Exploring Data and Descriptive Statistics (using Stata)

Oscar Torres-Reyna

Data Consultant

otorres@princeton.edu

Agenda…

- What is Stata
- Transferring data to Stata
- Excel to Stata
- Exercise 1: Data from ICPSR using the *Online Learning Center.*
- Exercise 2: Data from the *World Development Indicators & Global Development Finance* from the World Bank

Basic commands (review)

- Stata's screen
- First steps (working directory, log file, memory setting)
- Frequencies
- Crosstabulations
- Scatterplots/Histograms

This document is created from the following:
http://dss.princeton.edu/training/StataTutorial.pdf
http://dss.princeton.edu/training/DataPrep101.pdf
http://dss.princeton.edu/training/FindingData101.pdf

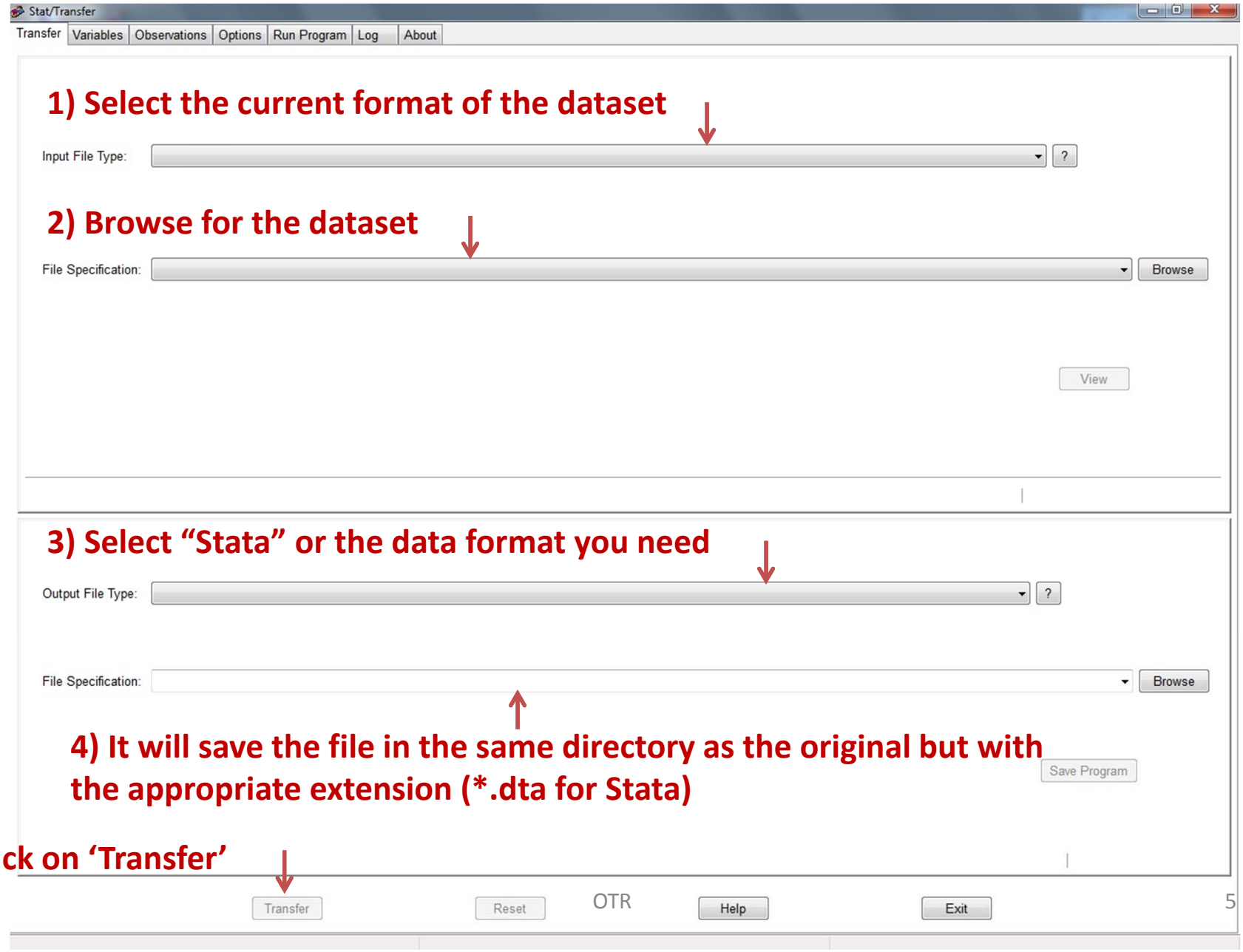
What is Stata?

- It is a multi-purpose statistical package to help you explore, summarize and analyze datasets.
- A dataset is a collection of several pieces of information called variables (usually arranged by columns). A variable can have one or several values (information for one or several cases).
- Other statistical packages are SPSS, SAS and R.
- Stata is widely used in social science research and the most used statistical software on campus.

Other data formats…

Features	Stata	SPSS	SAS	R
Data extensions	***.dta**	***.sav, *.por (portable file)**	***.sas7bcat, *.sas#bcat, *.xpt (xport files)**	***.Rdata**
User interface	Programming/point-and-click	Mostly point-and-click	Programming	Programming
Data manipulation	Very strong	Moderate	Very strong	Very strong
Data analysis	Powerful	Powerful	Powerful/versatile	Powerful/versatile
Graphics	Very good	Very good	Good	Excellent
Cost	Affordable (perpetual licenses, renew only when upgrade)	Expensive (but not need to renew until upgrade, long term licenses)	Expensive (yearly renewal)	Open source
Program extensions	*.do (do-files)	*.sps (syntax files)	*.sas	*.txt (log files)
Output extension	*.log (text file, any word processor can read it), *.smcl (formated log, only Stata can read it).	*.spo (only SPSS can read it)	(various formats)	*.R, *.txt(log files, any word processor can read)

Stat/Transfer: Transferring data from one format to another (available in the DSS lab)



Example of a dataset in Excel.

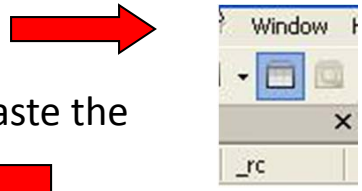
Variables are arranged by columns and cases by rows. Each variable has more than one value

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ID	Last Name	First Name	City	State	Gender	Student Status	Major	Country	Age	SAT	Average score (grade)	Height (in)	Newspaper readership (times/wk)
2	1	DOE01	JANE01	Los Angeles	California	Female	Graduate	Politics	US	30	2263	67	61	5
3	2	DOE02	JANE02	Sedona	Arizona	Female	Undergraduate	Math	US	19	2006	63	64	7
4	3	DOE16	JOE16	Elmira	New York	Male	Graduate	Math	US	26	2221	78	73	6
5	4	DOE17	JOE17	Lackawana	New York	Male	Graduate	Econ	US	33	1716	78	68	3
6	5	DOE18	JOE18	Defiance	Ohio	Male	Graduate	Econ	US	37	1701	65	71	6
7	6	DOE19	JOE19	Tel Aviv	Israel	Male	Graduate	Econ	Israel	25	1786	69	67	5
8	7	DOE20	JOE20	Cimax	North Carolina	Male	Graduate	Politics	US	39	1577	96	70	5
9	8	DOE03	JANE03	Liberal	Kansas	Female	Undergraduate	Politics	US	21	1842	87	62	5
10	9	DOE04	JANE04	Montreal	Canada	Female	Undergraduate	Math	Canada	18	1813	91	62	6
11	10	DOE05	JANE05	New York	New York	Female	Graduate	Math	US	33	2041	71	66	5
12	11	DOE21	JOE21	Hot Coffe	Mississippi	Male	Undergraduate	Econ	US	18	1787	82	67	3
13	12	DOE06	JANE06	Java	Virginia	Female	Graduate	Math	US	38	1513	79	59	5
14	13	DOE22	JOE22	Varna	Bulgaria	Male	Graduate	Politics	Bulgaria	30	1637	79	63	4
15	14	DOE23	JOE23	Moscow	Russia	Male	Graduate	Politics	Russia	30	1512	70	75	6
16	15	DOE07	JANE07	Drunkard Creek	New York	Female	Undergraduate	Math	US	21	1338	82	64	5
17	16	DOE08	JANE08	Mexican Hat	Utah	Female	Undergraduate	Econ	US	18	1821	80	63	3
18	17	DOE09	JANE09	Amsterdam	Holland	Female	Undergraduate	Math	Holland	19	1494	75	60	3
19	18	DOE10	JANE10	Mexico	Mexico	Female	Graduate	Politics	Mexico	31	2248	95	59	4
20	19	DOE11	JANE11	Caracas	Venezuela	Female	Undergraduate	Math	Venezuela	18	2252	92	68	5
21	20	DOE24	JOE24	San Juan	Puerto Rico	Male	Graduate	Politics	US	33	1923	95	63	7
22	21	DOE12	JANE12	Remote	Oregon	Female	Undergraduate	Econ	US	19	1727	67	62	7
23	22	DOE25	JOE25	New York	New York	Male	Undergraduate	Econ	US	21	1872	82	73	4
24	23	DOE13	JANE13	The X	Massachusetts	Female	Graduate	Politics	US	25	1767	89	68	6
25	24	DOE14	JANE14	Beijing	China	Female	Undergraduate	Math	China	18	1643	79	65	6
26	25	DOE26	JOE26	Stockholm	Sweden	Male	Undergraduate	Politics	Sweden	19	1919	88	64	4
27	26	DOE27	JOE27	Embarrass	Minnesota	Male	Graduate	Econ	US	28	1434	96	71	4
28	27	DOE28	JOE28	Intercourse	Pennsylvania	Male	Undergraduate	Math	US	20	2119	88	71	5
29	28	DOE15	JANE15	Loco	Oklahoma	Female	Undergraduate	Econ	US	20	2309	64	68	6
30	29	DOE29	JOE29	Buenos Aires	Argentina	Male	Graduate	Politics	Argentina	30	2279	85	72	3
31	30	DOE30	JOE30	Acme	Louisiana	Male	Undergraduate	Econ	US	19	1907	79	74	3

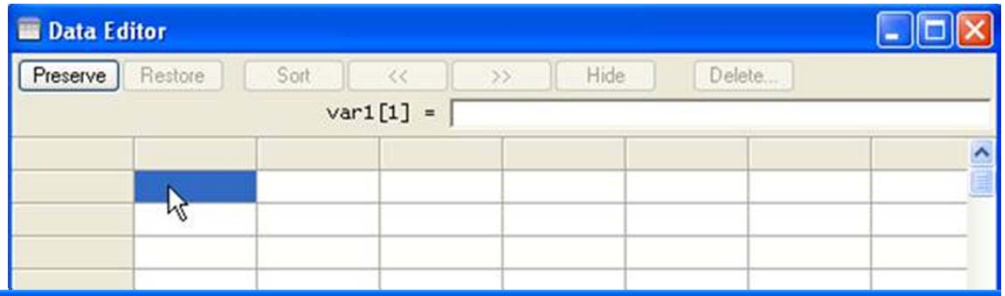
Path to the file: http://dss.princeton.edu/training/students.xls

Excel to Stata (copy-and-paste)

1 - To go **from Excel to Stata** you simply copy-and-paste data into the Stata's "Data editor" which you can open by clicking on the icon that looks like this:



2 - This window will open, is the data editor



3 - Press Ctrl-v to paste the data from Excel…

Data Editor

Preserve | Restore | Sort | << | >> | Hide | Delete...

id[1] = 1

	id	lastname	firstname	city	state	gender	studentstatus	major	country	age	sat	averagesco~e	heightin	newspaperr~k
1	1	DOE01	JANE01	Los Angeles	California	Female	Graduate	Politics	US	30	2263	67	61	5
2	2	DOE02	JANE02	Sedona	Arizona	Female	Undergraduate	Math	US	19	2006	63	64	7
3	3	DOE16	JOE16	Elmira	New York	Male	Graduate	Math	US	26	2221	78	73	6
4	4	DOE17	JOE17	Lackawana	New York	Male	Graduate	Econ	US	33	1716	78	68	3
5	5	DOE18	JOE18	Defiance	Ohio	Male	Graduate	Econ	US	37	1701	65	71	6
6	6	DOE19	JOE19	Tel Aviv	Israel	Male	Graduate	Econ	Israel	25	1786	69	67	5
7	7	DOE20	JOE20	Cimax	North Carolina	Male	Graduate	Politics	US	39	1577	96	70	5
8	8	DOE03	JANE03	Liberal	Kansas	Female	Undergraduate	Politics	US	21	1842	87	62	5
9	9	DOE04	JANE04	Montreal	Canada	Female	Undergraduate	Math	Canada	18	1813	91	62	6
10	10	DOE05	JANE05	New York	New York	Female	Graduate	Math	US	33	2041	71	66	5
11	11	DOE21	JOE21	Hot Coffe	Mississippi	Male	Undergraduate	Econ	US	18	1787	82	67	3
12	12	DOE06	JANE06	Java	Virginia	Female	Graduate	Math	US	38	1513	79	59	5
13	13	DOE22	JOE22	Varna	Bulgaria	Male	Graduate	Politics	Bulgaria	30	1637	79	63	4
14	14	DOE23	JOE23	Moscow	Russia	Male	Graduate	Politics	Russia	30	1512	70	75	6
15	15	DOE07	JANE07	Drunkard Creek	New York	Female	Undergraduate	Math	US	21	1338	82	64	5
16	16	DOE08	JANE08	Mexican Hat	Utah	Female	Undergraduate	Econ	US	18	1821	80	63	3
17	17	DOE09	JANE09	Amsterdam	Holland	Female	Undergraduate	Math	Holland	19	1494	75	60	3
18	18	DOE10	JANE10	Mexico	Mexico	Female	Graduate	Politics	Mexico	31	2248	95	59	4
19	19	DOE11	JANE11	Caracas	Venezuela	Female	Undergraduate	Math	Venezuela	18	2252	92	68	5
20	20	DOE24	JOE24	San Juan	Puerto Rico	Male	Graduate	Politics	US	33	1923	95	63	7
21	21	DOE12	JANE12	Remote	Oregon	Female	Undergraduate	Econ	US	19	1727	67	62	7
22	22	DOE25	JOE25	New York	New York	Male	Undergraduate	Econ	US	21	1872	82	73	4
23	23	DOE13	JANE13	The X	Massachusetts	Female	Graduate	Politics	US	25	1767	89	68	6
24	24	DOE14	JANE14	Beijing	China	Female	Undergraduate	Math	China	18	1643	79	65	6
25	25	DOE26	JOE26	Stockholm	Sweden	Male	Undergraduate	Politics	Sweden	19	1919	88	64	4
26	26	DOE27	JOE27	Embarrass	Minnesota	Male	Graduate	Econ	US	28	1434	96	71	4
27	27	DOE28	JOE28	Intercourse	Pennsylvania	Male	Undergraduate	Math	US	20	2119	88	71	5
28	28	DOE15	JANE15	Loco	Oklahoma	Female	Undergraduate	Econ	US	20	2309	64	68	6
29	29	DOE29	JOE29	Buenos Aires	Argentina	Male	Graduate	Politics	Argentina	30	2279	85	72	3
30	30	DOE30	JOE30	Acme	Louisiana	Male	Undergraduate	Econ	US	19	1907	79	74	3

Stata color-coded system

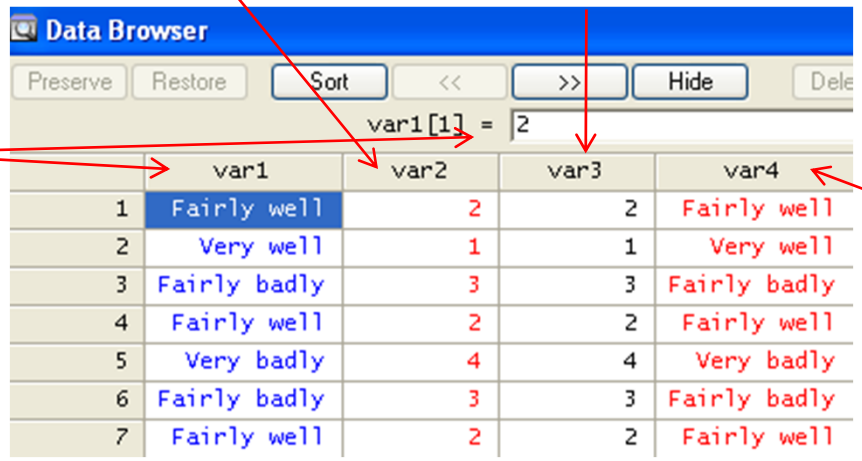
An important step is to make sure variables are in their expected format.

Stata has a color-coded system for each type. Black is for numbers, red is for text or string and blue is for labeled variables.

Var2 is a string variable even though you see numbers. You can't do any statistical procedure with this variable other than simple frequencies

Var3 is a numeric You can do any statistical procedure with this variable

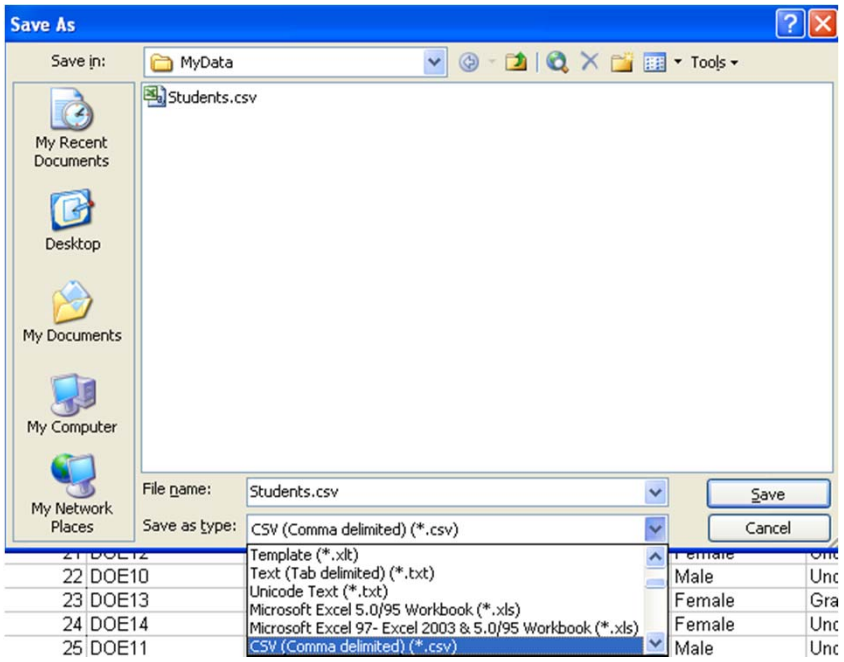
For var1 a value 2 has the label "Fairly well". It is still a numeric variable



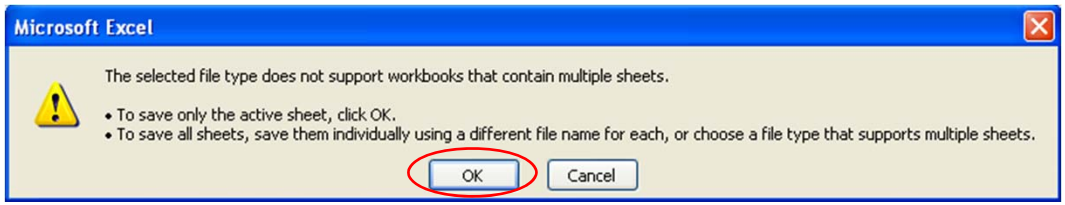
	var1	var2	var3	var4
1	Fairly well	2	2	Fairly well
2	Very well	1	1	Very well
3	Fairly badly	3	3	Fairly badly
4	Fairly well	2	2	Fairly well
5	Very badly	4	4	Very badly
6	Fairly badly	3	3	Fairly badly
7	Fairly well	2	2	Fairly well

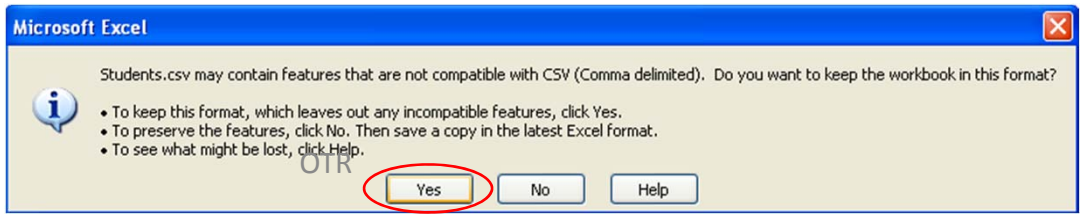
Var4 is clearly a string variable. You can do frequencies and crosstabulations with this but not statistical procedures.

Another way to bring excel data into Stata is by saving the Excel file as ***.csv** (comma-separated values) and import it in Stata using the `insheet` command.
In **Excel** go to `File->Save as` and save the Excel file as *.csv:



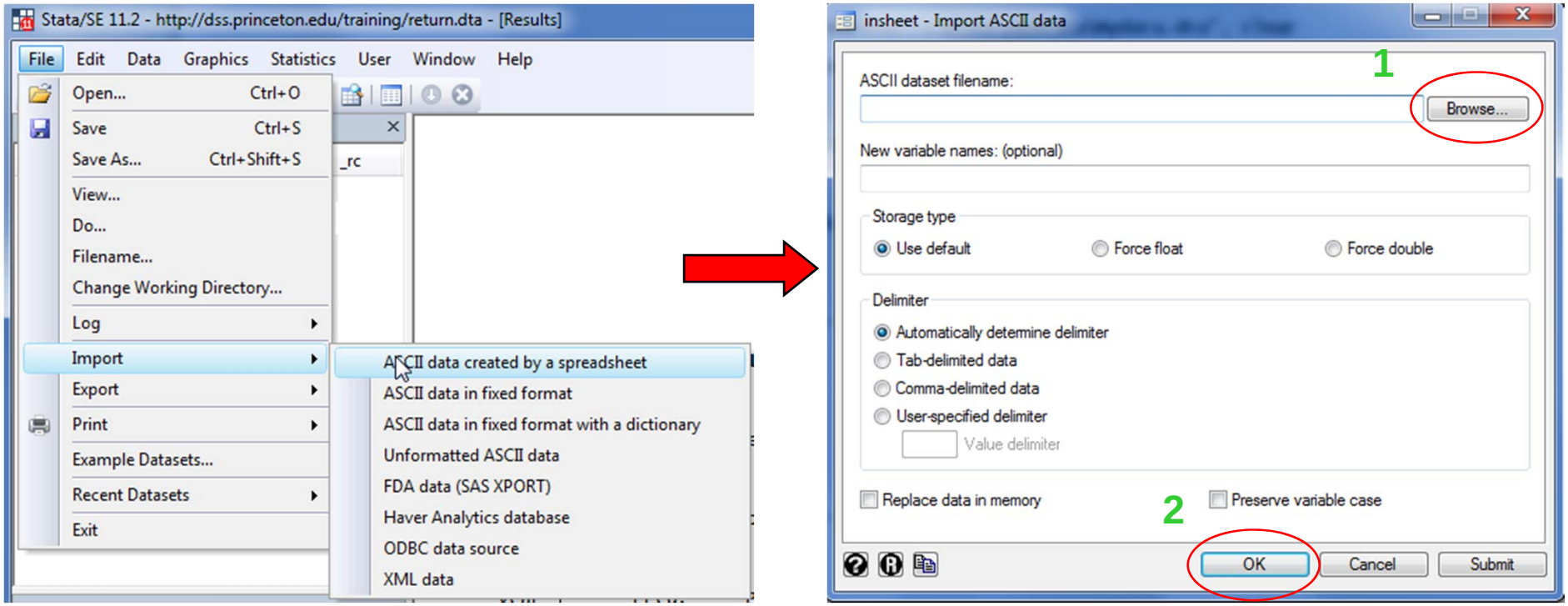
You may get the following messages, click OK and YES…





Go to the next page…

In **Stata** go to `File->Import->"ASCII data created by spreadsheet"`. Click on '`Browse`' to find the file and then OK.



An alternative to using the menu you can type:

```
insheet using "c:\mydata\mydatafile.csv"
```

Exercises

Exercise 1

Using the ICPSR Online Learning Center, go to guide on *Civic Participation and Demographics in Rural China (1990)*
http://www.icpsr.umich.edu/icpsrweb/ICPSR/OLC/guides/China/sections/a01

Got to the tab 'Dataset' and download the data (http://www.icpsr.umich.edu/icpsrweb/ICPSR/OLC/guides/China/sections/a02)

We'll focus on the first exercise on 'Age and Participation' and use the following variables:

- Respondent's year of birth (M1001)
- Village meeting attendance (M3090)

Activities:

- Create the variable 'age' for each respondent
- Create the variable 'agegroup' with the following categories: 16-35, 36-55 and 56-79

Questions:

- What percentage of respondents reported attending a local village meeting?
- Of those attending a meeting, which age group was most likely to report attending a village meeting?
- Of those attending a meeting , which group was most likely to report no village meeting attendance?

Source: Inter-university Consortium for Political and Social Research. Civic Participation and Demographics in Rural China: A Data-Driven Learning Guide. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], July, 31 2009. Doi:10.3886/China

Exercise 2

Got to the *World Development Indicators (WDI) & Global Development Finance (GDF)* from the World Bank (access from the library's *Articles and Databases,* http://library.princeton.edu/catalogs/articles.php)

Direct link to WDI/GDF http://databank.worldbank.org/ddp/home.do?Step=12&id=4&CNO=2

Get data for the United States and **all** available years on:

- Long-term unemployment (% of total unemployment)
- Long-term unemployment, female (% of female unemployment)
- Long-term unemployment, male (% of male unemployment)
- Inflation, consumer prices (annual %)
- GDP per capita (constant 2000 US$)
- GDP per capita growth (annual %)

See here to arrange the data as panel data http://dss.princeton.edu/training/FindingData101.pdf#page=21
For an example of how panel data looks like click here: http://dss.princeton.edu/training/DataPrep101.pdf#page=3

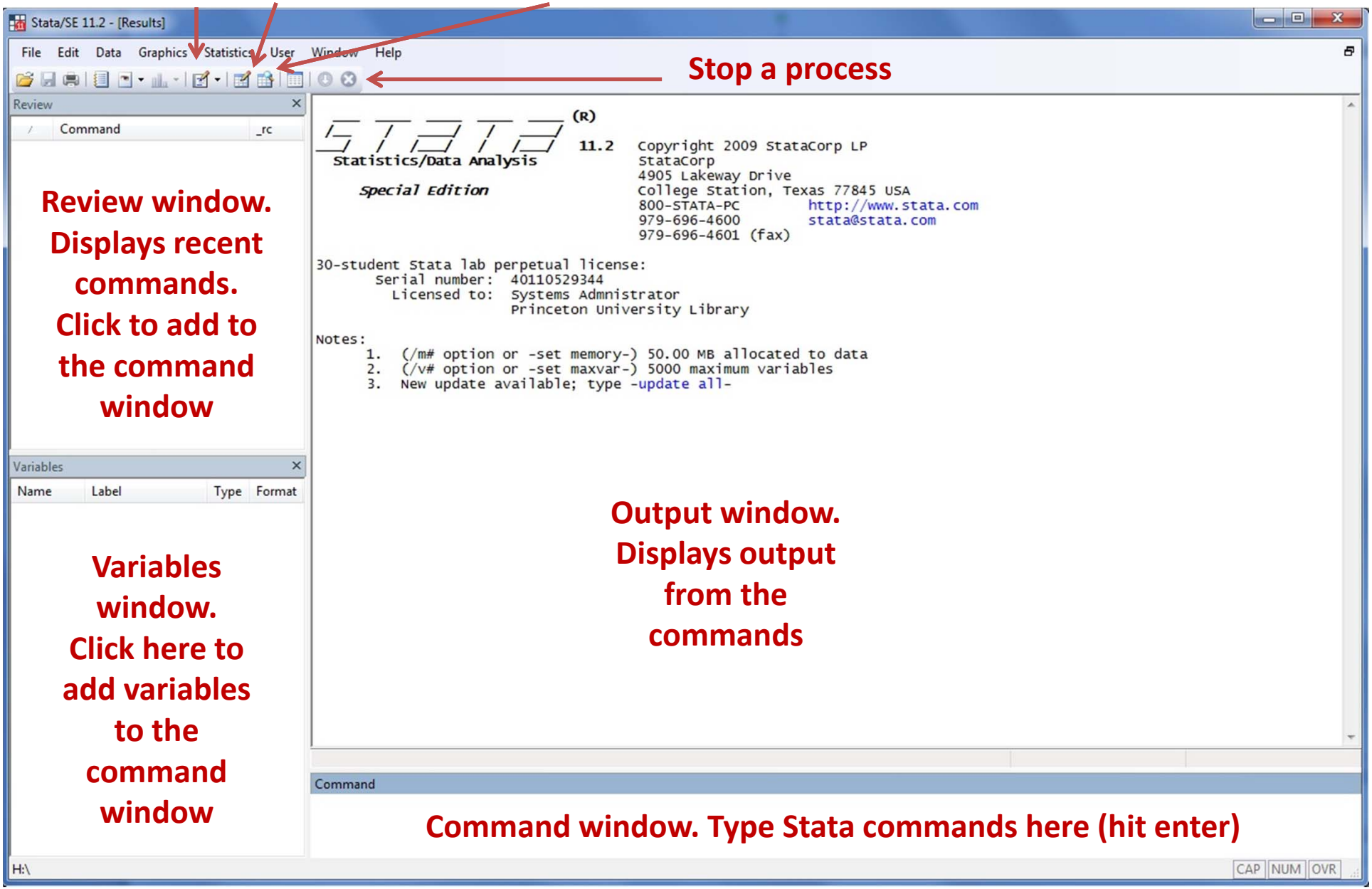
Activities:

- Rename the variables and explore the data (use describe, summarize)
- Create a variable called crisis where it takes the value of 17 for the following years: 1960, 1961, 1969, 1970, 1973, 1974, 1975, 1981, 1982, 1990, 1991, 2001, 2007, 2008, 2009. Replace missing with zeros (source: nber.org).
- Set as time series (see http://dss.princeton.edu/training/TS101.pdf#page=6)
- Create a line graph with unemployment rate (total, female and males) and crisis by year.

Questions:

- What do you see? Who tends to be more affected by the economic recessions?

Basic commands


Do-file editor
Open data editor
Open data browser
Stop a process
Stata/SE 11.2 - [Results]
File Edit Data Graphics Statistics User Window Help
Review
Command _rc
Review window. Displays recent commands. Click to add to the command window
Variables
Name Label Type Format
Variables window. Click here to add variables to the command window
(R)
Statistics/Data Analysis
11.2
Special Edition
Copyright 2009 StataCorp LP
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC http://www.stata.com
979-696-4600 stata@stata.com
979-696-4601 (fax)
30-student Stata lab perpetual license:
Serial number: 40110529344
Licensed to: Systems Admnistrator
Princeton University Library
Notes:
1. (/m# option or -set memory-) 50.00 MB allocated to data
2. (/v# option or -set maxvar-) 5000 maximum variables
3. New update available; type -update all-
Output window. Displays output from the commands
Command
Command window. Type Stata commands here (hit enter)
H:\
CAP NUM OVR

First steps: Working directory

To see your working directory, type

```
pwd
```

To change the working directory to avoid typing the whole path when calling or saving files, type:

```
cd c:\mydata
```

Use quotes if the new directory has blank spaces, for example

```
cd “h:\stata and data”
```

If you want to use the menu go to (useful with Macs):

```
File -> Change Working Directory…
```

First steps: log file

Create a ***log file***, sort of Stata's built-in tape recorder and where you can: 1) retrieve the output of your work and 2) keep a record of your work.

In the command line type:

```
log using mylog.log
```

This creates the file '`mylog.log`' in your working directory. You can read it using any word processor (notepad, word, etc.).

To close a log file type:

```
log close
```

To add more output to an existing log file add the option `append`, type:

```
log using mylog.log, append
```

To replace a log file add the option `replace`, type:

```
log using mylog.log, replace
```

Note that the option `replace` will delete the contents of the previous version of the log.

First steps: set the correct memory allocation

If you get the following error message while opening a datafile or adding more variables:

```
no room to add more observations
    An attempt was made to increase the number of observations beyond what is currently possible.  You have the
    following alternatives:

     1.  Store your variables more efficiently; see help compress.  (Think of Stata's data area as the area of a
         rectangle; Stata can trade off width and length.)

     2.  Drop some variables or observations; see help drop.

     3.  Increase the amount of memory allocated to the data area using the set memory command; see help memory.
```

You need to set the *correct memory allocation* for your data or the maximun number of variable allowed. Some big datasets need more memory, depending on the size you can type, for example:

`set mem 700m`

```
. set mem 700m

Current memory allocation

    settable       current value     description                    memory usage (1M = 1024k)
-------------------------------------------------------------------------------------------
set maxvar          5000           max. variables allowed                1.909M
set memory           700M          max. data space                     700.000M
set matsize          400           max. RHS vars in models               1.254M
                                                                       ---------
                                                                       703.163M
```

Note: If this does not work try a bigger number.

*To allow more variables type `set maxvar 10000`

First steps: Opening/saving Stata files (*.dta)

To open files already in Stata with extension *.dta, run Stata and you can either:

- Use the menu: go to `file->open`, or
- In the command window type `use "c:\mydata\mydatafile.dta"`

If your working directory is already set to `c:\mydata`, just type

```
use mydatafile
```

To save a data file from Stata go to file – save as or just type:

```
save, replace
```

If the dataset is new or just imported from other format go to `file -> save as` or just type:

```
save mydatafile
```

For ASCII data please see http://dss.princeton.edu/training/DataPrep101.pdf

Command: `describe`

To get a general description of the dataset and the format for each variable type `describe`

```
. describe

Contains data from http://dss.princeton.edu/training/students.dta
  obs:            30
 vars:            14                          29 Sep 2009 17:12
 size:         2,580 (99.9% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
id              byte   %8.0g                  ID
lastname        str5   %9s                    Last Name
firstname       str6   %9s                    First Name
city            str14  %14s                   City
state           str14  %14s                   State
gender          str6   %9s                    Gender
studentstatus   str13  %13s                   Student Status
major           str8   %9s                    Major
country         str9   %9s                    Country
age             byte   %8.0g                  Age
sat             int    %8.0g                  SAT
averagescoreg~e byte   %8.0g                  Average score (grade
heightin        byte   %8.0g                  Height (in)
newspaperread~k byte   %8.0g                  Newspaper readership
-------------------------------------------------------------------------------
```

Type `help describe` for more information…

Command: summarize

Type summarize to get some basic descriptive statistics.

```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
id	30	15.5	8.803408	1	30
lastname	0				
firstname	0				
city	0				
state	0				
gender	0				
studentsta~s	0				
major	0				
country	0				
age	30	25.2	6.870226	18	39
sat	30	1848.9	275.1122	1338	2309
averagesco~e	30	80.36667	10.11139	63	96
heightin	30	66.43333	4.658573	59	75
newspaperr~k	30	4.866667	1.279368	3	7

Zeros indicate string variables

Use ‘min’ and ‘max’ values to check for a valid range in each variable. For example, ‘age’ should have the expected values (‘don’t know’ or ‘no answer’ are usually coded as 99 or 999)

Type help summarize for more information…

Exploring data: frequencies

Frequency refers to the number of times a value is repeated. Frequencies are used to analyze categorical data. The tables below are *frequency tables*, values are in ascending order. In Stata use the command **`tab varname.`**

variable

```
. tab major
```

Major	Freq.	Percent	Cum.
Econ	10	33.33	33.33
Math	10	33.33	66.67
Politics	10	33.33	100.00
Total	30	100.00	

'Freq.' provides a raw count of each value. In this case 10 students for each major.

'Percent' gives the relative frequency for each value. For example, 33.33% of the students in this group are econ majors.

'Cum.' is the cumulative frequency in ascending order of the values. For example, 66.67% of the students are econ or math majors.

variable

```
. tab readnews
```

Newspaper readership (times/wk)	Freq.	Percent	Cum.
3	6	20.00	20.00
4	5	16.67	36.67
5	9	30.00	66.67
6	7	23.33	90.00
7	3	10.00	100.00
Total	30	100.00	

'Freq.' Here 6 students read the newspaper 3 days a week, 9 students read it 5 days a week.

'Percent'. Those who read the newspaper 3 days a week represent 20% of the sample, 30% of the students in the sample read the newspaper 5 days a week.

'Cum.' 66.67% of the students read the newspaper 3 to 5 days a week.

Type `help tab` for more details.

Exploring data: frequencies and descriptive statistics (using `table`)

Command `table` produces frequencies and descriptive statistics per category. For more info and a list of all statistics type `help table`. Here are some examples, type

```
table gender, contents(freq mean age mean score)
```

```
. table gender, contents(freq mean age mean score)
```

Gender	Freq.	mean(age)	mean(score)
Female	15	23.2	78.73333
Male	15	27.2	82

The mean age of females is 23 years, for males is 27. The mean score is 78 for females and 82 for males. Here is another example:

```
table major, contents(freq mean age mean sat mean score mean readnews)
```

```
. table major, contents(freq mean  age mean sat mean  score mean  readnews)
```

Major	Freq.	mean(age)	mean(sat)	mean(score)	mean(read~s)
Econ	10	23.8	1806	76.2	4.4
Math	10	23	1844	79.8	5.3
Politics	10	28.8	1896.7	85.1	4.9

Exploring data: crosstabs

Also known as *contingency tables*, crosstabs help you to analyze the relationship between two or more categorical variables. Below is a crosstab between the variable 'ecostatu' and 'gender'. We use the command **`tab var1 var2`**

Options 'column', 'row' gives you the column and row percentages.

var1 var2

```
. tab ecostatu gender, column row
```

Key
frequency *row percentage* *column percentage*

Status of Nat'l Eco	Gender of Respondent Male	Female	Total
Very well	90 60.40 14.33	59 39.60 7.92	149 100.00 10.85
Fairly well	337 50.30 53.66	333 49.70 44.70	670 100.00 48.80
Fairly badly	139 39.94 22.13	209 60.06 28.05	348 100.00 25.35
Very badly	57 29.84 9.08	134 70.16 17.99	191 100.00 13.91
Not sure	2 16.67 0.32	10 83.33 1.34	12 100.00 0.87
Refused	3 100.00 0.48	0 0.00 0.00	3 100.00 0.22
Total	628 45.74 100.00	745 54.26 100.00	1,373 100.00 100.00

The first value in a cell tells you the number of observations for each xtab. In this case, 90 respondents are 'male' and said that the economy is doing 'very well', 59 are 'female' and believe the economy is doing 'very well'

The second value in a cell gives you row percentages for the first variable in the xtab. Out of those who think the economy is doing 'very well', 60.40% are males and 39.60% are females.

The third value in a cell gives you column percentages for the second variable in the xtab. Among males, 14.33% think the economy is doing 'very well' while 7.92% of females have the same opinion.

NOTE: You can use `tab1` for multiple frequencies or `tab2` to run all possible crosstabs combinations. Type help tab for further details.

Exploring data: crosstabs (a closer look)

You can use crosstabs to compare responses among categories in relation to aggregate responses. In the table below we can see how opinions for males and females diverge from the national average.

```
tab ecostatu gender, column row
```

Key
frequency *row percentage* *column percentage*

Status of Nat'l Eco	Gender of Respondent Male	Female	Total
Very well	90 60.40 14.33	59 39.60 7.92	149 100.00 10.85
Fairly well	337 50.30 53.66	333 49.70 44.70	670 100.00 48.80
Fairly badly	139 39.94 22.13	209 60.06 28.05	348 100.00 25.35
Very badly	57 29.84 9.08	134 70.16 17.99	191 100.00 13.91
Not sure	2 16.67 0.32	10 83.33 1.34	12 100.00 0.87
Refused	3 100.00 0.48	0 0.00 0.00	3 100.00 0.22
Total	628 45.74 100.00	745 54.26 100.00	1,373 100.00 100.00

As a rule-of-thumb, a margin of error of ±4 percentage points can be used to indicate a significant difference (some use ±3).

For example, rounding up the percentages, 11% (10.85) answer 'very well' at the national level. With the margin of error, this gives a range roughly between 7% and 15%, anything beyond this range could be considered significantly different (remember this is just an approximation). It does not appear to be a significant bias between males and females for this answer.

In the 'fairly well' category we have 49%, with range between 45% and 53%. The response for males is 54% and for females 45%. We could say here that males tend to be a bit more optimistic on the economy and females tend to be a bit less optimistic.

If we aggregate responses, we could get a better picture. In the table below 68% of males believe the economy is doing well (comparing to 60% at the national level, while 46% of females thing the economy is bad (comparing to 39% aggregate). Males seem to be more optimistic than females.

RECODE of ecostatu (Status of Nat'l Eco)	Gender of Respondent Male	Female	Total
Well	427 52.14 67.99	392 47.86 52.62	819 100.00 59.65
Bad	196 36.36 31.21	343 63.64 46.04	539 100.00 39.26
Not sure/ref	5 33.33 0.80	10 66.67 1.34	15 100.00 1.09
Total	628 45.74 100.00	745 54.26 100.00	1,373 100.00 100.00

```
recode ecostatu (1 2 = 1 "Well") (3 4 = 2 "Bad") (5 6=3 "Not sure/ref"), gen(ecostatu1) label(eco)
```

Exploring data: crosstabs (test for associations)

To see whether there is a relationship between two variables you can choose a number of tests. Some apply to nominal variables some others to ordinal. I am running all of them here for presentation purposes.

```
tab ecostatu1 gender, column row nokey chi2 lrchi2 V exact gamma taub
```

Likelihood-ratio χ^2(chi-square) | *Goodman & Kruskal's γ (gamma)* | X^2(chi-square) | Cramer's *V* | Kendall's τ_b (tau-b) | Fisher's exact test

```
. tab ecostatu1 gender, column row nokey chi2 lrchi2 V exact gamma taub

Enumerating sample-space combinations:
stage 3:  enumerations = 1
stage 2:  enumerations = 16
stage 1:  enumerations = 0
```

RECODE of ecostatu (Status of Nat'l Eco)	Gender of Respondent Male	Female	Total
Well	427	392	819
	52.14	47.86	100.00
	67.99	52.62	59.65
Bad	196	343	539
	36.36	63.64	100.00
	31.21	46.04	39.26
Not sure/ref	5	10	15
	33.33	66.67	100.00
	0.80	1.34	1.09
Total	628	745	1,373
	45.74	54.26	100.00
	100.00	100.00	100.00

```
          Pearson chi2(2) =  33.5266   Pr = 0.000
 likelihood-ratio chi2(2) =  33.8162   Pr = 0.000
                Cramér's V =   0.1563
                     gamma =   0.3095   ASE = 0.050
           Kendall's tau-b =   0.1553   ASE = 0.026
            Fisher's exact =                 0.000
```

- For *nominal* data use chi2, lrchi2, V
- For *ordinal* data use gamma and taub
- Use `exact` instead of `chi2` when frequencies are less than 5 across the table.

X^2(chi-square) tests for relationships between variables. The null hypothesis (Ho) is that there is no relationship. To reject this we need a Pr < 0.05 (at 95% confidence). Here both chi2 are significant. Therefore we conclude that there is some relationship between perceptions of the economy and gender. `lrchi2` reads the same way.

Cramer's *V* is a measure of association between two nominal variables. It goes from 0 to 1 where 1 indicates strong association (for rXc tables). In 2x2 tables, the range is -1 to 1. Here the V is 0.15, which shows a small association.

Gamma and taub are measures of association between two ordinal variables (both have to be in the same direction, i.e. negative to positive, low to high). Both go from -1 to 1. Negative shows inverse relationship, closer to 1 a strong relationship. Gamma is recommended when there are lots of ties in the data. Taub is recommended for square tables.

Fisher's exact test is used when there are very few cases in the cells (usually less than 5). It tests the relationship between two variables. The null is that variables are independent. Here we reject the null and conclude that there is some kind of relationship between variables

Exploring data: descriptive statistics

For continuous data use descriptive statistics. These statistics are a collection of measurements of: *location* and *variability*. Location tells you the central value the variable (the mean is the most common measure of this) . Variability refers to the spread of the data from the center value (i.e. variance, standard deviation). Statistics is basically the study of what causes such variability. We use the command `tabstat` to get these stats.

```
tabstat age sat score heightin readnews, s(mean median sd var count range min max)
```

```
. tabstat  age sat score heightin readnews, s(mean median sd var count range min max)
```

stats	age	sat	score	heightin	readnews
mean	25.2	1848.9	80.36667	66.43333	4.866667
p50	23	1817	79.5	66.5	5
sd	6.870226	275.1122	10.11139	4.658573	1.279368
variance	47.2	75686.71	102.2402	21.7023	1.636782
N	30	30	30	30	30
range	21	971	33	16	4
min	18	1338	63	59	3
max	39	2309	96	75	7

Type `help tabstat` for a complete list of descriptive statistics

- The *mean* is the sum of the observations divided by the total number of observations.
- The *median* (p50 in the table above) is the number in the middle . To get the median you have to order the data from lowest to highest. If the number of cases is odd the median is the single value, for an even number of cases the median is the average of the two numbers in the middle.
- The *standard deviation* is the squared root of the variance. Indicates how close the data is to the mean. Assuming a normal distribution, 68% of the values are within 1 sd from the mean, 95% within 2 sd and 99% within 3 sd
- The *variance* measures the dispersion of the data from the mean. It is the simple mean of the squared distance from the mean.
- *Count* (N in the table) refers to the number of observations per variable.
- *Range* is a measure of dispersion. It is the difference between the largest and smallest value, max – min.
- *Min* is the lowest value in the variable.
- *Max* is the largest value in the variable.

Exploring data: descriptive statistics

You could also estimate descriptive statistics by subgroups (i.e. gender, age, etc.)

```
tabstat age sat score heightin readnews, s(mean median sd var count range min max) by(gender)
```

```
. tabstat  age sat score heightin readnews, s(mean median sd var count range min max) by(gender)

Summary statistics: mean, p50, sd, variance, N, range, min, max
  by categories of: gender (Gender)
```

gender	age	sat	score	heightin	readnews
Female	23.2	1871.8	78.73333	63.4	5.2
	20	1821	79	63	5
	6.581359	307.587	10.66012	3.112188	1.207122
	43.31429	94609.74	113.6381	9.685714	1.457143
	15	15	15	15	15
	20	971	32	9	4
	18	1338	63	59	3
	38	2309	95	68	7
Male	27.2	1826	82	69.46667	4.533333
	28	1787	82	71	4
	6.773899	247.0752	9.613978	3.943651	1.302013
	45.88571	61046.14	92.42857	15.55238	1.695238
	15	15	15	15	15
	21	845	31	12	4
	18	1434	65	63	3
	39	2279	96	75	7
Total	25.2	1848.9	80.36667	66.43333	4.866667
	23	1817	79.5	66.5	5
	6.870226	275.1122	10.11139	4.658573	1.279368
	47.2	75686.71	102.2402	21.7023	1.636782
	30	30	30	30	30
	21	971	33	16	4
	18	1338	63	59	3
	39	2309	96	75	7

Type `help tabstat` for more options.

Examples of frequencies and crosstabulations

Frequencies (tab command)

```
. tab gender
```

Gender	Freq.	Percent	Cum.
Female	15	50.00	50.00
Male	15	50.00	100.00
Total	30	100.00	

In this sample we have 15 females and 15 males. Each represents 50% of the total cases.

Crosstabulations (tab with two variables)

```
. tab gender studentstatus, column row
```

Key

frequency
row percentage
column percentage

Gender	Student Status: Graduate	Undergrad	Total
Female	5 33.33 33.33	10 66.67 66.67	15 100.00 50.00
Male	10 66.67 66.67	5 33.33 33.33	15 100.00 50.00
Total	15 50.00 100.00	15 50.00 100.00	30 100.00 100.00

Average SAT scores by gender and major. Notice, 'sat' variable is a continuous variable. The first cell reads the average SAT score for a female whose major is econ is 1952.3333 with a standard deviation 312.43, there are only 3 females with a major in econ.

```
. tab gender major, sum(sat)
```

Means, Standard Deviations and Frequencies of SAT

Gender	Major: Econ	Math	Politics	Total
Female	1952.3333 312.43773 3	1762.5 317.99326 8	2030 262.25052 4	1871.8 307.58697 15
Male	1743.2857 155.6146 7	2170 72.124892 2	1807.8333 288.99994 6	1826 247.07518 15
Total	1806 219.16559 10	1844 329.76928 10	1896.7 287.20687 10	1848.9 275.11218 30

Three way crosstabs

bysort *var3*: tab *var1 var2*, colum row

bysort *studentstatus*: tab *gender major*, colum row

```
. bysort  studentstatus: tab gender major, column row
```

-> studentstatus = Graduate

Key
frequency *row percentage* *column percentage*

Gender	Major Econ	Math	Politics	Total
Female	0 0.00 0.00	2 40.00 66.67	3 60.00 37.50	5 100.00 33.33
Male	4 40.00 100.00	1 10.00 33.33	5 50.00 62.50	10 100.00 66.67
Total	4 26.67 100.00	3 20.00 100.00	8 53.33 100.00	15 100.00 100.00

-> studentstatus = Undergraduate

Key
frequency *row percentage* *column percentage*

Gender	Major Econ	Math	Politics	Total
Female	3 30.00 50.00	6 60.00 85.71	1 10.00 50.00	10 100.00 66.67
Male	3 60.00 50.00	1 20.00 14.29	1 20.00 50.00	5 100.00 33.33
Total	6 40.00 100.00	7 46.67 100.00	2 13.33 100.00	15 100.00 100.00

Three way crosstabs with summary statistics of a fourth variable

Average SAT scores by gender and major for graduate and undergraduate students. The third cell reads: The average SAT score of a female graduate student whose major is politics is 2092.6667 with a standard deviation of 2.82.13, there are 3 graduate female students with a major in politics.

```
. bysort  studentstatus: tab gender major, sum(sat)
```

-> studentstatus = Graduate

Means, Standard Deviations and Frequencies of SAT

Gender	Major			
	Econ	Math	Politics	Total
Female	.	1777	2092.6667	1966.4
	.	373.35238	282.13531	323.32924
	0	2	3	5
Male	1659.25	2221	1785.6	1778.6
	154.66819	0	317.32286	284.3086
	4	1	5	10
Total	1659.25	1925	1900.75	1841.2
	154.66819	367.97826	324.8669	300.38219
	4	3	8	15

-> studentstatus = Undergraduate

Means, Standard Deviations and Frequencies of SAT

Gender	Major			
	Econ	Math	Politics	Total
Female	1952.3333	1757.6667	1842	1824.5
	312.43773	337.01197	0	305.36872
	3	6	1	10
Male	1855.3333	2119	1919	1920.8
	61.711695	0	0	122.23011
	3	1	1	5
Total	1903.8333	1809.2857	1880.5	1856.6
	208.30979	336.59952	54.447222	257.72682
	6	7	2	15

First steps: Quick way of finding variables (`lookfor`)

You can use the command lookfor to find variables in a dataset, for example you want to see which variables refer to education, type:

lookfor educ

```
. lookfor educ

                storage  display     value
variable name   type     format      label      variable label
------------------------------------------------------------------------------
educ            byte     %10.0g                 Education of R.
```

`lookfor` will look for the keyword 'educ' in the variable name and labels. You will need to be creative with your keyword searches to find the variables you need.

It always recommended to use the codebook that comes with the dataset to have a better idea of where things are.

First steps: Subsetting using conditional 'if'

Sometimes you may want to get frequencies, crosstabs or run a model just for a particular group (lets say just for females or people younger than certain age). You can do this by using the conditional 'if', for example:

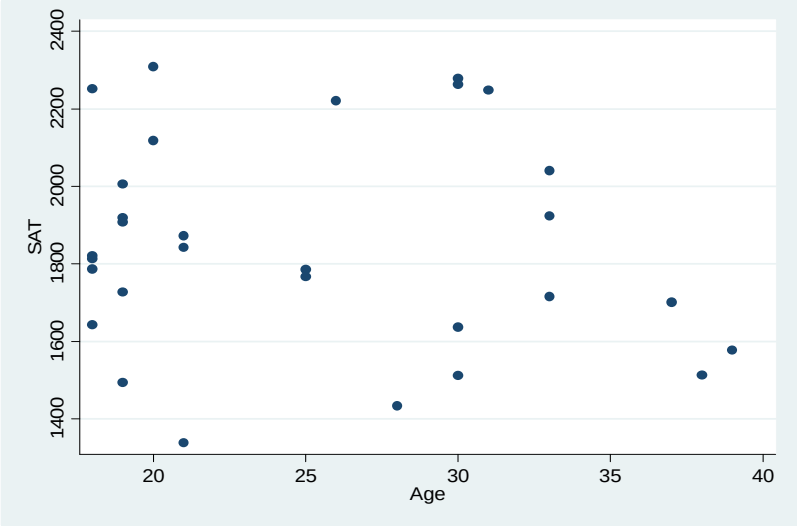
```
/*Frequencies of var1 when gender = 1*/
tab var1 if gender==1, column row

/*Frequencies of var1 when gender = 1 and age < 33*/
tab var1 if gender==1 & age<33, column row

/*Frequencies of var1 when gender = 1 and marital status = single*/
tab var1 if gender==1 & marital==2 | marital==3 | marital==4, column row

/*You can do the same with crosstabs: tab var1 var2 … */

/*Regression when gender = 1 and age < 33*/
regress y x1 x2 if gender==1 & age<33, robust

/*Scatterplots when gender = 1 and age < 33*/
scater var1 var2 if gender==1 & age<33
```

"`if`" goes at the end of the command BUT before the comma that separates the options from the command.

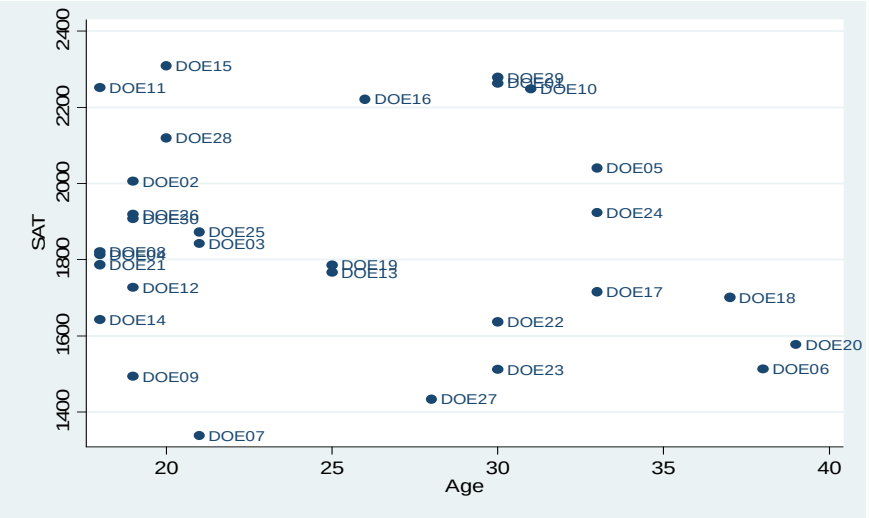
Graphs: scatterplot

Scatterplots are good to explore possible relationships or patterns between variables and to identify outliers. Use the command `scatter` (sometimes adding `twoway` is useful when adding more graphs). The format is `scatter y x`. Below we check the relationship between SAT scores and age. For more details type `help scatter` .

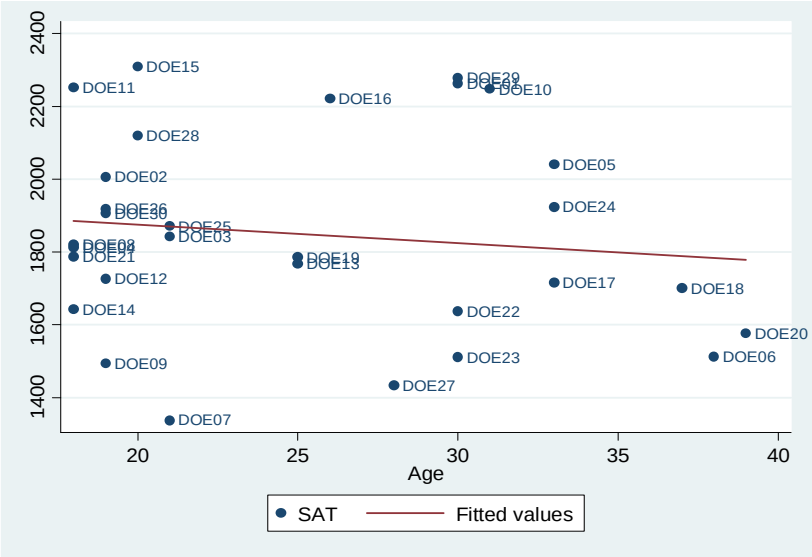
`twoway scatter sat age`



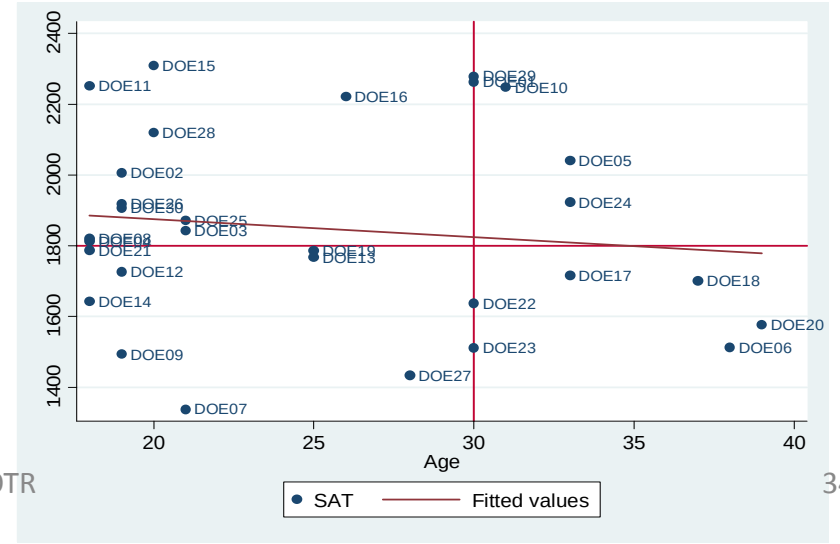
`twoway scatter sat age, mlabel(last)`



`twoway scatter sat age, mlabel(last) || lfit sat age`

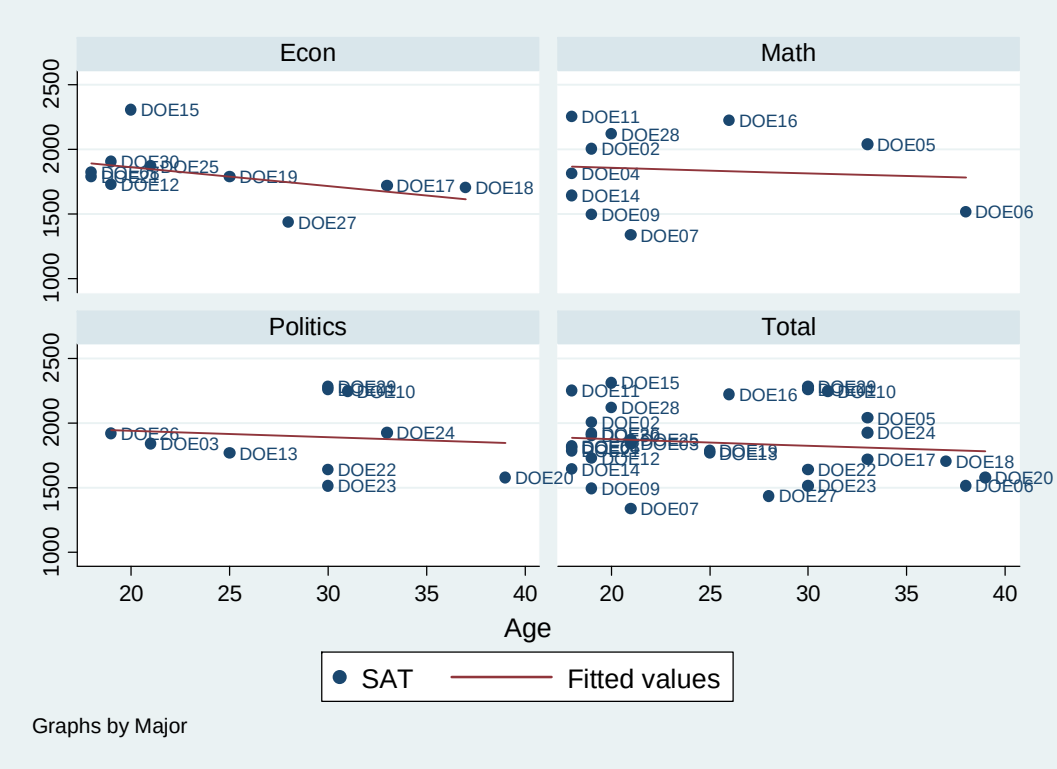


`twoway scatter sat age, mlabel(last) || lfit sat age, yline(30) xline(1800)`



By categories

```
twoway scatter sat age, mlabel(last) by(major, total)
```

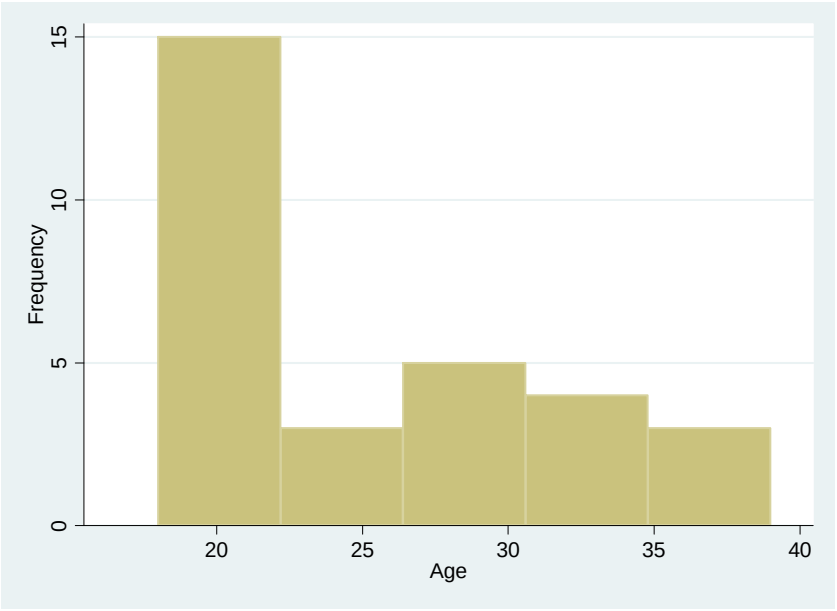


Go to http://www.princeton.edu/~otorres/Stata/ for additional tips

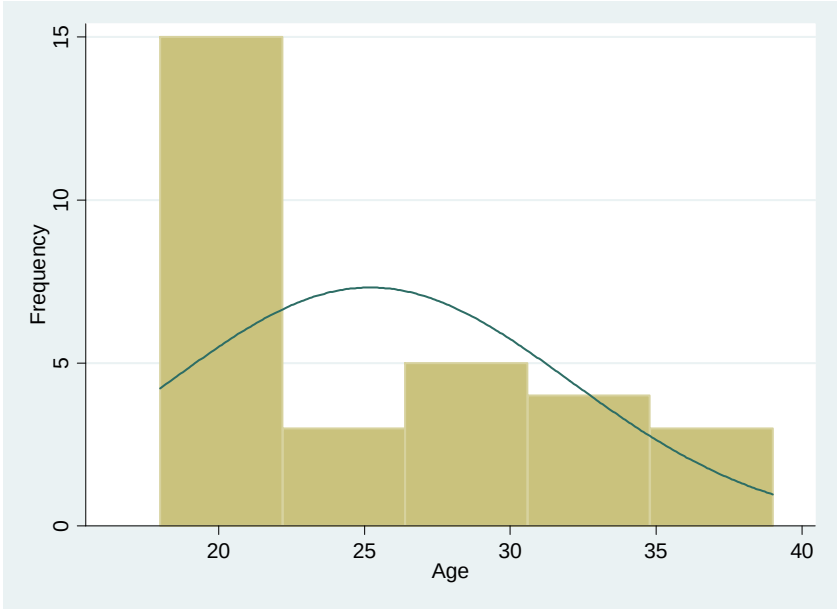
Graphs: histogram

Histograms are another good way to visually explore data, especially to check for a normal distribution. Type `help histogram` for details.

`histogram age, frequency`



`histogram age, frequency normal`



Frequently used Stata commands

Category	Stata commands
Getting on-line help	**help** **search**
Operating-system interface	**pwd** **cd** **sysdir** **mkdir** **dir / ls** **erase** **copy** **type**
Using and saving data from disk	**use** **clear** **save** **append** **merge** **compress**
Inputting data into Stata	**input** **edit** **infile** **infix** **insheet**
The Internet and Updating Stata	**update** **net** **ado** **news**
Basic data reporting	**describe** **codebook** **inspect** **list** **browse** **count** **assert** **summarize** **Table (tab)** **tabulate**
Data manipulation	**generate** **replace** **egen** **recode** **rename** **drop** **keep** **sort** **encode** **decode** **order** **by** **reshape**
Formatting	**format** **label**
Keeping track of your work	**log** **notes**
Convenience	**display**

Type `help [command name]` in the windows command for details

Source: http://www.ats.ucla.edu/stat/stata/notes2/commands.htm

Is my model OK? (links)

Regression diagnostics: A checklist
http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter2/statareg2.htm

Logistic regression diagnostics: A checklist
http://www.ats.ucla.edu/stat/stata/webbooks/logistic/chapter3/statalog3.htm

Times series diagnostics: A checklist (pdf)
http://homepages.nyu.edu/~mrg217/timeseries.pdf

Times series: dfueller test for unit roots (for R and Stata)
http://www.econ.uiuc.edu/~econ472/tutorial9.html
http://dss.princeton.edu/training/TS101.pdf#page=19

Panel data tests: heteroskedasticity and autocorrelation

- http://www.stata.com/support/faqs/stat/panel.html
- http://www.stata.com/support/faqs/stat/xtreg.html
- http://www.stata.com/support/faqs/stat/xt.html
- http://dss.princeton.edu/online_help/analysis/panel.htm

I can't read the output of my model!!! (links)

Data Analysis: Annotated Output
http://www.ats.ucla.edu/stat/AnnotatedOutput/default.htm

Data Analysis Examples
http://www.ats.ucla.edu/stat/dae/

Regression with Stata
http://www.ats.ucla.edu/STAT/stata/webbooks/reg/default.htm

Regression
http://www.ats.ucla.edu/stat/stata/topics/regression.htm

How to interpret dummy variables in a regression
http://www.ats.ucla.edu/stat/Stata/webbooks/reg/chapter3/statareg3.htm

How to create dummies
http://www.stata.com/support/faqs/data/dummy.html
http://www.ats.ucla.edu/stat/stata/faq/dummy.htm

Logit output: what are the odds ratios?
http://www.ats.ucla.edu/stat/stata/library/odds_ratio_logistic.htm

Topics in Statistics (links)

What statistical analysis should I use?
http://www.ats.ucla.edu/stat/mult_pkg/whatstat/default.htm

***Statnotes: Topics in Multivariate Analysis*, by G. David Garson**
http://www2.chass.ncsu.edu/garson/pa765/statnote.htm

Elementary Concepts in Statistics
http://www.statsoft.com/textbook/stathome.html

Introductory Statistics: Concepts, Models, and Applications
http://www.psychstat.missouristate.edu/introbook/sbk00.htm

Statistical Data Analysis
http://math.nicholls.edu/badie/statdataanalysis.html

Stata Library. Graph Examples (some may not work with STATA 10)
http://www.ats.ucla.edu/STAT/stata/library/GraphExamples/default.htm

Comparing Group Means: The T-test and One-way ANOVA Using STATA, SAS, and SPSS
http://www.indiana.edu/~statmath/stat/all/ttest/

Useful links / Recommended books

- DSS Online Training Section http://dss.princeton.edu/training/
- UCLA Resources to learn and use STATA http://www.ats.ucla.edu/stat/stata/
- DSS help-sheets for STATA http://dss/online_help/stats_packages/stata/stata.htm
- *Introduction to Stata* (PDF), Christopher F. Baum, Boston College, USA. "A 67-page description of Stata, its key features and benefits, and other useful information." http://fmwww.bc.edu/GStat/docs/StataIntro.pdf
- STATA FAQ website http://stata.com/support/faqs/
- Princeton DSS Libguides http://libguides.princeton.edu/dss

Books

- *Introduction to econometrics* / James H. Stock, Mark W. Watson. 2nd ed., Boston: Pearson Addison Wesley, 2007.
- *Data analysis using regression and multilevel/hierarchical models* / Andrew Gelman, Jennifer Hill. Cambridge ; New York : Cambridge University Press, 2007.
- Applied Regression Analysis and Generalized Linear Models, Second Edition. John Fox, Sage, 2008
- *Econometric analysis* / William H. Greene. 6th ed., Upper Saddle River, N.J. : Prentice Hall, 2008.
- *Designing Social Inquiry: Scientific Inference in Qualitative Research* / *Gary King, Robert O. Keohane, Sidney Verba, Princeton University Press*, 1994.
- *Unifying Political Methodology: The Likelihood Theory of Statistical Inference* / Gary King, Cambridge University Press, 1989
- *Statistical Analysis: an interdisciplinary introduction to univariate & multivariate methods* / Sam Kachigan, New York : Radius Press, c1986
- *Statistics with Stata (updated for version 9)* / Lawrence Hamilton, Thomson Books/Cole, 2006