Getting to know your variables

Jane E. Miller, PhD

Why is it important to get to know your variables?

- Each variable measures
  - A specific concept
    - Numeric values have particular meanings that differ depending on the nature of that concept
  - In a particular context
    - Place, time, and group to whom do the #s pertain
  - In specific units
  - Collected with a particular study design
    - Affects prevalence of and reasons for missing values

Example of failing to get to know variables

- In a nationally representative survey sample from a developing country circa 2002.
  - Data set downloaded from a research data web site; not cleaned or evaluated before use.
  - Birth weight in grams observed range up to 9999 with a mean of 8000
- First red flag: Implausible as an actual birth weight, given its meaning and units. 9,999 grams ~= 22 lbs.
  - 9999 was a code for missing value
- Lesson: Must become familiar with what a particular value means for that concept, context and units.

Second red flag

- 2/3 of sample had a birth weight value of 9999
  - Very high value for a substantial share of the sample
  - Unlikely to be explained solely by
    - outliers
    - data entry errors
- Lesson: Look at study documentation and questionnaire to find out why this distribution was observed.
  - Occurred due to a skip pattern designed to minimize recall bias in birth weight reporting.

Resources needed for this exercise

- Documentation on the data source
  - Description of study design
  - Questionnaire
  - Codebook for electronic data file
- Electronic file of database
- Statistical software
- Research question
- Articles, books, etc. on the topic
  - Dependent and key independent variables

Getting to know variables is project-specific

Attributes of data and variables to become familiar with prior to analysis

Analytic sample

- Before becoming acquainted with variables in the analysis, impose any limits on the analytic sample related to the research question.
- Exclude cases
  - to whom the topic does not pertain
  - that are part of a group with too few cases
  - for whom a key variable was not collected

Context of measurement

- When, where, who, e.g., family income will be
  - Higher now than it was 200 years ago in a given place and group
  - Higher in a currently developed than developing country
  - Higher in a sample of all households than in a sample of low-income households

Unit of analysis

- Do data pertain to
  - Individual person?
  - Family?
  - Census tract?
  - Institution?
- Knowing unit of analysis helps ascertain plausible range of values
  - e.g., number of persons in a family will be much lower than the population of a census tract or a school

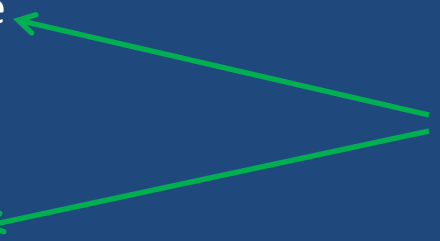
Labeling, coding, and missing value information for the variables

- To help create a comprehensive record of information on each of the variables in the analysis, fill out a grid like this one, which is available online.

Variable name (e.g. acronym on the 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?**
DOCLY	Saw doctor last year	Nominal	1 = yes 2 = no	1, 2	7 = refused 8 = don't know 9 = missing	None for this variable	Original
BWGRMS	Birth weight	Ratio	Grams	0–6000	9999 = missing	Asked only about children < age 5 years.	Original

Level of measurement

- Categorical variables are classified into categories or ranges.
  - Nominal, e.g., gender, race
  - Ordinal, e.g., age group, income range
- Continuous variables
  - Measured in numeric units, but not grouped.
  - Two types of continuous variables:
    - Interval
      - Zero is not lowest possible value
      - e.g., temperature °Fahrenheit
    - Ratio
      - Zero is lowest possible value
      - e.g., temperature °Kelvin, height, weight



Helps to anticipate limits on range of values

Units of measurement

- System of measurement: Metric, British or other?
  - E.g., income in dollars or Euros or yen?
- Level of aggregation
  - E.g., income per hour or per week or per year?
- Scale
  - E.g., income in dollars or thousands of dollars or millions of dollars?

Missing values

- Missing values on a variable can occur because they are
  - Not applicable for some respondents
  - Missing by design (e.g., modules given only to a subset of the overall sample)
  - Item non-response
- Identify missing values as such in the electronic database, so they are treated correctly during analysis.

Plausible values for the concept being measured

A value of 10,000

- Makes sense in at least some contexts for
  - Annual family income in dollars
  - Population of a census tract
  - An annual death rate per 100,000 persons
- Does NOT make sense for
  - Hourly income in dollars
  - Height of a person, in inches
  - Number of persons in a family
  - A Likert scale item
  - A proportion
  - An annual death rate per 1,000 persons

Another example of plausible values

A value of –1

- Makes sense in at least some contexts for
  - Temperature in degrees Fahrenheit or Celsius
  - Change in rating on a 5 point scale
  - Change in death rate per 100,000 persons
  - Percentage change in annual family income
- Does NOT make sense for
  - Temperature in degrees Kelvin
  - Number of persons in a family
  - A Likert scale item
  - A proportion

Becoming acquainted with the concepts under study

- To identify plausible ranges of values for each of the dependent and key independent variables, read the literature.
- Definitional limits
  - E.g., a proportion of a whole must fall between 0 and 1
- Conceptually plausible range
  - E.g., birth weight must be positive but low enough that an infant of that size could conceivably be born!
- Context of measurement (who, when, where)

Descriptive statistics

- After
  - Imposing restrictions on analytic sample
  - Filling in missing value codes for each variable
- Complete a grid of d-statistics on each variable to compare across
  - Analytic data set
  - Codebook
  - Articles or books on the topic

Variable name	#of valid cases for that variable (excl. missing values)	Observed values from data set				Values & range consistent w/ codebook?	Reference values from external source			
		For continuous variables			For categorical variables: Frequency distribution		For continuous variables			For categorical variables: Frequency distribution
		Min	Max	Mean			Min	Max	Mean	

Check each distribution against the codebook for the original source

- Check the distribution of values observed in the analytic sample for each variable against the codebook for the data set.
  - range and/or mean values for continuous variables
  - frequency distribution of categorical variables
  - # cases with missing values, by reason for missing value
- If any distributions are inconsistent, do NOT analyze the data until discrepancies are resolved!

Check each distribution against the literature on similar variables

- Track down information in the published literature on each of the main variables for a similar population.
- If the values in the data are substantially different from those used in other studies of the same concepts, do NOT analyze the data until discrepancies are resolved!

Identify reasons for inconsistencies

- Explain possible reasons for discrepancies between their data and similar data sets, e.g.,:
  - Population studied, e.g., substantially different time, place, and/or subgroup
  - Units of analysis, e.g., family instead of individual
  - Units of measurement, e.g., metric instead of British units
  - Scale, e.g., grams instead of kilograms
  - Transformations of the variables, e.g., percentiles instead of original value

Reasons for getting to know your variables, redux

- These attributes of the analytic sample and variables are essential information for
  - Data preparation
    - Inclusion criteria for the analytic sample
    - Creation of new variables
  - Choice of pertinent descriptive and multivariate statistics
  - Design of correct charts and tables
  - Writing correct prose
- Even experienced researchers should complete this assignment when undertaking a project with a new topic or data set.

Exercise yields key information for a research paper on the topic

- Reading the literature on the topic yields information needed for the
  - introduction
  - literature review
  - discussion sections of a paper
- Detailed knowledge of study design and variables from documentation, questionnaire and codebook provides information needed in the
  - data and methods
  - results sections of a paper

Suggested readings

- Miller, J. E. 2013. The Chicago Guide to Writing about Multivariate Analysis, 2nd Edition.
  - chapter 4 on levels of measurement, units, standards and cutoffs
  - chapters 7 and 10 on choice of contrasts to suit the variable
  - chapter 13 on data and methods
  - chapters 4 and 13 on missing values and missing by design
- Chambliss, Daniel F., and Russell K. Schutt. 2012. Making Sense of the Social World: Methods of Investigation, 4th Edition. Thousand Oaks, CA: Sage Publications, or other research methods book for information on
  - study design, conceptualization, and measurement

Suggested online resources

Suggested podcasts:

- Reporting one number (re: units)
- Comparing two numbers or series of numbers (re: levels of measurement)
- Defining the Goldilocks problem

Online materials available at

http://press.uchicago.edu/books/miller/multivariate/index.html

Contact information

Institute for Health, Health Care Policy and Aging Research

Rutgers University

112 Paterson Street

New Brunswick NJ 08901

jmiller@ifh.rutgers.edu

(848) 932-6730