

# Introduction to Stata

Adam Okulicz-Kozaryn  
Harvard-MIT Data Center (HMDC)

Brought to you by:  
**MIT Libraries' Social Science Data Services**  
<http://libraries.mit.edu/guides/subjects/data>

# Outline

Preliminaries

Stata Basics

Import/Export

Labels

Do-Files

Variables Manipulations

Data Description

Missing Values

## Assumptions and Disclaimers

- ◇ This is **Introduction** to Stata
- ◇ Assumes no/very little knowledge of Stata
- ◇ Not appropriate for people already well familiar with Stata
- ◇ Computer paths pertain to default lab setup; If you have laptop adjust paths accordingly
- ◇ Your level of knowledge will differ from the mean – If you are ahead of time experiment with command features described in help files

# Outline

Preliminaries

Stata Basics

Import/Export

Labels

Do-Files

Variables Manipulations

Data Description

Missing Values

## Class Website

- ◇ [http://stathelp.iq.harvard.edu/stata\\_intro](http://stathelp.iq.harvard.edu/stata_intro)
- ◇ More detailed information
- ◇ Good for self-study
- ◇ More advanced topics
- ◇ Links to resources

## Preliminaries

- ◇ Feel free to interrupt, especially if lost
- ◇ Learn how things work and how to get help
- ◇ Share code and use others code  
(Learn by example !)
- ◇ My replication code is available on class website
- ◇ The goal is *\*not\** to memorize commands

## Statistics is the Future !

“I keep saying that the sexy job in the next 10 years will be statisticians.”

NYT: Hal Varian, chief economist at Google

<http://www.nytimes.com/2009/08/06/technology/06stats.html>

- ◇ More and more data, e.g. surveys, blogs, twitter
- ◇ Academia more quantitative, e.g. pol sci
- ◇ Industry more quantitative, e.g. google
- ◇ In fact, even qualitative data (pictures, text, etc.) is rich quantitative data and we can analyze it as quantitative data. In fact, everything can be quantified. Any examples of non-quantifiable things ?

## Why Stata

- ◇ Powerful. No need to learn any other software; Sufficient for vast majority of projects: data analysis, data management and graphics.
- ◇ User friendly (Good GUI, Built-In Documentation)
- ◇ Great user community: Listserv, websites, etc.
- ◇ Reasonable cost



## Why Stata (subjective)



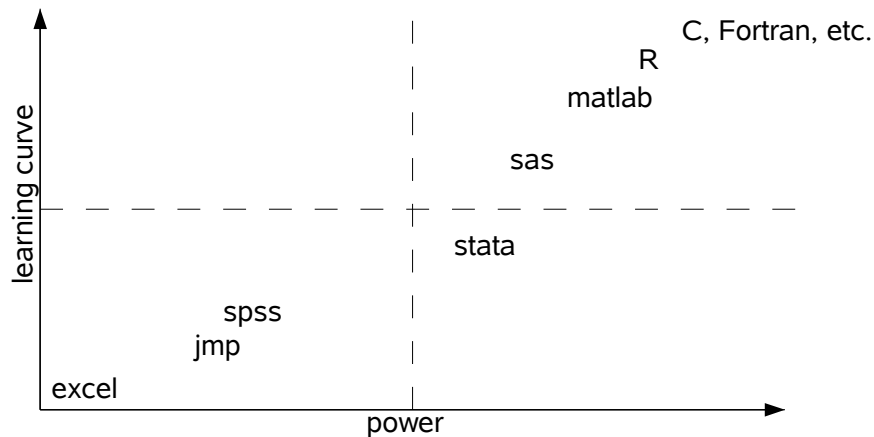
sas

spss



R

## Why Stata (subjective)



## Which Stata

<b>Stata Editions</b>	<b># observations</b>	<b># variables</b>
Small(Student version)	1,000	99
Intercooled (Standard version)	Based on RAM in your computer	2,047
SE (For large datasets)		32,767
MP (Multi-processor)		32,767

- ◇ Most people need Stata-IC (Intercooled)
- ◇ Small Stata is useless !

## How Do I Get Stata ?

- ◇ Athena (you use it now)
- ◇ Your Department IT
- ◇ HMDC Labs
- ◇ RCE (Research Computing Environment)
- ◇ Buy it: educational or grad plan. Again, IC is usually what you want <http://www.stata.com/order/new/edu/gradplans/gp-campus.html>

# Outline

Preliminaries

Stata Basics

Import/Export

Labels

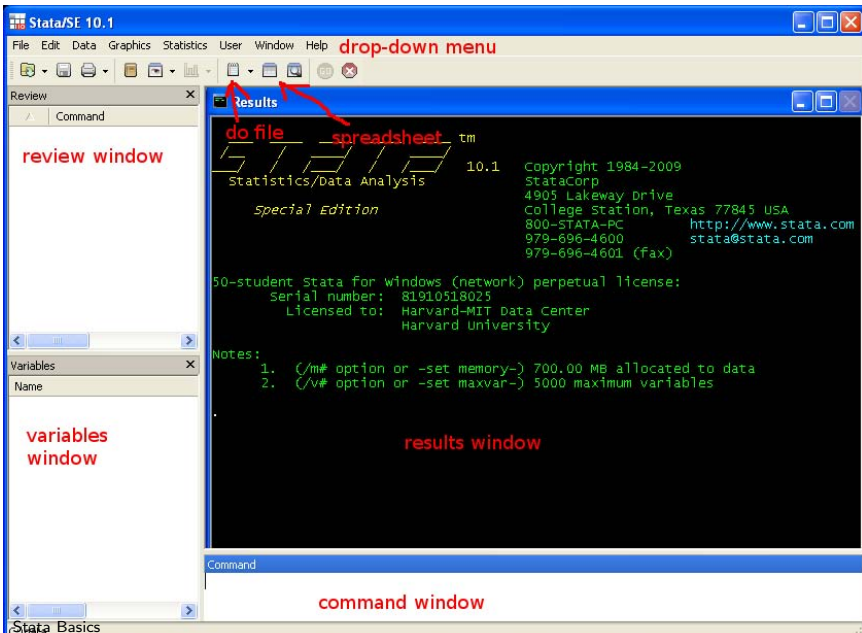
Do-Files

Variables Manipulations

Data Description

Missing Values

# Stata Interface



## Exercise 0: Data for Today

- ◇ Find class materials [http://stathelp.iq.harvard.edu/stata\\_intro](http://stathelp.iq.harvard.edu/stata_intro) You are MIT, **not** Harvard
- ◇ Right-click, Save Link As, and put in your home directory.  
Mine is:  
`/afs/athena.mit.edu/user/a/k/akozaryn`  
Yours will be similar except that instead “akozaryn” there will be your ID at the end; Or just click “home icon”.

## Exercise 0: Data for Today

- ◇ Then start xstata, by typing in terminal

```
add stata
```

```
xstata
```

- ◇ Check where you are; Type in Stata

```
pwd
```

- ◇ Check what you have

```
ls
```



## Exercise 0: Data for Today

- ◇ You should see there **Mit\_stata\_intro.zip**; If not try `cd ~` and then `ls`  
If you still do not see **Mit\_stata\_intro.zip** raise your hand
- ◇ Unzip files by typing in Stata `sh(unzip Mit_stata_intro.zip)`  
Raise your hand if Stata says something bad:  
“unzip: cannot <do something>”

# Getting Help

## ◇ Stata Help Files

- . **help** if you know command name, e.g. **help regress** [useful]
- . **search** if you do not know , e.g. **search regression** [not useful]

## ◇ Built-in pdf documentation

## ◇ Do web search e.g. "stata, dummy variables" [very useful]

# Stata Command Syntax

- ◇ `<command> <variables> , <options>`  
`describe var1 var2, detail`
- ◇ `<variables>` and `<options>` are optional
- ◇ Command specific syntax is in help files,  
e.g. `help describe`

## Tips

- ◇ Make sure you have enough memory when you start stata  
`set mem 500m, perm`
- ◇ Use drop-down menus instead of command line to run Stata if you are a beginner. It will still produce code.
- ◇ Learn abbreviations, e.g. `d` for `describe`, they are underlined in help files
- ◇ Press Page-UP to get previous command in Command Window

## Data for Today

- ◇ Data we use is a subset of General Social Survey:  
<http://www.norc.org/GSS+Website/>
- ◇ Probably the most comprehensive social science data for the U.S.
- ◇ It is very exciting data set
- ◇ We will look today at income, education and gender across U.S. regions

# Outline

Preliminaries

Stata Basics

**Import/Export**

Labels

Do-Files

Variables Manipulations

Data Description

Missing Values

# Paths

- ◇ To import data you need path
- ◇ To get path right-click file

## Importing Stata Data files .dta (Windows)

- ◇ Safe to put path in quotes. Use “clear” in case there is already data in memory  
use “C:\files\gss.dta”, clear
- ◇ Note “Review” and “Variables” Windows



## Importing Stata Data files .dta (Linux)

- ◇ To import/export data :
- ◇ Change dir to where are the files

```
cd ~/files
```

- ◇ See where you are

```
pwd
```

- ◇ See what you have

```
dir or ls
```

- ◇ No need for quotes if no spaces

```
use gss.dta, clear
```

## Exporting Stata Data files .dta

- ◇ Use “replace” in case there is old version of this file on hard drive; replace will not prompt if the file exists

```
save mydata.dta, replace
```

- ◇ To maintain compatibility with <Stata10

```
saveold mydata.dta, replace
```

## Text File Types

- ◇ Data often comes as text file. E.g. **.tab .csv .dat .raw .txt**
- ◇ **.tab** is TAB delimited file
- ◇ **.csv** is Comma Separated Values file
- ◇ But do not trust suffixes
- ◇ Check yourself by opening file with text editor, such as **Stata do-file editor** – if it opens in text editor it is... a text file

## Delimited, ASCII (text file)

- ◇ Stata will usually figure delimiter out
- ◇ Assuming it is in current directory:

```
insheet using gss.csv, clear
```

```
insheet using gss.tab, clear
```

```
outsheet using mydata.csv, replace comma
```

## Fixed Format, ASCII (text file) [extra]

- ◇ **.raw, .dat, ...** They will either tell you or open it in text editor and figure yourself
- ◇ You need a dictionary that specifies variables columns
- ◇ There are several ways to do it...
- ◇

```
infix rate 1-4 speed 6-7 str country 9-11 using highway.raw
```

## Import/Export Tips

- ◇ Use the following commands often:
- ◇ `d`
- ◇ `sum`
- ◇ `edit`
- ◇ `list` (Stata will list variables; Press “–more–” to get more or green arrow in menu. Press red cross in menu to break)

## Import/Export Tips Cont'd

- ◇ Use GUI: File-Open/Import/Export
- ◇ Copy-Paste between Excel and Stata Data Editor
- ◇ Use Stat-Transfer
- ◇ Let's do Exercise 1

# Outline

Preliminaries

Stata Basics

Import/Export

Labels

Do-Files

Variables Manipulations

Data Description

Missing Values



## Variable Names

- ◇ Are we in the right directory?

```
pwd
```

- ◇ insheet using gss.csv, clear

- ◇ Ugly

```
d
```

- ◇ rename v1 hh\_inc

- ◇ Nice

```
d
```

## Variable Labels

- ◇ `label var hh_inc "household income"`
- ◇ `d`
- ◇ You can search labels; useful  
`lookfor income`
- ◇ There are also **value labels** – labels of values that a variable takes on – we will talk about them in data management class

## Tips

- ◇ Give variables short names
- ◇ Labels prevent confusion later and for other people
- ◇ Labels automatically appear on graphs, regressions, etc.
- ◇ Use `lookfor` if you have many variables
- ◇ Let's do Exercise 2

# 10 Minutes Break

# Outline

Preliminaries

Stata Basics

Import/Export

Labels

Do-Files

Variables Manipulations

Data Description

Missing Values

## Research Philosophy

- ◇ Replication is **necessary** for Science  
Scientific results need to be documented
  - . People make mistakes
  - . People forget
  - . People lie
- ◇ Other scientist should be able to replicate your results.  
You too

## Implications of Research Philosophy

- ◇ GUI and Command Window OK for playing around
- ◇ Copy-paste from review window or from results window to do-file
- ◇ By saving commands in do-file you document results
- ◇ Do-file should contain *\*all\** (correct) commands you executed
- ◇ Do-file should produce final results (e.g. regression results) from raw data (e.g. data you downloaded)

## Do-File Basics

- ◇ Do-File is just a text file (**.do**) containing commands
- ◇ Let's close Stata and open it again
- ◇ Click “New do-file editor” icon
- ◇ New window pops up. File-Open... and open stata\_intro.do
- ◇ It has all the code we used and will use today
- ◇ Note the preamble and comments
- ◇ Highlight code you want to run and press Ctrl-D



## Do-File Basics Cont'd

- ◇ You can have several do-files opened at the same time: In do-file editor: File-New
- ◇ You can copy-paste between do-file editor and command window, review window and results window
- ◇ To save do-file, go to File-Save As...
- ◇ You can open do-file with Stata do-file editor as well as with any other text editor (e.g. Notepad)

## Do-File Tips

- ◇ Always have preamble in do-file as in our example
- ◇ Use comments !

```
*comment
```

```
/*comment
```

```
block*/
```

# Outline

Preliminaries

Stata Basics

Import/Export

Labels

Do-Files

**Variables Manipulations**

Data Description

Missing Values

# Operators

- ◇ == equal to (status quo)
- ◇ = used for assigning values
- ◇ != not equal to
- ◇ > greater than
- ◇ >= greater than or equal to
- ◇ & and
- ◇ | or
- ◇ replace hi\_ses=1 if (educ==7 | y==10) & inc>=10
- ◇ Let's have a look at the do-file

## Tips

- ◇ Beware of missing values: Come to our Data Management Class
- ◇ Understand your data: level of measurement, coding
- ◇ Use often: `d`; `sum`; `edit`; `tab` and `tab, nola`
- ◇ Use `lookfor`, especially if you have many variables
- ◇ Let's do Exercise 3

# Outline

Preliminaries

Stata Basics

Import/Export

Labels

Do-Files

Variables Manipulations

Data Description

Missing Values

# Fun

- ◇ This is where fun begins
- ◇ We may use data to answer interesting questions, e.g.:
- ◇ Do women make less than men ?
- ◇ Is the income gap bigger in North-East than in South ?
- ◇ Does education really help with income ?
- ◇ At home go to <http://www.norc.org/GSS+Website/> and use full GSS dataset

## Descriptive Statistics

- ◇ Do you understand what a variable is describing ? For instance, variable 'education' may measure years of schooling or highest degree obtained on scale from 1 to 4
- ◇ Measurement ? Is income in \$ or thousands of \$?
- ◇ Does it make sense ? Can a person be -9 years old?
- ◇ What are the implications for your statistical analysis? (Number of observations, missing values, etc.)
- ◇ Let's see do-file



## Tips

- ◇ `tab` is Stata workhorse; See `help tab` for useful options
- ◇ Also see GUI: Statistics–Summaries, Tables and Tests

## Tips Cont'd

- ◇ Again, use often: `d`; `sum`; `edit`; `tab` and `tab, nola`
- ◇ Do not do inferential statistics (e.g. regressions) before doing descriptive statistics (e.g. histograms, scatterplots, frequency tables and cross-tabs)
- ◇ Let's do exercise 4

## More Information

For further information see: our class website

[http://stathelp.iq.harvard.edu/stata\\_intro](http://stathelp.iq.harvard.edu/stata_intro)

and especially this section:

[http:](http://stathelp.iq.harvard.edu/stata_intro#Extras)

[//stathelp.iq.harvard.edu/stata\\_intro#Extras](http://stathelp.iq.harvard.edu/stata_intro#Extras)

- ▶ Paper replication code
- ▶ Stata useful commands
- ▶ Software comparison
- ▶ And much more...

# Outline

Preliminaries

Stata Basics

Import/Export

Labels

Do-Files

Variables Manipulations

Data Description

Missing Values

## Missing Values

- ◇ Most data sets have missing values
- ◇ Missing value is blank or empty value
- ◇ We have no information for a particular observation
- ◇ For instance, a person declined to report his income
- ◇ Missing value is NOT 0; e.g. if income is 0 it is not missing: we have information that a person does not have income
- ◇ If it is missing we do not know
- ◇ Stata labels missing as ".", or ".a", ".b", etc.

## Missing Values

- ◇ Let's load data with missing values  
`use gss_missing.dta`
- ◇ Tabulate income  
`tab inc`
- ◇ Use "mi" option to see the missing values  
`tab inc, mi`
- ◇ **Always** use "mi" option with tabulate
- ◇ You will also see missings in the spreadsheet  
`edit`

## Missing Values

- ◇ Stata treats missings as a very big number
- ◇ For instance, if income is coded from 1 to 26 and we generate high income, this is **wrong**:
- ◇ `gen hi_inc=0`
- ◇ `replace hi_inc=1 if inc>15` it would be 1 for >15 and for missing
- ◇ It should be:
- ◇ `gen hi_inc=.`
- ◇ `replace hi_inc=1 if inc>15 & hi_inc<26`
- ◇ `replace hi_inc=0 if inc>0 & hi_inc<16`

# Thank You !

- ◇ Please fill evaluations AND give us some comments/feedback – we do care for these classes and want to make them better
- ◇ Come to other classes we offer and tell your friends about our classes

[http://www.iq.harvard.edu/statistical\\_software\\_2009\\_2010](http://www.iq.harvard.edu/statistical_software_2009_2010)



## A Word From Our Sponsor !

- ▶ MIT's membership in HMDC provided by schools and departments at MIT
- ▶ Institute for Quantitative Social Science  
<http://iq.harvard.edu>
- ▶ Data Collection, Management  
[http://www.iq.harvard.edu/data\\_collection\\_management\\_analysis](http://www.iq.harvard.edu/data_collection_management_analysis)
- ▶ Research Computing [http://www.iq.harvard.edu/research\\_computing](http://www.iq.harvard.edu/research_computing)
- ▶ Computer Labs (Software, Books)  
<http://www.iq.harvard.edu/facilities>
- ▶ Training <http://www.iq.harvard.edu/training>