

Stata Workshop 2: Data Manipulation and Visualization

Tao Wang, SSQ @ Swarthmore, swarthmore.edu/ssq

1. Workshop Objectives

This workshop builds on the skills from our first session. By the end of this workshop, you will be able to:

- Label variables and values to create clear, professional outputs.
 - Import, clean, and prepare a dataset for analysis.
 - Use powerful commands like `codebook`, `inspect` and `tabulate` to explore variables.
 - Create new variables from existing data using `generate`, `replace` and `egen`.
 - Visualize the distribution of data with histograms, bar charts, and pie charts.
 - Use if qualifiers to perform operations on a subset of your data.
-

2. Commands to be Mastered

Data Labeling & Management

- `label variable varname "label text"`: Attaches a descriptive label to a variable.
- `label define labelname # "text"`: Creates a set of labels for numeric values.
- `label values varname labelname`: Applies a defined set of labels to a variable.
- `rename old_varname new_varname`: Renames a variable.
- `drop varlist`: Deletes specified variables from the dataset.
- `order varlist`: Re-arranges the order of variables in the dataset.

Descriptive Statistics & Exploration

- `codebook varname`: Provides a detailed report on a variable's properties.
- `inspect varname`: Gives a quick summary and text-based histogram of a variable.
- `count`: Reports the number of observations in the dataset.
- `tabulate cat_variable`: Creates a frequency table for a categorical variable.

Variable Creation

- `generate newvar = expression`: Creates a new variable.
- `replace varname = expression if condition`: Modifies the values of an existing variable, but only for observations that meet the if condition.

Graphics

- `histogram discrete_variable, discrete`: Creates a histogram for a discrete variable.
 - `graph bar, over(cat_variable)`: Creates a bar chart for a categorical variable.
 - `graph pie, over(cat_variable)`: Creates a pie chart for a categorical variable.
 - `twoway (scatter y_var x_var) (lfit y_var x_var)`: Creates a scatter plot with a linearly fitted line.
-

3. Workshop Exercises

Exercise 1:

Find the documentation for the variable `etotapx4` on the BLS website. Rename and label the variable accordingly.

```
rename etotapx4 outlay
```

```
label variable outlay "Total family outlay in the previous quarter"
```

The documentation for the variable can be located in the [Dictionary for Interview and Diary Surveys \(XLSX\)](#).

Exercise 2:

Create a new variable that represents the quintile of the family's income, and label the variable and values appropriately.

```
egen q_inc = cut(inc), group(5)
```

```
xtile q_inc = inc, nq(5)
```

Either command can be used to create an indicator for quintiles. However they do not always produce the same results due to different treatment of values near cutoffs. Note also that the numerical values assigned by the two commands for each quintile is off by 1.

Use the “Variables Manager” or the `label` commands to define and assign value labels.

Exercise 3:

Use appropriately numerical or graphical analysis to investigate the relationship between poverty rate and family size.

```
corr size poverty
```

A stacked bar chart or scatter plot can also be used.