Getting to Know Your Variables:
An Exercise to Prepare Students to Undertake Data Analysis

Jane E. Miller Ph.D.[1]

November 2013

[1]Professor, Institute for Health, Health Care Policy and Aging Research, and Bloustein School of Planning and Public Policy, Rutgers University. 112 Paterson Street, New Brunswick NJ 08901. Voicemail: (848) 932-6730; fax (732) 932-1253; email: jmiller@ifh.rutgers.edu.

ABSTRACT

This paper outlines a series of steps students should conduct to get to know their dataset and variables before analyzing them, in order to avoid common errors of misinterpretation. Initial steps involve learning about context, unit of analysis, levels of measurement, units, coding, and missing values from documentation about the data source. Later steps require becoming familiar with literature on the topic, and using a dataset to estimate descriptive statistics and create histograms for each variable to compare against the codebook and data from a related population. A suggested schedule of readings and homework assignments for a semester-long course is structured to provide opportunities for students to receive feedback on early steps while working toward a complete research paper about an application of statistical analysis.

KEYWORDS

Data analysis; descriptive statistics; study design; teaching; variables.

As experienced researchers know, each variable in a statistical analysis measures a specific concept in a particular context collected with a specific study design and data collection instrument. Because concepts, context, and study design all affect the valid range and interpretation of numeric values for those variables, it is important that students develop the habit of getting to know each of their variables before analyzing their data. Just as knowledge of the place, time, and circumstances in which someone is living or working can foster a better understanding of their behaviors, knowledge about the context pertaining to one's data can help avert blunders in data analysis. And just as someone's relationships with people are more successful if they get to know them as individuals rather than working from stereotypes, students' data analysis projects will be more effective if they get acquainted with individual variables before working with them.

Too often, however, students simply run statistics on their data without stopping to learn their substantive, real-world meaning or to check the range of values. On a more advanced level, some students forge ahead and interpret multivariate regression coefficients without stopping to consider whether a one-unit increase in an independent variable makes sense for the level of measurement and range of values in that specific variable (Miller 2013a, chapter 10).

For more concrete illustrations of why it is a bad idea to treat all variables as if they were the same, consider the following examples of how not all variables can take on all values. A value of 10,000 makes sense in at least some contexts (places, times, or groups) for annual family income in dollars, the population of a census tract, or an annual death rate per 100,000 persons. However, with rare exceptions, a value of 10,000 does *not* make sense for hourly income in dollars or birth weight in grams, and never fits number of persons in a family, a Likert-type item,[i] a proportion, or an annual death rate per 1,000 persons. A value of –1 makes sense for temperature in degrees Fahrenheit or Celsius, *change* in rating on a 5 point scale, change in a death rate, or percentage change in income, but is completely nonsensical for temperature in degrees Kelvin, number of persons in a family, a proportion, or a death rate. Thus, it is essential that students learn not to think about variables in their analyses as generic, but instead to

understand the specific concepts behind each of the variables they are studying, which will help them identify reasonable levels and ranges for each variable in their analysis.

Failing to become acquainted with one's data can lead to many mistakes in data preparation, model specification, and interpretation of results. This story about the experience of a young research trainee will illustrate: She came to me in the ninth week of a ten-week training program, puzzled by the results of her multivariate regression. She was analyzing predictors of birth weight using a nationally representative survey sample from a developing country circa 2002, which she had downloaded from a research data website but hadn't cleaned or evaluated before analyzing it. In the sample, birth weight in grams ranged up to 9,999 with a mean of 8,000 in the sample. Had she taken the time to look up the expected range of values for that concept (birth weight) and units (grams), she would have immediately seen a red flag because 9,999 grams is roughly 22 lbs., which is a typical weight for a 1 year-old, not a newborn!

A second warning sign was that two-thirds of the sample had a birth weight value of 9,999 – a very high value for such a substantial share of a sample, and one that is unlikely to be explained by either outliers or data entry errors alone. By examining the study documentation and questionnaire, she discovered that this distribution occurred due to a skip pattern designed to minimize recall bias, such that only mothers of children under age 5 years were asked about birth weight. Children aged 5 through 17 years should have been omitted from the analytic sample because the dependent variable was missing (9999) for them. If the student had familiarized herself with the concepts, units, context, and study design related to her topic and data, she could have averted the need to rerun all of her statistical analyses and rewrite major sections of her paper at the last minute to reflect the correct sample and values.

In this paper, I outline a series of steps to be conducted whenever a student or other researcher undertakes a project on a topic or dataset that is new to them. Getting acquainted with variables is a multi-step process involving several resources about the data source and topic under study. Information on attributes such as levels of measurement, range of reasonable values, skip patterns, and other missing values is essential for data preparation, including exclusion criteria for the analytic sample and creation of new variables; choice of pertinent descriptive and inferential statistics; design of correct charts and tables; and writing correct prose descriptions for the data and methods, results, and discussion sections of a research paper.

This “getting to know your data and variables” exercise incorporates a wide range of research methods concepts as well as univariate statistics. To ensure adequate time for students to master the definitions and concepts before applying them to their own data, the steps for the exercise should be spread out over a semester-long graduate or advanced undergraduate methods course or research internship. Results and feedback on early steps will inform later steps in the exercise, so it is best undertaken as part of a project in which each student is using one dataset to analyze a single research question rather than separate exercises using different data and topics for each task. Different students can study different questions or data, but each student should trace one topic and dataset throughout the entire project. To guide students through drafting sections of the paper as they complete the related tasks, I have included a suggested schedule of readings and assignments at the end of this paper.

RESOURCES NEEDED FOR THIS EXERCISE

Early in this exercise, students must commit to a specific research question so they can identify the relevant variables in their dataset and conduct focused searches to identify articles, books, web sites etc. on their topic. In addition, they will need documentation on the data source, including a description of study design, the questionnaire or other form(s) used to collect the data, a codebook for the dataset, a copy of the electronic data file, and statistical software to analyze it. Finally, they should have access to standard methods textbook such as Chambliss and Schutt (2012), Treiman (2009) or Wooldrige (2012) that define and illustrate the concepts mentioned in the instructions for the exercise.

BACKGROUND INFORMATION ON DATA AND VARIABLES

Confirming Availability of Variables

The first step is for students to confirm that the variables they would like to use for their analysis are available in their dataset. Sometimes new versions of existing variables need to be created from variables in the dataset (Miller 2013b), such as creating an age group variable from a continuous measure of age, but to do so, there must be a measure of that concept (age) in the dataset.

Research Question

Next, students should write out their research question, including the dependent variable(s), the key independent variable(s) and any major hypothesized potential confounders, mediators, or control variables. Doing so will help them identify the full set of variables that they should include in this exercise.

Attributes of the Dataset

Before getting to know characteristics of the individual variables, students should spend some time learning about attributes of their dataset. Information on these attributes will be used in later steps of this assignment when students search for articles or reports on their topic in a similar context in order to identify external reference values of their variables to compare against observed values in their own dataset.

Restrictions on Analytic Sample

In many cases, students will need to impose limits on the original dataset to create an analytic sample to whom their research question pertains (Chambliss and Schutt 2012; Miller 2013a, chapter 13), such as limiting the sample to particular demographic traits, minimum test scores, or having a specific disease. It might also mean excluding subgroups that don't meet minimum sample size, as when aren't enough cases in one or more subgroups of a key variable to provide sufficient statistical power and it would not be theoretically sensible to combine them with other subgroups in the analysis. In some instances, it may be necessary to exclude cases for whom a key variable was not collected, as in the birth weight example above. Have students read the literature on their topic to learn how other researchers have identified pertinent exclusion criteria, and then make notes about these exclusions and the reasons for each, given their research question and dataset. In the electronic copy of their dataset, students should impose any needed restrictions and save the syntax used to make those exclusions, which they will need when verifying results and writing the data and methods sections of their research papers (Treiman 2009).

Context of the Data

The next step in getting familiar with the data is to identify the context -- when, where, and to whom it pertains -- also known as "the W's" (Miller 2004, 2013a). This information is critical because knowing the topic alone is often insufficient to help students recognize unrealistic values of their variables. For example, suppose you are teaching a course about global economic patterns and have assigned your students to analyze income using any dataset they can find that includes a measure of annual family income. If one student is studying Brazil today, another is studying the US today, another the US 100 years ago, another a sample of recent GED recipients in one American city, and another the salaries of NFL football players, they will observe vastly different levels and ranges of income. An annual income that would be absurdly high for a recent GED earner would be just as absurdly low for a recent NFL first round draft pick.

Much of the information about when, where, and who will come from the documentation for the dataset, which should explain the study design and sampling plan. In some cases, one or more of the W's from the original dataset will have been modified while creating the student's analytic sample in order to suit their research question, as noted above.

Unit of Analysis

Next, students should identify the unit of observation, e.g., whether the data pertain to individual persons, families, census tracts, institutions, or some other level of aggregation (Chambliss and Schutt 2012; Miller 2004, 2013a). Knowing units of observation helps ascertain plausible range of values for their variables. For example, the number of persons in a family will be much lower than the population of a census tract or a school. Information on unit of observation should be confirmed in the documentation and labeled in the electronic dataset.

Attributes of individual variables

Next, students should familiarize themselves with the attributes of all of the variables to be used in their analysis, much of which can be gleaned from the study documentation. Some variables they will be analyzing in the same form in which they appeared in the original dataset. Others might be new variables they created from variables in the original dataset (Miller 2013b), such as:

- collapsing multi-category variables into fewer categories, e.g., five-category race simplified into three-category race, or a dummy (binary) indicator of poor/non-poor from a multi-category measure of poverty status;
- categorical versions of continuous variables, e.g., age group from continuous age;
- aggregated variables, e.g., income calculated from several sources, or scales that combine responses to multiple items;
- calculated variables such as family income as a percentage of the Federal Poverty Level (FPL), computed from income, family size and age composition, and FPL thresholds,
- variables transformed by taking logarithms, standardizing, or changing scale.

If they created new variables for their analysis, have students save the syntax so they can check for errors and revise how those variables were created, if necessary (Treiman 2009). Have them include rows in Table 1 to show units, coding, etc., of both the original (source) variables from which the new variables were created and the new version they will use in their analysis.

Insert Table 1 about here

Have students start by downloading[ii] or creating an electronic version of Table 1, which is a grid for organizing information about the labeling, coding, units, and missing value information on each of their variables, with one row for each variable. The major row headings in the table organize the variables based on whether they are dependent variable(s), key independent variables, mediating, confounding, or control variables for their particular research question. Also include rows for variables that constitute filter questions (e.g., used to restrict their analytic sample; Chambliss and Schutt 2012) or sampling weights used in their analysis.

Variable Name and Label

For each variable, students should fill in the appropriate section (row heading) of Table 1 with the *variable name* (acronym, often limited to 8 characters) used to identify the variable in their dataset, which they can find in the codebook for the dataset and verify in the electronic dataset. They should also fill in the *variable label* – a longer descriptive phrase that helps convey the substantive meaning of the variable (often limited to 30 characters). If they rename an item in their dataset with a more informative variable name (e.g., "gender" instead of Q117), have them include the original question name in the variable label so they can track it back to the documentation and original dataset.

Units and Categories

The next step is to fill in units and/or categories for every variable in the analysis. For all continuous variables, the pertinent information includes the system of measurement (e.g., income in dollars, Euros, or Yuan), the level of aggregation (e.g., hourly, monthly, or annual income), and the scale of measurement (e.g., income in dollars, thousands of dollars, or millions of dollars). For all nominal or ordinal variables, instead fill in category names and their associated numeric codes used in the dataset.

For example, household structure might be coded 1= married couple family household; 2 = female householder family, 3=male householder family, 4 = female householder non-family, and 5 = male householder non-family. For some ordinal variables such as income group or age group, units will pertain as well. For others, such as letter grade or words describing frequency (e.g., "never," "rarely," "sometimes," "often," "always"), units are not relevant. Information on units and coding can be extracted from the codebook and documentation. See Miller (2004 or 2013a) or Chambliss and Schutt (2012) for more on levels of measurement and units.

Missing Value Codes and Reasons

For all variables in Table 1, have students read the documentation and questionnaire to learn about skip patterns and other design issues that lead to some cases in their sample not having a response to one or more questions due to *valid skips* (Chambliss and Schutt 2012; Miller 2013a, chapter 13). One type of skip pattern occurs when a question does not pertain to some respondents. For instance, if a respondent reported that they were not working in the week prior to the survey, any questions related to the number of hours worked or wages during that period are irrelevant for them, so they should receive codes for not applicable to the latter questions. Another type of skip patterns arises in surveys that use a split sample design to administer specialized topic modules only to a randomly selected portion of the overall sample. For example, the 2011 American Time Use Survey asked a subsample a detailed "Leave Module" about wage and salary workers' access to paid and unpaid leave and the ability to adjust their work schedules and locations (Bureau of Labor Statistics 2013).

In addition, students should look in the codebook for codes that identify item *non-response* – when a respondent did not answer a question that was asked of them (Chambliss and Schutt 2012). Most datasets will designate separate numeric (or sometimes alphanumeric) codes for each of these reasons so users can distinguish among them, e.g., 97 = not applicable; 98 = module not administered to case; and 99 = item non-response. After filling in missing value codes for each of their variables into Table 1, students should update the electronic dataset if necessary to identify the various missing value codes so those values are treated correctly during statistical analysis.

Once information on missing values has been recorded for each variable in Table 1 and the dataset, students should exclude cases with missing values on key independent and dependent variables from their analytic sample. Such exclusions will affect both the sample size and composition, and should be imposed before statistics are run on the sample. Have students note the exclusion criteria and the number of cases excluded based on missing values for each variable so they can describe these steps accurately in their methods section (Miller 2013a, chapter 13).

Plausible Range of Values

The next step is to fill in information on the plausible range of values for each variable so students can identify any out-of-range values that occur in the data. The range of credible values can be affected by definitional limits, what is conceptually plausible, and the context of measurement (Miller 2013a, chapter 10). For instance, the percentage of a whole must *by definition* fall between 0 and 100; however a percentage *change* can be negative or exceed 100. Many other variables also cannot assume negative values. For example, an index constructed by summing 20 items each of which could range from 0 to 3 will have a mathematically-defined minimum of 0 and maximum of 60.

The *conceptually plausible range* is topic-specific, as with infant birth weight, which is limited by physiological and anatomical constraints. It is also *unit-specific*: for example, live births in the United States have birth weight in grams that range from about 400 to 5,900 (Miller 2013a, table 5.4), but the corresponding range in ounces is 14 (less than 1 lb.) to 208 (13 lbs.) Finally, context (when, where and who is in a sample) will affect the range of reasonable values, as with the income examples cited earlier.

Emphasize to students that the range of values that is definitionally possible and conceptually plausible can and often does differ from the *observed* range of values in their data, which they will investigate in a later step. For instance, although in theory a widely used depression scale (the CESD

scale) could range from 0 to 60 points, in the general population the mean is between 8 and 10 and scores above 20 are rarely observed (Radloff and Locke 1986). To help students learn about the substantively relevant range of values for each of their variables, have them consult the published literature on their topic.

DESCRIPTIVE STATISTICS

Once students have received feedback from their instructor or research supervisor on the conceptual attributes of each variable in Table 1, have them complete Table 2 by filling in descriptive statistical information from their dataset, the codebook for the original data source, and one or more reference sources related to each of their key variables.

Insert Table 2 about here

Using the electronic copy of their datasets, have students run unweighted descriptive statistics on each of the variables involved in their analysis. They should then fill information from the statistical output into Table 2 on the number of cases for which there are valid (non-missing) values of each variable, the observed minimum, maximum, mean and standard deviation values for each continuous variable, and the frequency distribution for each categorical variable. Remind them to save their output for constructing tables, charts, and prose descriptions for the results section of their papers, and for verifying their results and the steps taken to generate those results. To make it easier for them to find the pertinent syntax and output for each section of the paper, suggest that students save separate files for recoding, descriptive statistics, and bivariate syntax and output, naming each according to its purpose, such as "data preparation syntax," or "univariate output".

Frequency Distribution Charts

A critical next step is create a chart to display the distribution of each variable because summary statistics such as mean, median, mode and standard deviation alone can obscure important issues in the distribution of observed values (Miller 2004, 2013a, chapter 4). Consider the example of a student who used instructions from a published article to create an acculturation scale for use as a predictor variable in a multivariate regression. For a nationally representative sample of Latinos, the mean of the acculturation scale was 4.56 with a standard deviation of 2.23. However, a histogram (Figure 1) revealed that the distribution was highly unusual, with three small approximately normal distributions just above values of 0, 2 and 4, gaps between those distributions, and spikes at exact values of 6.0 and 7.0.[iii] As a consequence, the constructed "scale" variable was neither a continuous variable for which one-unit increases could be sensibly interpreted, nor a categorical variable with categories that could be modeled in the regression (Miller 2013a, chapters 9 and 15; Wooldrige 2012).

Insert Figure 1 about here

By examining the distribution of the scale and the original component variables, the student determined that in order to meet both conceptual criteria and empirical assumptions about the distributions of variables in the model, two of the component variables (the continuous variable and one of the categorical variables) should be entered as separate predictors in the regression model. Help your student avoid analyzing "funky" variables like this one by requiring that they display the distribution of each variable and evaluate them for plausibility using the criteria described next.

Comparison against the Codebook for the Data Source

Have students check the distribution of values for *each* of their variables against the codebook for the dataset**.** If distributions are inconsistent between those sources, they should read through the codebook to identify possible reasons for discrepancies such as differences in units of measurement, scale (e.g., grams instead of kilograms), or numeric missing values codes that they neglected to identify as such in their dataset. Also have them consider exclusions they imposed on their sample, which might explain why such differences might actually be correct. For example, if they have restricted the sample to persons aged

25 to 34, the distribution of annual income would be expected to have a lower mean and narrower range than the distribution of annual incomes for all ages of adults from the original sample.

Comparison with a Related Population

To become familiar with which values are realistic for each of their variables, assign students to track down descriptive statistical information on those variables from the published literature on their topic for a sample that is similar to theirs in terms of location, time period, and demographic characteristics (those W's again!), but is based on a different sample. Into Table 2, they should fill in information from those reference sources on values against which to check plausibility of range and central tendency, or percentage distribution, as well as information about who, when, and where studied.

Table 3 is an example of a comparison of a study dataset (the NHANES) against national data for a similar period. The footnotes to Table 3 provide information about sample restrictions, definitions of variables, and calculations used to make the comparison. As in that example, 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 national data.

Insert Table 3 about here

Students should then compare the distribution of values for each variable in their data against the reference values from the external source of information about that variable. If those values are inconsistent, have them read through the codebook and literature to identify possible reasons for discrepancies, such as different units of observation or measurement (system of measurement, level of aggregation, or scale) between their sample and the reference population. They should also watch for transformations such as logged values, percentiles, or multiples of standard deviations rather than original units, which might explain observed differences between values observed for their sample and those from the reference population.

If the values in their dataset are substantially different from those listed in the codebook or in data from other studies of their topic, students should refrain from analyzing the data until they understand the reasons for those discrepancies, make any needed corrections in the dataset, and describe them in their notes about the data preparation.

SUMMARY

Getting acquainted with one's data and variables is an essential step for ensuring that statistical analysis to address a research question is conceived and interpreted based on specific information about the data at hand. Table 4 presents a suggested timeline of readings and homework assignments that has been field tested in undergraduate and graduate methods courses and an undergraduate research training program at a large state university. Many of these steps to prepare, clean, and perform consistency checks are behind-the-scenes work that does not need to be reported in their papers, but are crucial for ensuring that the data they analyze make sense for their specific topic and data, and that their statistics and interpretation thereof are based on correct information.

Insert Table 4 about here

Although the steps involved in this exercise are extensive and time consuming, it is time well spent because each step yields information that should be included on a paper describing the analysis. Reading the literature on the topic will provide information needed for the introduction, literature review, and discussion sections. A detailed understanding of study design and variables from documentation, questionnaire, and codebook will provide information for a comprehensive data section, appropriate model specification, interpretation of statistical results, and discussion of study strengths and limitations. The exercise will help students learn the steps needed to conduct a correct, complete application of a statistical analysis, which they can use throughout their careers whenever they undertake an analysis of a topic or dataset that is new to them.

ACKNOWLEDGEMENTS

An earlier version of this paper was presented at the Statistical Literacy session at the 2013 Joint Statistical Meetings in Montreal, Canada. I would like to thank session participants for their comments, and Dawne Mouzon and Brittany Battle for feedback on earlier drafts of this manuscript.

REFERENCES

Bureau of Labor Statistics. 2013. "American Time Use Survey User's Guide: Understanding ATUS 2003 to 2012." Available online at http://www.bls.gov/tus/atususersguide.pdf Accessed October 2013.

Chambliss, Daniel F., and Russell K. Schutt. 2012. M*aking Sense of the Social World: Methods of Investigation, 4th Edition.* Thousand Oaks, CA: Sage Publications.

Miller, Jane E. 2013a. *The Chicago Guide to Writing about Multivariate Analysis, 2nd Edition*. Chicago: University of Chicago Press.

-----. 2013b. "Planning How to Create the Variables You Need from the Variables You Have."

-----. 2004. *The Chicago Guide to Writing about Numbers*. Chicago: University of Chicago Press.

Radloff, Lenore S., and Ben Z. Locke. 1986. "The Community Mental Health Assessment Survey and the CES-D Scale." In *Community Surveys of Psychiatric Disorders,* edited by M. M. Weissman, J. K. Myers, and C. E. Ross. New Brunswick, NJ: Rutgers University Press.

Treiman, Donald J. 2009. *Quantitative Data Analysis: Doing Social Research to Test Ideas*. San Francisco: Jossey-Bass.

US Department of Health and Human Services. 1997. *National Health and Nutrition Examination Survey, III, 1988–1994.* CD-ROM Series 11, no. 1. Hyattsville, MD: National Center for Health Statistics, Centers for Disease Control and Prevention.

Ventura, Stephanie J., Joyce A. Martin, Sally C. Curtin, and T. J. Mathews. 1999. "Births: Final Data for 1997." *National Vital Statistics Report* 47 (18). Hyattsville, MD: National Center for Health Statistics.

Wooldridge, Jeffrey M. 2012. *Introductory Econometrics: A Modern Approach.* 5th ed. Mason, OH: Thomson South-Western.

NOTES

[i] Likert-type items are ordinal variables that measure the extent of a person's beliefs, attitudes, or feelings towards some topic, often on a 5-point scale ranging from "strongly agree" to "strongly disagree" or similar coding (Chambliss and Schutt 2012). For example, "Rate the extent of your agreement with the following statement: 'Automatic weapons should be banned.' Strongly disagree, disagree, neutral, agree, strongly agree."

[ii] The grids for Tables 1 and 2 can be downloaded from http://www.press.uchicago.edu/books/miller/multivariate/App4_10.1.pdf

[iii] Closer examination revealed that the acculturation "scale" had combined one continuous variable that was measured as a proportion (bounded between 0 and 1) with two categorical variables that each took on integer values of 0 and 2, and another categorical variable that assumed values of 0, 1 and 2. Scales should combine variables that share a common level of measurement and coding scheme. Moreover, the component variables were related to one another such that it was mathematically impossible for anyone to have values between 1.0 and 2.0, between 3.0 and 4.0, between 5.0 and 6.0, or between 5.0 and 6.0, yielding the bizarre distribution shown in Figure 1.

TABLE 1. Labeling, coding, and missing value information to be filled in for each variable							
Variable name (e.g., acronym in your dataset)	**Variable label**	**Level of measurement (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?**
Dependent Variables							
Independent Variables							
Key predictor(s)							
Potential confounders or mediators							
Control variables							
Illustrative Examples							
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	500–6000	9999 = missing	Asked only about children < 5 years old at time of survey.	From source data
CESDSCRE	Depression scale score	Ratio	Points	0–60	99 = missing	Asked only of adults	Created from items ##-##.

TABLE 2. Univariate statistics to be filled in for each variable from data, codebook, and external reference source

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

Notes to Table 2

[a] 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.

[b] SD: standard deviation

TABLE 3. Example table comparing values of variables in analytic dataset and a comparison population

	1988–1994 NHANES III sample[abc]	**All US births,1997**[d]
Birth weight		
Median (grams)	3,402	3,350
% Low birth weight (<2,500 grams)	6.8	7.5
Race/ethnicity		
Non-Hispanic white	73.4	68.4[e]
Non-Hispanic black	16.9	17.0
Mexican American	9.7	14.6
Mother's age		
% Teen mother	12.5	12.7
Mother's education		
Median (years)	12.0	12.8
% <High school	21.6	22.1
% +High school	35.0	32.4
Mother smoked while pregnant (%)	24.5	13.2
Number of cases	9,813	3,880,894

Source: Adapted from Miller (2013a), Table 5.5

[a] Weighted to population level using weights provided with the NHANES III; sample size is unweighted.
[b] Information for NHANES III is calculated from data extracted from National Center for Health Statistics (US Department of Health and Human Services, 1997).
[c] Includes non-Hispanic white, non-Hispanic black, and Mexican American infants with complete information on family income, birth weight, maternal age, and education.
[d] Information for all US births is from Ventura et al. (1999).
[e] For consistency with the NHANES III sample, racial composition of US births is reported as a percentage of births that are non-Hispanic white, non-Hispanic black, or Mexican American, excluding births of other Hispanic origins or racial groups. When all racial/ethnic groups are considered, the racial composition is 60.1% non-Hispanic white, 15.0% non-Hispanic black, 12.9% Mexican American, 5.4% other Hispanic origin, and 6.6% other racial groups.

TABLE 4. Suggested order and timing of assignments for getting to know your variables				
Week #	Step(s)	Readings	Comments	Assignment
1	Confirm that measures of the desired concepts are available in the dataset. Name the dataset and list the W's. Identify the unit of observation. Write the research question and identify IV, DV, and levels of measurement for each.	Chambliss and Schutt 2012, Chapter 3 Miller 2004, Chapter 4 Codebook for dataset		One page description of dataset, context, unit of analysis, research question, and key variables.
2	Write first part of the methods section: Describe the study design and context of the dataset.	Chambliss and Schutt 2012, Chapters 2, 4, and 5 Documentation for dataset		
3	Write next part of methods section: Identify and explain restrictions on analytic sample to fit the research question and dataset.			Draft of methods section: study design and restrictions on analytic sample.
4	Conduct background readings to learn about substantively plausible ranges of the independent and dependent variables. Impose restrictions on analytic dataset in electronic copy of dataset. Save syntax.	Literature search on topic	Sample restrictions to suit research question and dataset.	
5	Fill in Table 1, sections on units, categories, and missing values. Define system missing values for each variable into the electronic copy of the dataset.	Codebook for dataset		Completed Table 1.
6	Create histograms for each variable. Fill descriptive statistics into Table 2. Fill codebook information into Table 2.	Codebook for dataset	Run descriptive statistics on all variables in the analysis.	Histograms for dependent and key independent variables.

7	Fill information on outside reference source into Table 2. Write next part of methods section: Description of variables.	Literature search on topic		Completed Table 2. Draft description of variables for methods section.
8	Revise electronic copy of dataset to correct for discrepancies between data, codebook and references. Save syntax. Redo descriptive statistics for any revised variables, and fill into Table 2.		E.g., make corrections to sample exclusions, missing value codes.	
9 and 10	Conduct statistical analysis for paper. Write results section of paper.	Chambliss and Schutt 2012, Chapter 8 Miller 2004, Chapters 5 and 9	Univariate description of sample, and bivariate analyses to test main hypothesis. Advanced students also turn in multivariate results.	Draft results section: Tables and prose description of how statistical results address research question.
11	Finalize data and methods section of paper: describe analytic plan, statistical methods.	Miller 2004, Chapter 10	See assignment from week 2.	
12	Write strengths and limitations for discussion section of paper	Miller 2004, Chapters 10 and 11		Final paper, including introduction, revised data and methods, results, and discussion sections; complete list of references including all those used to complete Tables A and B.

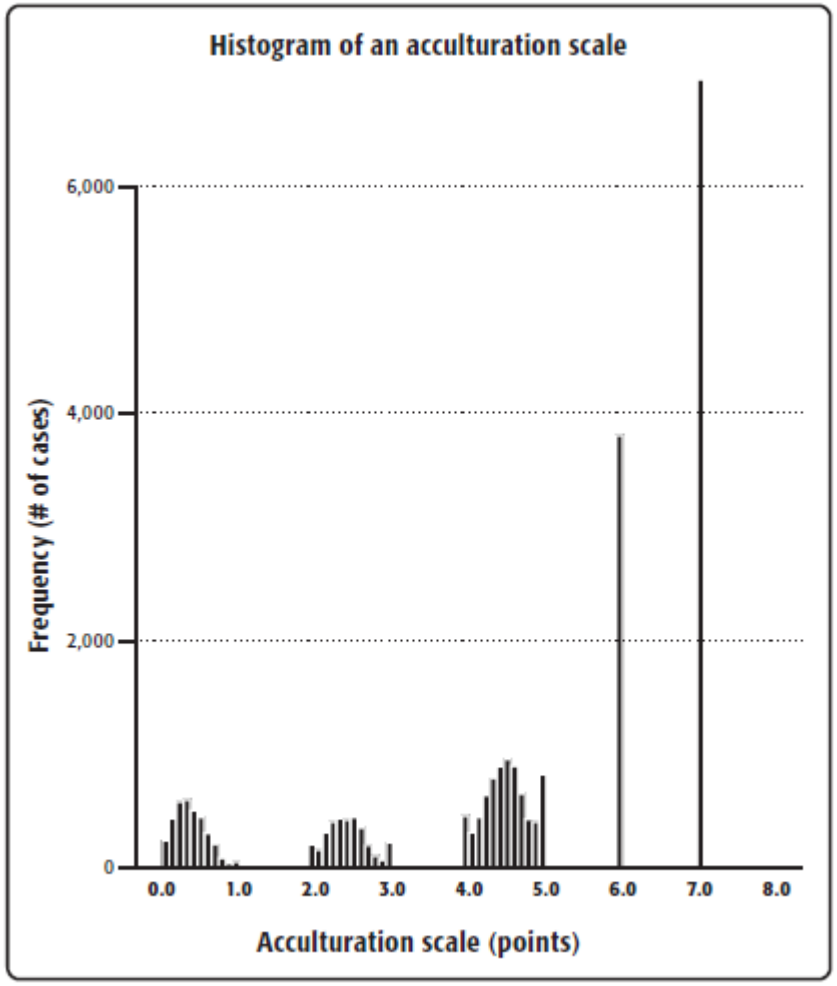

Figure 1. Frequency distribution for an incorrectly-conceived acculturation scale