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

Jane E. Miller Ph.D.

ABSTRACT

This appendix outlines a series of steps students should conduct to get to know their data set 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 data set 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.

INTRODUCTION

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. 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,[1] 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 on the Kelvin scale, 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

---

[1] 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.' Possible responses include strongly disagree, disagree, neutral, agree, strongly agree."

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 looked 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 appendix, I outline a series of steps to be conducted whenever a student or other researcher undertakes a project on a topic or data set 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 data set 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 sets, but each student should trace one topic and data set throughout the entire project. To guide students through drafting sections of the paper as they complete the related tasks, at the end of this appendix I have included a suggested schedule of readings and assignments.

RESOURCES NEEDED FOR THIS EXERCISE
Early in this exercise, students must commit to a specific data set and research question so they can identify the relevant variables in that data set 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 data set, 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) or Treiman (2009) that define and illustrate the concepts mentioned in the instructions for the exercise.

BACKGROUND INFORMATION ON DATA AND VARIABLES
Research Question
The first step is for students to write out their research question, including the dependent variable(s), the key independent variable(s) and any hypothesized potential confounders, mediators, or control variables, filter questions and sampling weights. Doing so will help them identify the full set of variables to include in this exercise.

Availability of Variables

Next, they should confirm that the variables they would like to use for their analysis are available in their data set. Sometimes new versions of existing variables need to be created from variables in the data set (Miller 2015b), such as creating an age group variable from a continuous measure of age. To create that new variable, there must be a measure of that concept (in this case, age) in the data set.

Attributes of the Data set

Before getting to know characteristics of the individual variables, students should spend some time learning about attributes of their data set. 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 data set.

Restrictions on the Analytic Sample

In many cases, students will need to impose limits on the original data set to create an analytic sample to whom their research question pertains, such as limiting the sample to particular demographic traits, minimum test scores, or having a specific disease (Miller 2015a, chapter 10). In other instances, restricting the sample will involve excluding subgroups that don't meet minimum sample size, as when there 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. Or 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 data set. Then, in the electronic copy of their data set they 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 those data pertain – also known as "the W's" (Miller 2015a, chapter 2). 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 data set 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 to whom the data pertain will come from the documentation for the data set, which should explain the study design and sampling plan. In some cases, one or more of the W's from the original data set 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 analysis, e.g., whether the data pertain to individual persons, families, census tracts, institutions, or some other level of aggregation (Chambliss and Schutt 2012, chapter 2; Miller 2015a, chapter 4). Knowing the unit of analysis helps ascertain a 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 analysis should be confirmed in the documentation and labeled in the electronic data set.

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. Have them download[2] or create an electronic version of table 1, which is a template of 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 table 1 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, chapter 7) or sampling weights used in their analysis.

Some variables they will be analyzing in the same form in which they appeared in the original data set. Others might be new variables they created from variables in the original data set (Miller 2015b), 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). They should also 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(s) they will use 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 data set, which they can find in the codebook for the data set and verify in the electronic data set. 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 data set 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 data set.

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, fill in category names and their associated numeric codes used in the data set. For example, household structure might be coded 1= married couple 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

---

[2] The instructions for this assignment and grids for Tables 1 and 2 can be downloaded from the web site of supplemental materials for *The Chicago Guide to Writing about Numbers, 2nd Edition*: http://www.press.uchicago.edu/books/miller/numbers/index.html. Appendix4_10A.

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 (2015a) 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 2015a, chapter 10). 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. Most data sets 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 data set 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 data set, students should exclude cases with missing values on key independent and dependent variables from their analytic sample.[3] 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 2015a, chapter 10).

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 2015a, 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 2015a, 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.

---

[3] If imputed values of the variable(s) are available on the data set, they might choose to use those versions instead of omitting cases with missing values on the original variable. Advanced students might elect to impute values themselves using standard algorithms. If so, imputation flags should be included in the analysis, with notes on the frequency of imputation for each variable filled into table 2, and the imputation strategy described in the final paper. See Miller (2015a, Chapter 10) and Westat (1994) for more on using imputed values.

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													
		Observed values from data set[4]						Reference values from external source					
		For continuous variables				For categorical variables	Values & range consistent w/ codebook?	For continuous variables				For categorical variables	Values & range consistent w/ external source?
Variable name (e.g. acronym on your data set)	**# valid cases for variable (excl. missing values)**	Min.	Max.	Mean	SD[5]	Frequency distribution (%)		Min.	Max.	Mean	SD[b]	Frequency distribution (%)	
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

---

[4] 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.

[5] SD: standard deviation

Emphasize 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 CES-D 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 data set, the codebook for the original data source, and one or more reference sources related to each of their key variables.

Using the electronic copy of their data sets, 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".
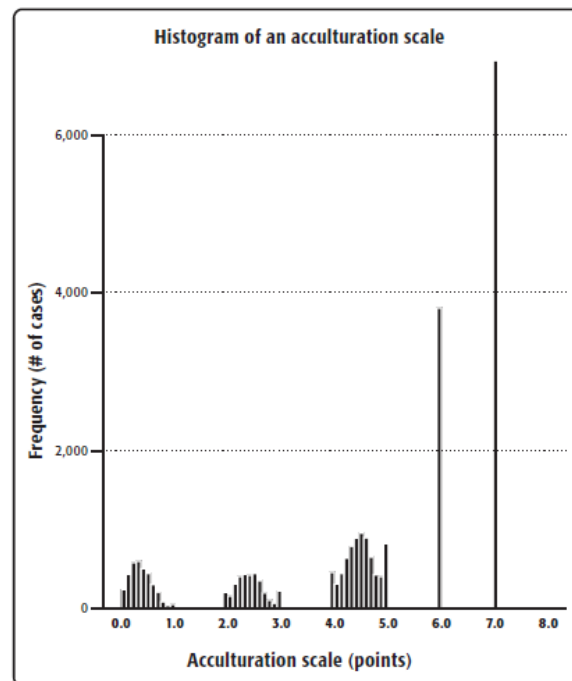

Figure 1. Frequency distribution for an incorrectly-conceived acculturation scale

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 2015a, 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 3 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.[6] 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.

 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 have been analyzed separately rather than used to construct the scale. Help your student avoid analyzing "funky" variables like that 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 data set. If distributions are inconsistent between those sources, they should examine 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 data set. 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.

 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

---

[6] 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.

units, which might explain observed differences between values observed for their sample and those from the reference population.

If the values in their data set 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 data set, and describe them in their notes about the data preparation.

Table 3 is an example of a comparison of a study data set (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 student's data were collected using a complex design that involved stratification or disproportionate sampling, their statistics should be weighted before they can be compared with national data.

Table 3. Example table comparing values of variables in analytic data set 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 (2015a), Table 5.5
Notes to Table 3
[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.

SUMMARY

Getting acquainted with one's data set 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 students' papers, but that is 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.

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, guidance on a suitable analytic plan, 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 data set that is new to them.

ACKNOWLEDGEMENTS

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, D. F., and R. K. Schutt. 2012. M*aking Sense of the Social World: Methods of Investigation, 4*[th] *Edition*. Thousand Oaks, CA: Sage Publications.

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

-----. 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.

Radloff, L. S., and B. 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.

Westat. 1994. "National Health and Nutrition Examination Survey III, Section 4.1: Accounting for Item Non-Response Bias." In *NHANES III Reference Manuals and Reports.* Washington, DC: US Department of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, National Center for Health Statistics.

Table 4. Suggested schedule of assignments for getting to know your variables in a semester-long course

Week	Step(s)	Readings	Comments	Assignment
1	• Orientation to research project			
2	• Write the research question and identify IV, DV, and levels of measurement for each. • Confirm that measures of the desired concepts are available in the data set. • Name the data set and list the W's. • Identify the unit of analysis.	Chambliss and Schutt 2012, Ch. 3 Miller 2015a, Ch. 4 Codebook for data set		One page description of data set, context, unit of analysis, research question, and key variables.
3	• Write first part of the methods section: Describe the study design and context of the data set.	Chambliss and Schutt 2012, Chs. 2, 4, and 5 Documentation for data set		
4	• Write next part of methods section: Identify and explain restrictions on analytic sample to fit the research question and data set.			Draft of methods section: study design and restrictions on analytic sample.
5	• Conduct background readings to learn about substantively plausible ranges of the independent and dependent variables. • Impose restrictions on analytic data set in electronic copy of data set. Save syntax.	Literature search on topic	Sample restrictions to suit research question and data set.	
6	• Fill in table 1, sections on units, categories, and missing values. • Define system missing values for each variable into the electronic copy of the data set.	Codebook for data set		Completed table 1.
7	• Create histograms for each variable. • Fill descriptive statistics into table 2. • Fill codebook information into table 2.	Codebook for data set	Run descriptive statistics on all variables in the analysis.	Histograms for dependent and key independent variables.

8	• Fill information on outside reference source into table 2. • Write description of variables for methods section.	Literature search on topic		Completed table 2. Draft description of variables for methods section.
9	• Revise electronic copy of data set 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.	Revised table 2.
10, 11	• Conduct statistical analysis for paper. • Write results section of paper.	Chambliss and Schutt 2012, Ch. 8 Miller 2015a, Chs. 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.
12	• Finalize data and methods section of paper: Describe analytic plan, statistical methods.	Miller 2015a, Ch. 10	See assignment from week 2.	
13	• Write strengths and limitations for discussion section of paper	Miller 2015a, Chs. 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 1 and 2.