

# Stata Workshop 1: Introduction to Stata

Tao Wang, SSQ @ Swarthmore, [swarthmore.edu/ssq](http://swarthmore.edu/ssq)

---

## 1. Workshop Objectives

Welcome! By the end of this workshop, you will be able to:

- Have a running copy of Stata on your personal computer.
  - Understand the function of Stata's main windows and file types (.dta, .do, log files).
  - Set a working directory and create a log file to document your work.
  - Import data into Stata from a spreadsheet (.csv file).
  - Inspect your data using descriptive commands.
  - Calculate summary statistics.
  - Create a new variable based on a mathematical expression.
  - Create a histogram to visualize the distribution of a variable.
  - Calculate the linear correlation coefficient between two variables.
  - Create a simple scatterplot to visualize data.
  - Save your commands in a do-file for future use.
- 

## 2. Commands to be Mastered

### Workflow & File Management

- `cd "path"`: Change directory to your project folder.
- `log using "filename", replace`: Starts a log file to record commands and output.
- `clear`: Clears any data from Stata's memory.
- `help commandname`: Shows the help file for any command.

### Data Input & Output

- `import delimited "filename", options`: Imports data from an Excel file.
- `use "filename"`: Loads a Stata .dta data file.
- `save "filename", replace`: Saves the current data as a Stata .dta file.

### Data Inspection & Analysis

- `describe`: Shows a summary of your dataset (variable names, types, etc.).
- `browse`: Opens the Data Editor in a read-only "browse" mode.
- `summarize`: Calculates key summary statistics (mean, std. dev., min, max).
- `summarize varname, detail`: Provides detailed summary statistics for a variable.

- `correlate var1 var2`: Reports linear correlation for two variables.
- `generate varname = expression`: Generate a new variable.

## Graphics

- `Histogram varname`: Creates a simple histogram.
  - `twoway (scatter y_var x_var)`: Creates a simple scatterplot
- 

## 3. Workshop Exercises

Exercise 1:

Show the summary statistics for consumption. Find the mean, variance, standard deviation, and the five number summary. What kind of variable is consumption?

```
summarize con, detail  
su c, d
```

Exercise 2:

Generate a new variable that represents the natural log of disposable income, and create a histogram of the variable. Describe the shape of the distribution.

```
gen logdinc = log(dinc)  
hist logdi
```

Exercise 3:

Find the linear correlation coefficient between log disposable income and log consumption. Create a scatter plot to show the relationship.

```
corr logdi logc  
scatter logc logdi
```