Introduction to Commands, Basic Descriptive Statistics

Stata Commands

- Stata is organized around built-in commands
- Commands are “verbs” that perform an action
  - Actions that manipulate data
  - Actions that analyze data
  - Actions that create graphics
- Advanced users can write their own commands and share them with others (.ado files)

Command Structure

- Commands, like verbs, have different syntax and only work properly in certain contexts:
  - Some can be used without objects (sleep or **`describe`**)
  - Some must have one of more objects (give or **`label`**)
  - Some really only work with objects as well as options (take umbrage or **`encode`**)

Command Basics

```
[by] command variable [if] [in], options
```

- This is the basic syntax of most Stata commands
- Stata commands have this built-in syntax to easily allow for the most common ways you might want to manipulate or analyze your data
- We will take a look at each one of these components one at a time

Command `[if]`

```
command variable if expression
```

- The if qualifier allows you to perform a command on a subset of your observations defined by the expression
- Adding an if is optional, not necessary. Without it, the command will be performed on all observations of a variable in the dataset
- If expressions are often used to define new variables or modifying existing variables, but could also be used to present analyses of subsets

Command `[if]`

- **An example of using if in a variable generation step:**

```
generate great_headroom = 1 if headroom > 3
```

```
replace great_headroom = 0 if headroom <= 3
```

- This defines a new variable `great_headroom` as 1 if the headroom variable is greater than 3 and 0 if the headroom variable is less than or equal to 3

- **An example of using if to perform an analysis on a subset of observations:**

```
summarize price if headroom > 3
```

- This performs and outputs the summarize command on only observations whose headroom variable is greater than 3

Command `[in]`

```
command variable in indices
```

- The in qualifier allows you to perform a command on a subset of observations based on their indices

- Adding an in is optional, and is most often used alongside a list command to take a look at certain low or high values:

```
list price weight in 1/5
```

- Remember that this indexing can change depending on how the observations are sorted — it is good practice to only use in after an explicit **`sort`** command

Command `by`

```
by variable_name: command ...
```

```
bysort variable_name: command ...
```

- The **by** prefix command allows you to perform stratified commands across values of a `variable_name`
- If the data is not sorted by the variable_name, an error will usually occur. To automatically sort, use the **`bysort`** command

```
bysort foreign: summarize mpg
```

- This will give summaries of the `mpg` variable stratified by the foreign `variable`
- This prefix works with continuous and categorical variables (but only really makes sense with categorical ones)

Command `options`

`command` `. . . , options`

- Almost all commands have options that allow the user to alter the performance of the command, display less or more detailed results of a command, or override regular Stata behavior

- Options are often unique to a command, but here are a couple common ones:

`command` `. ,` `replace` - overwrites the current file / variable

`command` `. ,` `clear` - clears away old data when loading or reading in files

`command` `. ,` `gen(`*`newvar`*`)` - uses the output of the command to create a new variable with name `newvar`

`command` `. ,` `detail` - prints more detailed output of a command

Explore Your `options`

- If you are wondering if you can do something in Stata, the best way to find out is through exploring the `options` in the **`help`** documentation for a command that is *close* to what you want to do

- Let's explore some basic statistical commands

summarize

```
summarize [variable_name]
```

- This command displays summary statistics for a variable (or all variables in a data set)

```
summarize variable_name, detail
```

- Provides more detailed summary statistics on a variable

```
bysort variable_name: summarize variable_name, detail
```

- Creates detailed stratified summary statistics for a variable

correlate

```
correlate variable_name1 variable_name2 ...
```

- This command calculates the correlation (or correlation matrix) between variables
- It needs at least two `variable_names` to work

```
pwcorr [variable_name1 variable_name2 ...]
```

- With no specifications, it will create a pairwise correlation matrix for whole dataset

```
pwcorr [variable_name1 variable_name2 ...], sig
```

- The sig option also calculates the significance of a correlation

tabulate

```
tabulate variable_name1 variable_name2
```

- This command will create a one-way or two-way table of values (depending on the number of `variable_names` given)

- A very commonly used command in epidemiology and a good first step to check on cell size for analysis

```
tabulate foreign great_headroom if price < 7000
```

- Using the if qualifier we can look at a table of a subset of observations

tabstat

```
tabstat variable_name1 variable_name2 ...
```

- This command creates a very customizable table of summary statistics for variables in a dataset

- Using `help tabstat` and clicking on the statistics options we can look through all the possible ways to build up a table

ameans

```
ameans variable_name1 ...
```

- This command creates a table of pythagorean means with confidence intervals

- Is there a difference between **ameans**, **gmeans**, **hmeans**? How can we check?

Exercises (1)

1. Titanic Data

   A. Open up your titanic.do file and run it, but change it so that it keeps passengers of all ages and survival statuses.

   B. Create a new categorical age variable over_30. Observations with an age over 30 should be assigned 1, those under 30 should be given a 0. Watch out for missing!

   C. Give your new variable and its values appropriate labels.

   D. Create a 2x2 table of over_30 variable and the survival value. Create a note for the over_30 variable which indicates how many people over 30 survived the titanic.

   E. In one command, have Stata find the mean ages of people who survived and people who did not survive. Add this information as a note to the survived variable.

   F. What is the sex variable's type? Create a new variable that can be used by Stata commands.

   G. Record the number of females who did not survive as a note in your new sex variable

Exercises (2)

1. Movie Metadata

    A. Open up your movies.do file and run it.

    B. Create a table that reports the mean, count, 25th percentile, 75 percentile, and range of all the continuous variables in the dataset. (remember **help**)

    C. Explore if any of the continuous variables are correlated, include their statistical significance.

    D. Re-create part B, but perform the command across categories of the expensive variable.