BASIC DATA MANAGEMENT IN SPSS

(POINT AND CLICK)

By Joni Ricks

(IDRE Stat Consulting)

www.ats.ucla.edu/stat/seminars/special/SPSS_data.pdf

WHAT WE WILL COVER TODAY!

- Opening Excel Files in SPSS
- SPSS orientation and navigation
- Basic Data Management and Data Checking
  - Renaming variables
  - Labeling variables and values
  - Subsetting data
  - Recoding variables
  - Creating new variables
  - Missing information
  - Sorting
  - Keeping and dropping variables
- Saving data and SPSS file types

OUR DATA

- High School and Beyond
- N=200
- 13 Variables
- Student Demographics and Achievement including test scores

OPENING EXCEL FILES IN SPSS

- SPSS can read many data file types
- Excel often used to store data
- A properly formatted Excel file is easily opened in SPSS*

*If you need information on formatting please see:

- http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Excel_file_set_up.htm

HOW TO OPEN AN EXCEL FILE IN SPSS

- **File -> Open Data**
- From **"Files of type"**, select Excel .xls format
- Select File: **c:\...\hs0.xls**
- **Open**
- **Check** -> "Read variable names from the first row of the data"
- **Click** on OK
- Your data set should now open in SPSS

A FEW ITEMS TO NOTE

1) If you do not check **“Read variable names from the first row of the data”**, SPSS will read that first row as true data.

2) OUTPUT window is opened when loading data.
   - OUTPUT window in SPSS prints underlying syntax generated from any procedure.

SYNTAX WINDOW

- All SPSS procedures executed using syntax code
  - point-and-click interface generates and runs syntax for you.
- SPSS syntax files
  - text file that contains the commands
  - Not needed if only using point-and-click
  - Benefit: easy documentation.
  - All SPSS commands end in “.”
  - Comments are preceded by “*” or “Note”
    - Comments are not executed by SPSS
    - For example : *Mean age by treatment group.

DATA EDITOR

- Editor displays opened dataset
- At the bottom of the DATA EDITOR are two tabs:
  1. Data View
  2. Variable View
- Let's explore these

DATA VIEW

- Data displayed as spreadsheet
  - Variables as columns
    - Names on top
  - Records (observations) as rows
- Check that the number of rows match between SPSS and Excel
  - Mismatch may signify loading error

VARIABLE VIEW

- Lists variables and their properties
  1. **Type:** numeric vs string (aka character)
  2. **Label**: label variable name with explanatory text.
     1. Ex. SBP could be labeled “Systolic Blood Pressure”
  3. **Values**: label variable values with explanatory text.
     1. 3=High , 2=medium 1=Low
  4. **Missing**: indicates missing values for that variable.
     - For example, we can tell SPSS “-9” is a missing value for gender.
  5. **Measure:** defines variable scaling
     - Options are scale, nominal or ordinal.
     - Important for analysis stage

THE IMPORTANCE OF RENAME

- Analysis easier when variables have meaningful names.
- Gender a good candidate for renaming
  - Gender currently coded: 0=Males and 1=Females.
- Rename variable to make meaning of “1” obvious
  - Rename gender → female
    - Meaning of “1” now obvious

HOW TO RENAME?

- Go to Data Editor window
  - **Click** on "Variable View" tab
  - Double-click **gender**
  - Cursor should appear
  - Change **gender** to **female**
- Short variable names make cleaner output
  - Do not include special characters/symbols

LABELING VARIABLES

- Good variable names are short and informative
  - For example acronyms for medical terms (e.g. SBP, DBP)
- Sometimes hard to find a pithy name
  - Variable labels permit longer explanations of variables.

HOW TO ADD VARIABLE LABELS?

- Very easy!
- **Click** on "Variable View" tab
  - Bottom of Data Editor
- Type in label in **Label** column
  - For the variable **schtyp**, we can type "Type of school" in the label column.

WHY DO I CARE ABOUT VARIABLE LABELS?

Without Labels

female * schtyp Crosstabulation

Count

		schtyp		Total
		1	2	
female	0	77	14	91
	1	91	18	109
Total		168	32	200

With Labels

Female * Type of School Crosstabulation

Count

		Type of School		Total
		1	2	
Female	0	77	14	91
	1	91	18	109
Total		168	32	200

VALUES LABELS

- Can also label specific values of variables
  - Often done with categorical variables.
  - For example, we can label values of schtyp
    - 1 = public, 2 = private
- Value labels make output much easier to read and interpret.

HOW TO ASSIGN VALUE LABELS?

- Value labels assigned in “Variable View”
- To label a variable’s values:
  - Click on the values cell for the schtyp
  - Once the cell is selected, click on the small box with “…”
  - Type “1” in the Value box and "public" in the Label box, and then click on Add.
  - Do the same for the next value label.
- We will repeat this process for the variable **ses.**

WHY DO I CARE ABOUT VALUE LABELS?

Without labels

ses * Type of School Crosstabulation

Count

		Type of School		
		1	2	Total
ses	1	45	2	47
	2	76	19	95
	3	47	11	58
Total		168	32	200

With labels

Socioeconomic Status * Type of School Crosstabulation

Count

		Type of School		
		Public	Private	Total
Socioeconomic Status	Low	45	2	47
	Middle	76	19	95
	High	47	11	58
Total		168	32	200

VISUALIZING SUBSETS OF OBSERVATIONS

- Let's examine the records of students who earned read scores > 60.
- SPSS calls this **selecting cases**
  - **Data ->Select Cases**
  - **select** "if condition is satisfied" **if** read >= 60
- Creates a SPSS internal filter variable shown in Data View

PERFORMING ANALYSIS ON A SUBSET

Before Subset

Socioeconomic Status * Type of School Crosstabulation

Count

		Type of School		Total
		Public	Private	
Socioeconomic Status	Low	45	2	47
	Middle	76	19	95
	High	47	11	58
Total		168	32	200

After Subset

Socioeconomic Status * Type of School Crosstabulation

Count

		Type of School		Total
		Public	Private	
Socioeconomic Status	Low	7	1	8
	Middle	15	4	19
	High	24	5	29
Total		46	10	56

REMOVING FILTER

- Data
  - Select Cases
    - Choose “All Cases”

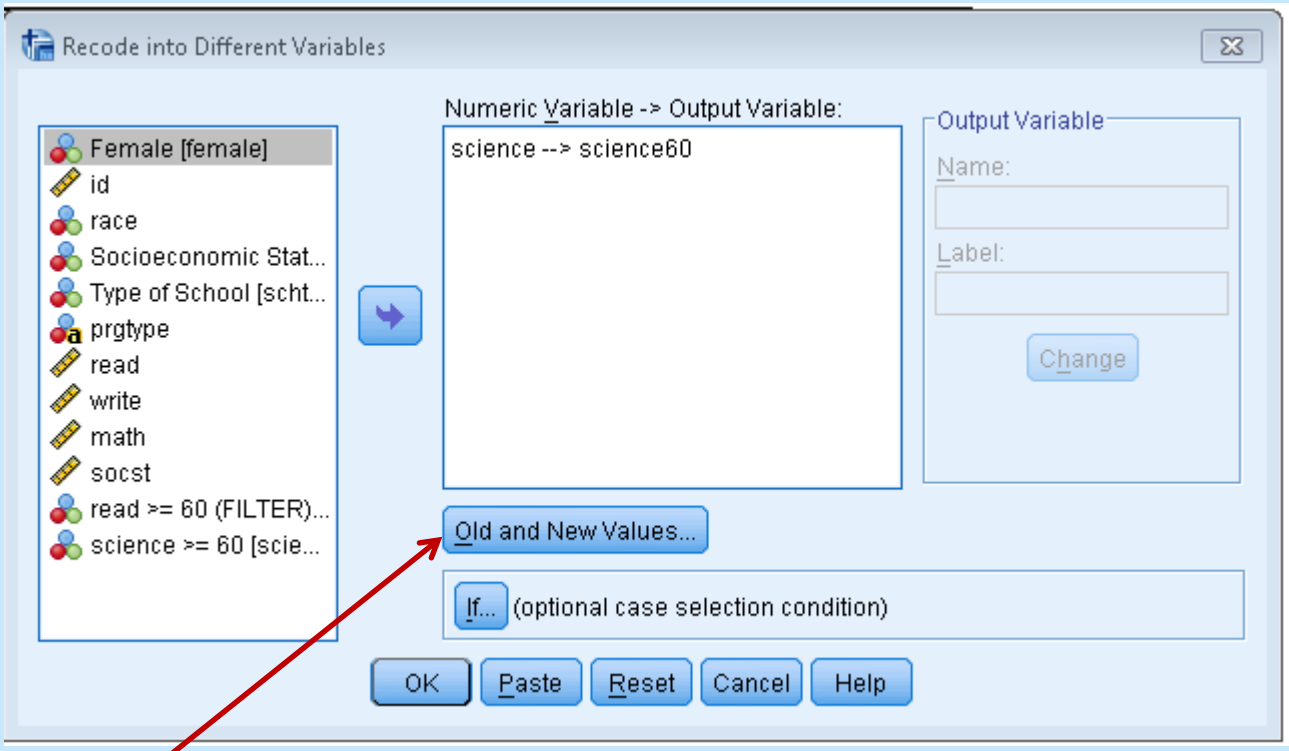
RECODING VARIABLES

- Continuous to categorical
- Collapsing categories
- Renumber categories

HOW TO RECODE A VARIABLES

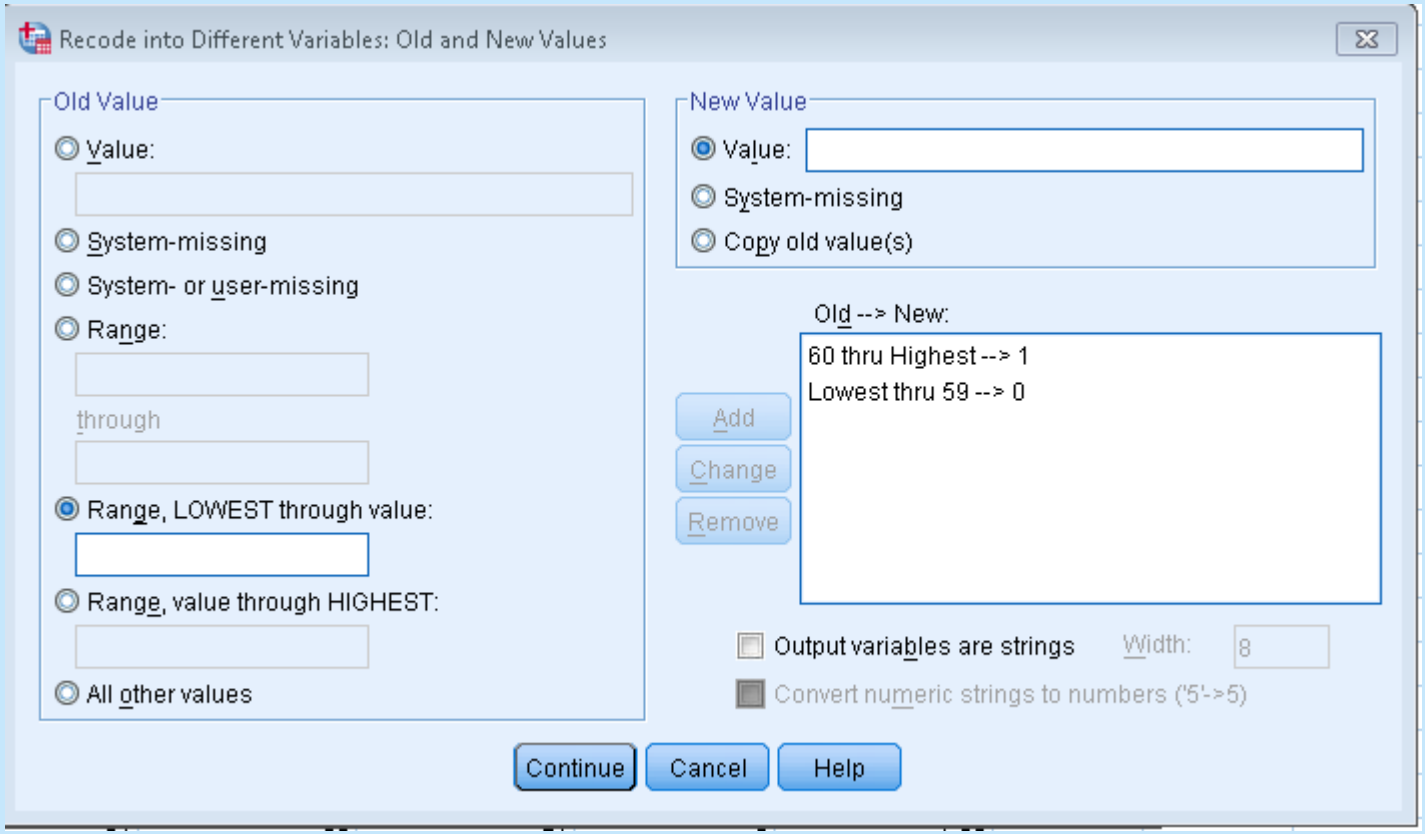
- **Transform…**
- 2 options
  1. **Recode Into different variables…**
     - Creates a new variable (recommended)
  2. **Recode Into same variables..**
     - Over-writes the current variable

SELECT VARIABLE FROM DATA AND NAME/LABEL NEW VARIABLE



Click on "Old and New Values"

PROVIDE OLD VALUE OR RANGE TO BE RECODED AND NEW VALUE



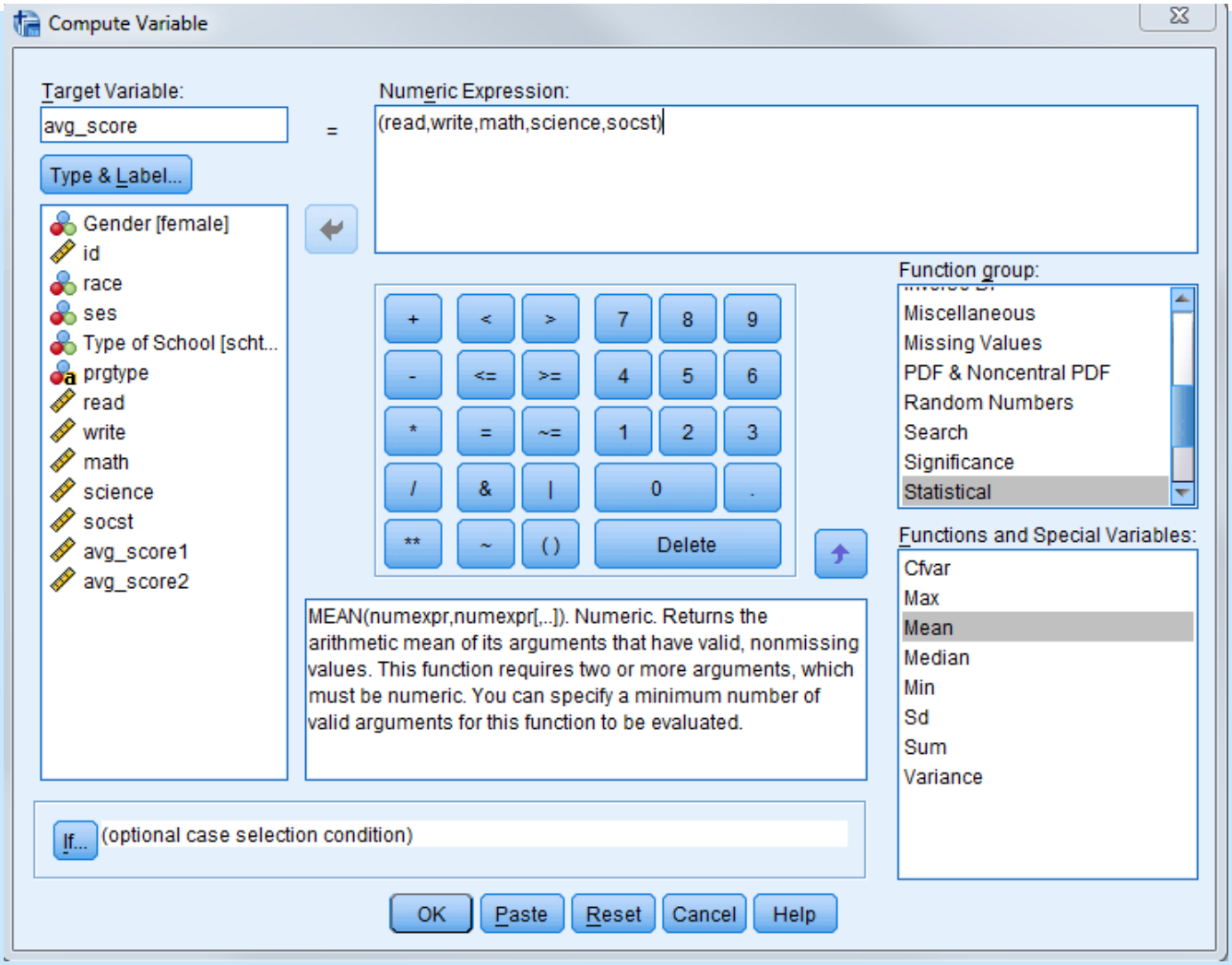
DATA CHECK: MAKE SURE NEW VARIABLE WAS RECODED CORRECTLY

- ALWAYS check variables after data manipulation.
- A simple way we would be to check our new variable through a frequency table.

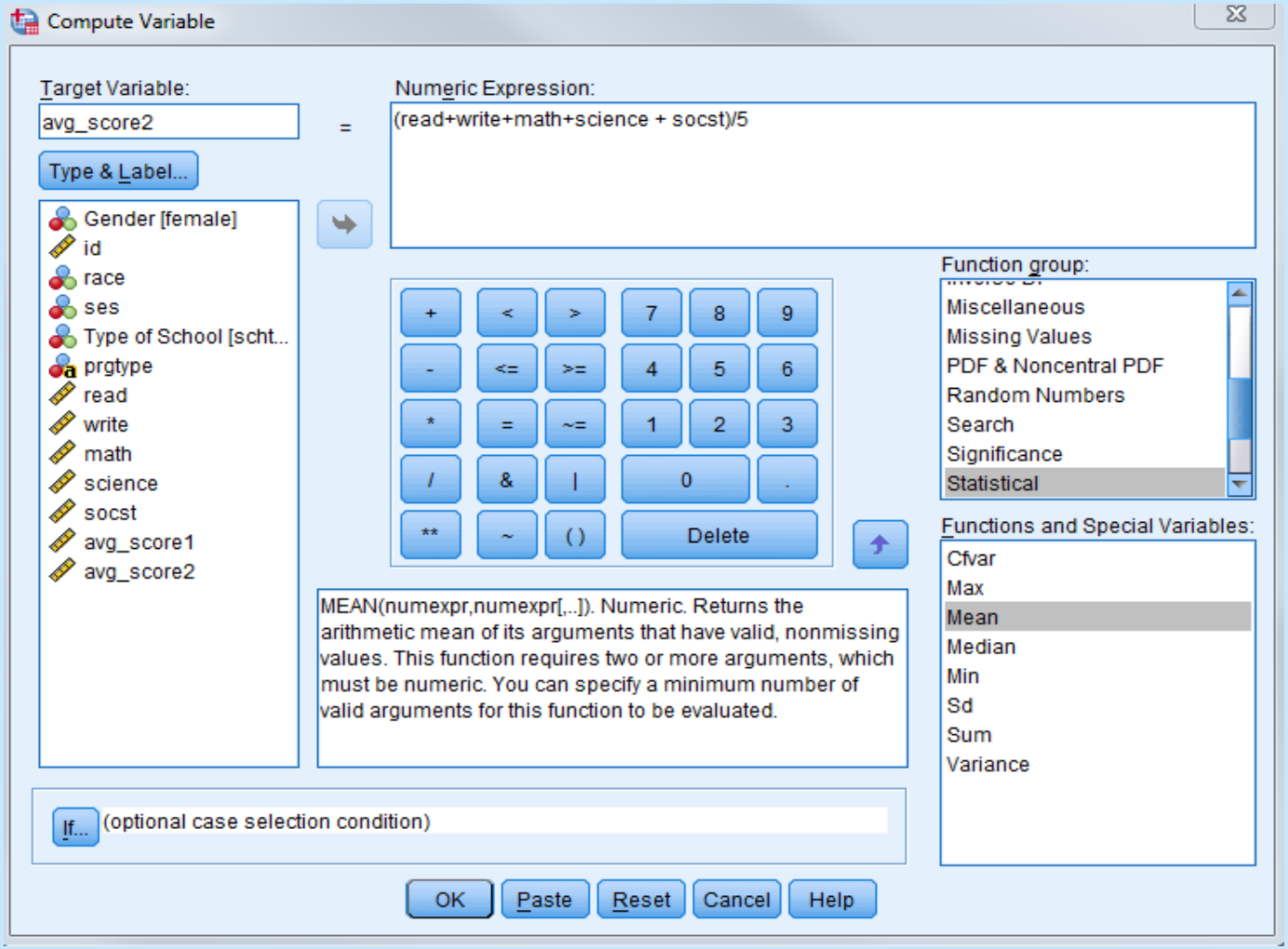
HOW TO CREATE NEW VARIABLES?

- **Transform**
- **Compute…**
  - Creates variables through mathematical functions or standard arithmetic

CREATE AN AVERAGE SCORE VARIABLES USING SPSS FUNCTION



CREATE AN AVERAGE SCORE VARIABLE USING ARITHMETIC



DATA CHECK: MAKE SURE NEW VARIABLE WAS CREATED CORRECTLY

- Let's use "Descriptives" to check the mean, min, max and standard deviation for the variables we just created.
- Without filtering, SPSS will run descriptives on all available cases

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
avg_score1	200	33.00	68.50	52.3732	8.17543
avg_score2	195	33.00	68.00	52.2379	8.08136
Valid N (listwise)	195				

- It appears that we generated some missing values
- "Valid N(listwise)" denotes that only 195 our of 200 observations have valid, non-missing values on both variables.

DEALING WITH MISSING VALUES IN SPSS

- Two types of missing value indicators in SPSS

1. **System-missing**

2. **User-defined**

SYSTEM MISSING

- These are values automatically recognized as missing by SPSS.
  - Represented by a “.”
  - Note: Blank cells in Excel are read in as system-missing

USER-DEFINED MISSING

- Numeric values chosen by user to denote missing
  - Must be defined as missing for SPSS.
    - Define them in **Missing** column of **Variable View**
  - Common user-defined values are -8, -9, -99, -9999, etc.

WHY DO I CARE?
USER-DEFINED MISSING VALUES
EFFECT ON VARIABLE CREATION

Without
user-defined missing

race

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-9	1	.5	.5	.5
	-8	1	.5	.5	1.0
	1	24	12.0	12.0	13.0
	2	11	5.5	5.5	18.5
	3	20	10.0	10.0	28.5
	4	141	70.5	70.5	99.0
	5	2	1.0	1.0	100.0
	Total	200	100.0	100.0	

With
User-defined missing

race

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	24	12.0	12.1	12.1
	2	11	5.5	5.6	17.7
	3	20	10.0	10.1	27.8
	4	141	70.5	71.2	99.0
	5	2	1.0	1.0	100.0
	Total	198	99.0	100.0	
Missing	-9	1	.5		
	-8	1	.5		
	Total	2	1.0		
Total		200	100.0		

SORTING DATA

- What if we want our records listed by ID, from 1 to 200.

- **Data->Sort Cases**
  - Choose Variable(s) you want to **Sort by**
  - Choose **Sort Order**
  - Have option to save the new sorted file (optional)
  - Paste Syntax (optional)

KEEP/DROP VARIABLES

- Perhaps you only want specific variables included in your dataset?

- Two ways of dropping variables in SPSS
  1. In **Variable View,** click on the row number of the variable you want to drop.
     - Then right click and choose **Clear.**

  2. Use syntax to save your file with a /**drop** or /**keep** subcommand.

LAST THING

- Don’t forget to save your data!
- File extensions
  - Data Editor (.sav files)
  - Output Viewer (.spv files)
  - Syntax Editor (.sps files)

ADDITIONAL RESOURCES

- Our website: http://www.ats.ucla.edu/stat/
- Take a look at the links:
  - Introduction to SPSS seminar:
    - http://www.ats.ucla.edu/stat/spss/notes/default_22.htm
  - Learning Modules
    - http://www.ats.ucla.edu/stat/spss/modules/default.htm
- Help with Data set-up in Excel
  - http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Excel_file_set_up.htm