points from a longitudinal study), note that as well.
4. 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 1. Add rows as necessary to accommodate additional variables. See illustrative examples below.
5. Read the questionnaire to identify skip patterns (see column marked “\*” in table 1).
6. 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).
7. 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).
8. In your methods section, include information on the reasons for, and numbers of, valid skips and other missing values.

1. If you created new variables for your analysis:
2. Fill in the requested information into the column marked † in table 1. See Miller 2015b for guidance on planning how to create new variables from variables available on the data set.
3. Add rows to table 1 to show units, coding, etc., of original (source) variables.
4. Save syntax for all created variables so you can check for errors and modify syntax if necessary.
5. Prepare your analytic data set:
6. 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.
7. Recode or create new variables that replace missing value codes (e.g., 999) with “system missing” values.
8. Run unweighted descriptive statistics on each of the variables listed above for your data set
9. Fill them into an electronic copy of table 2.
10. Save the output for future reference.
11. Create a frequency distribution chart of each variable in Table 2. Examine it for outliers, unusual values, or unexpected distributions. If you find any, check the documentation and instructions for variable creation.
12. Check the distribution of values for ***each*** variable against the codebook for the data set (see columns marked **†** in table 2**).** If they are inconsistent read through the codebook and literature to identify possible reasons for discrepancies, such as:
13. Units of measurement
14. Scale (e.g., grams instead of kilograms)
15. Transformations (e.g., logged values, percentiles, multiples of standard deviations rather than original units)
16. Skip patterns
17. Other missing values
18. Track down information from the published literature (see columns marked \* in table 2) on ***each*** of your variables ***for a similar population*** (e.g., similar location, date(s), and demographic groups).
19. Fill information into table 2 on values against which to check plausibility of range and central tendency, or percentage distribution.
20. 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:
21. Unit of analysis
22. System of measurement (e.g., British or metric)
23. Scale of measurement (e.g., grams instead of kilograms)
24. 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.**

**Suggested readings:**

J. E. Miller. 2015a. *The Chicago Guide to Writing about Numbers, 2nd Edition.* Chicago: University of Chicago Press.

* Chapter 2, section on "W's"
* Chapter 4, sections on
  + Types of variables
  + Units
  + Distributions
* Chapter 10, 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.

J.E. Miller, 2015b. "Planning how to create the variables you need from the variables you have." Available online at http://press.uchicago.edu/books/miller/numbers/index.html.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1. 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 | 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 |
| SRH | Respondent's self-rated health | Ordinal | 1 = Excellent  2 = Very good  3 = Good  4 = Fair  5 = Poor | 1-5 | 9 = missing | None for this variable | From source data |
| CESDSCRE | Depression scale score | Ratio | Continuous | 0–60 | 99 = missing | Asked only of adults | Created from items ##-##. |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | | | | |
| Table 2. Univariate statistics for each variable from data, codebook, and external reference source | | | | | | | | | | | | | |
| **Variable name** (e.g. acronym on your data set) | # valid cases for variable (excl. missing values) | Observed values from data set[[1]](#footnote-1) | | | | | Values & range consistent w/ codebook? | Reference values from external source | | | | | Values & range consistent w/ external source? |
| **For continuous variables** | | | | **For categorical variables**  Frequency distribution (%) | **For continuous variables** | | | | **For categorical variables**  Frequency distribution (%) |
| Min. | Max. | Mean | SD[[2]](#footnote-2) | Min. | Max. | Mean | SDb |
| **Dependent Variables** | |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Independent Variables** | |  |  |  |  |  |  |  |  |  |  |  |  |
| *Key predictor(s)* |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| *Other variables* |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Illustrative Examples: Univariate Statistics** | | | | | | | | | | | | | |
| DOCLY | 1,000 | NA | 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 |
| SRH | 100 | 1 | 5 | NA | NA | 40% excellent  33% very good  17% good  7% fair  3% poor | Yes |  |  |  |  | 38% exc.  35% very gd  15% good  5% fair  7% poor | 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 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 2015a, chapter 10). See Table 3 for an example. [↑](#footnote-ref-1)
2. SD: standard deviation [↑](#footnote-ref-2)