MSc Social Research Methods

Pre-sessional Statistics Bootcamp

Burak Sonmez & Dingeman Wiertz

UCL – September 2022

Before We Start

- Log into your computer with your SSO

- Go to the pre-sessional Moodle page and download:
  - The slides for this bootcamp
  - The dataset we will use (“WDI_Data.dta”)

- Make sure to save the dataset in an easy-to-access folder
  - A place you can also access from elsewhere (e.g., N-drive)

Objectives of This Bootcamp

Overarching goal:

Ensuring we have some sort of common starting point + that we are ready to take on the quantitative MSc modules

For this purpose, we will:

- Refresh key statistical concepts

  E.g., sampling distributions, hypothesis testing

- Obtain basic familiarity with Stata

Statistical Inference

Statistical inference is used to learn from incomplete/imperfect data

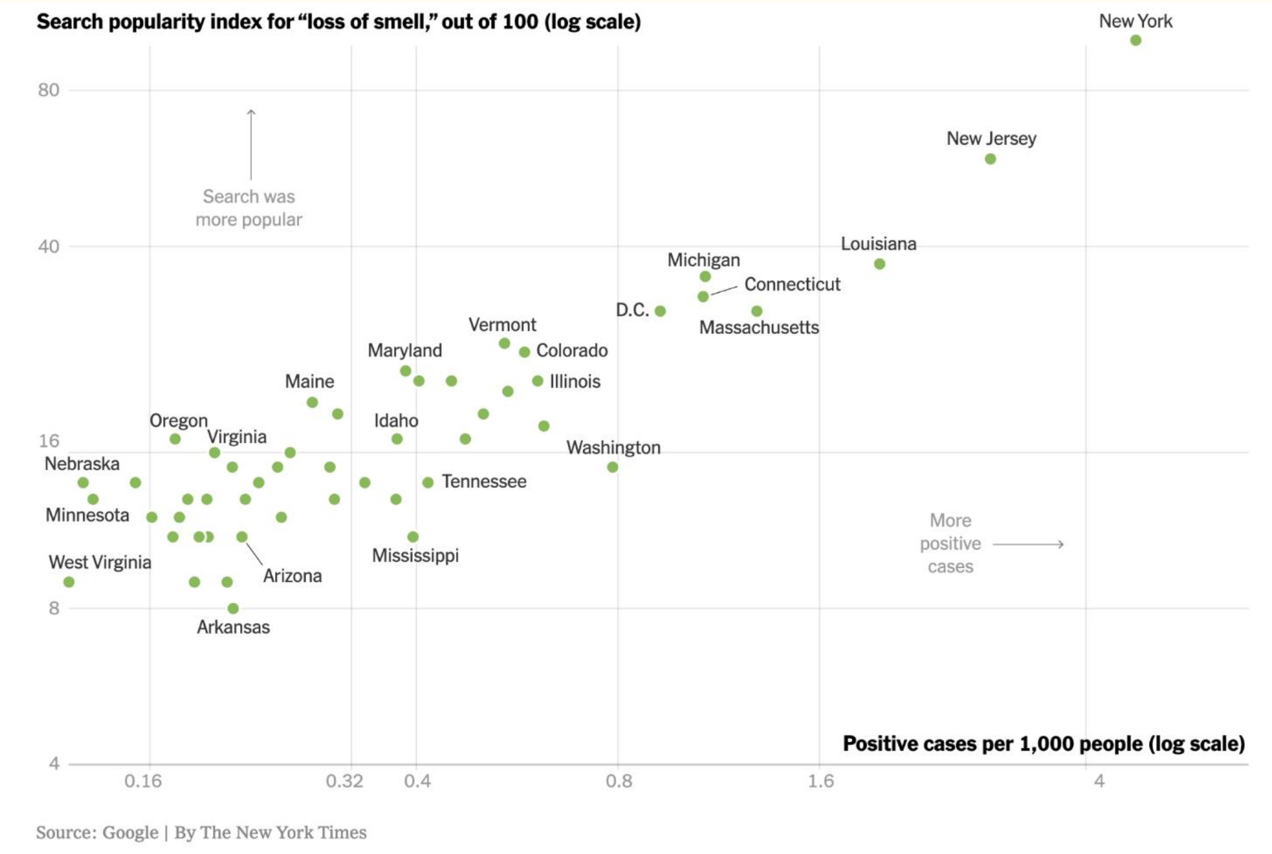
One way of thinking about statistical inference:

We wish to learn some characteristics of a population (e.g., the mean and standard deviation of the heights of all women in the UK), which we must estimate from a sample or subset of that population

Two types of statistical inference:

1. **Descriptive inference**: What is going on, or what exists?
2. **Causal inference**: Why is something going on, why does it exist?

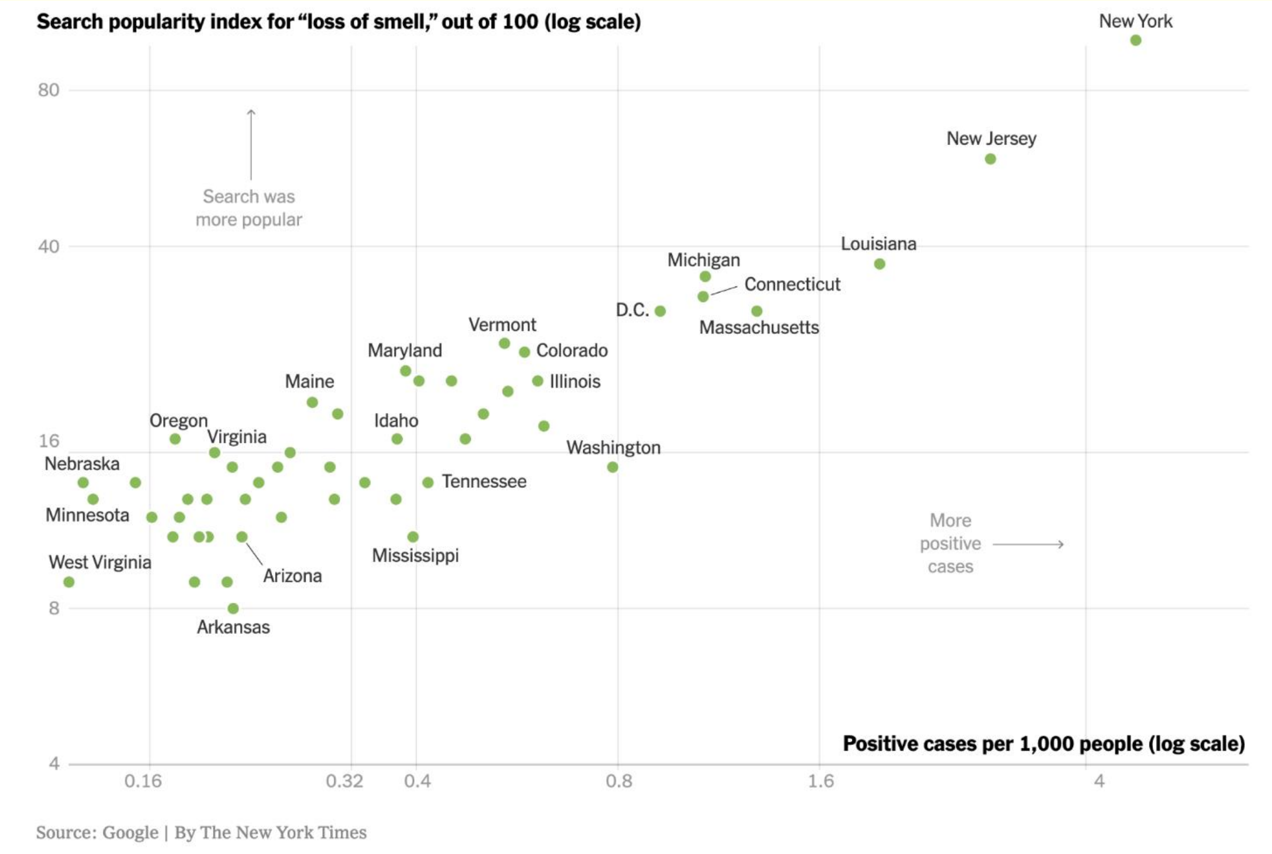
Descriptive vs. Causal Inference



Source: Google | By The New York Times

Descriptive inference: Does the search popularity of "loss of smell" increase as Covid cases are higher?

Descriptive vs. Causal Inference



Source: Google | By The New York Times

Causal inference: Why do people search more for "loss of smell" in some American states than in others?

Variables and Observations

A **variable** is anything that can vary across units of analysis:

It is a characteristic that can have multiple values / categories

E.g., sex, age, diameter, financial revenue, temperature

A **unit of analysis** is the major entity that you are analyzing:

E.g., individuals, objects, schools, countries

An **observation** is the value of a particular variable for a particular unit (sometimes a unit is in its entirety referred to as observation)

E.g., King Charles III is 73 years of age

Different Types of Variables

Continuous / interval-ratio variables:

They have an ordering, they can take on infinitely many values, and you can do calculations with them

E.g., income, age, weight, time

Categorical variables:

Each observation belongs to one out of a fixed number of categories

1. Ordinal variables: there is a natural ordering of the categories
2. Nominal variables: there is no natural ordering of the categories

E.g., education level, Likert scales, gender, vote choice

Introducing Stata

Stata is a powerful software package that allows you to do:

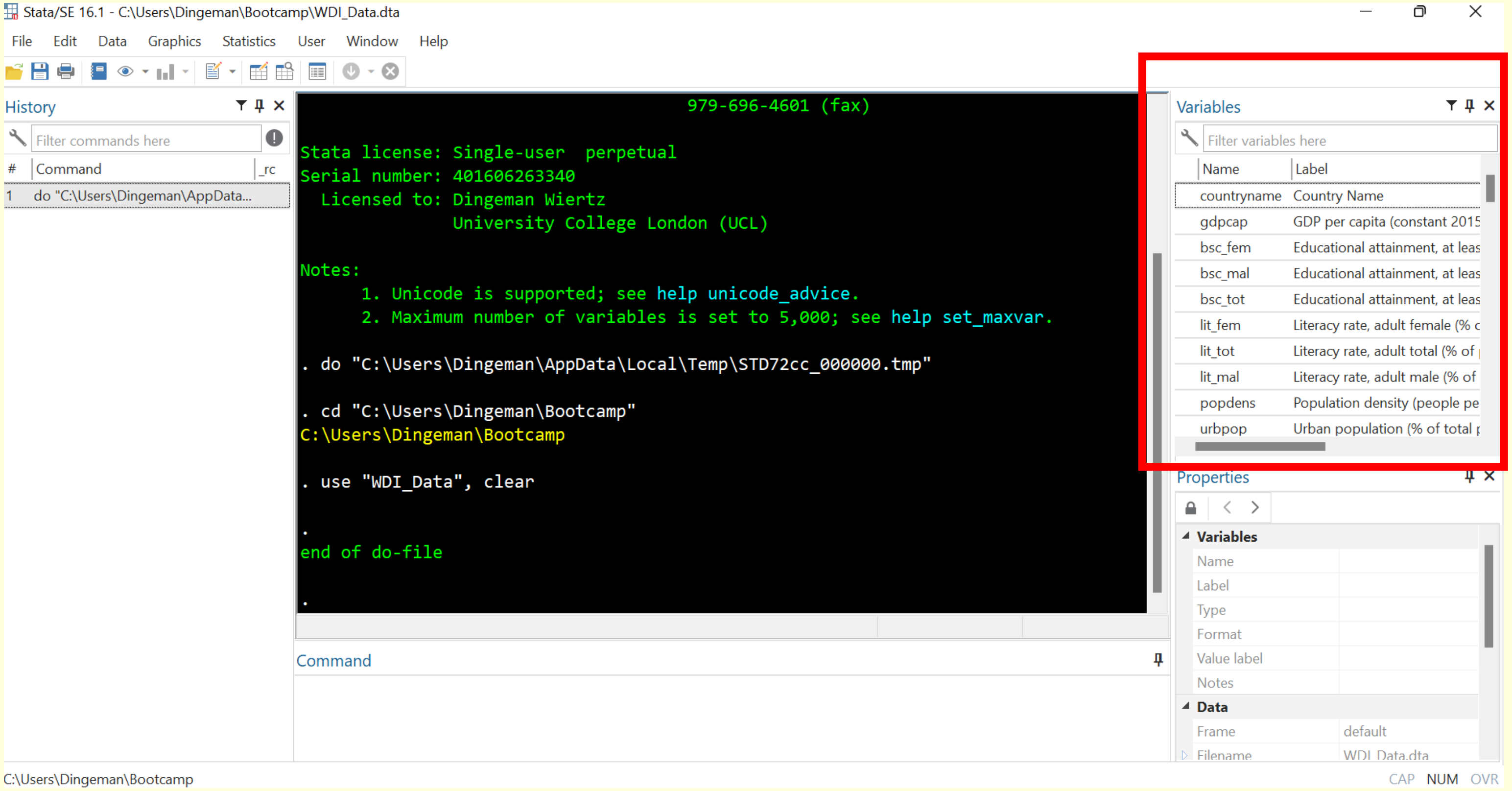
- Data management and manipulation
- Data visualization
- Statistical analysis

Stata has a compact command syntax that facilitates reproducibility

Other tools to conduct data analysis include **R**: a programming language and environment for statistical computing and graphics

- To some degree, Stata skills are transferable to such alternative platforms

Stata’s Interface: Variables Window



#	Command	_rc
1	do "C:\Users\Dingeman\AppData...	

```
                                      979-696-4601 (fax)

Stata license: Single-user  perpetual
Serial number: 401606263340
  Licensed to: Dingeman Wiertz
               University College London (UCL)

Notes:
      1. Unicode is supported; see help unicode_advice.
      2. Maximum number of variables is set to 5,000; see help set_maxvar.

. do "C:\Users\Dingeman\AppData\Local\Temp\STD72cc_000000.tmp"

. cd "C:\Users\Dingeman\Bootcamp"
C:\Users\Dingeman\Bootcamp

. use "WDI_Data", clear

.
end of do-file

.
```

Name	Label
countryname	Country Name
gdpcap	GDP per capita (constant 2015
bsc_fem	Educational attainment, at leas
bsc_mal	Educational attainment, at leas
bsc_tot	Educational attainment, at leas
lit_fem	Literacy rate, adult female (% c
lit_tot	Literacy rate, adult total (% of
lit_mal	Literacy rate, adult male (% of
popdens	Population density (people pe
urbpop	Urban population (% of total p

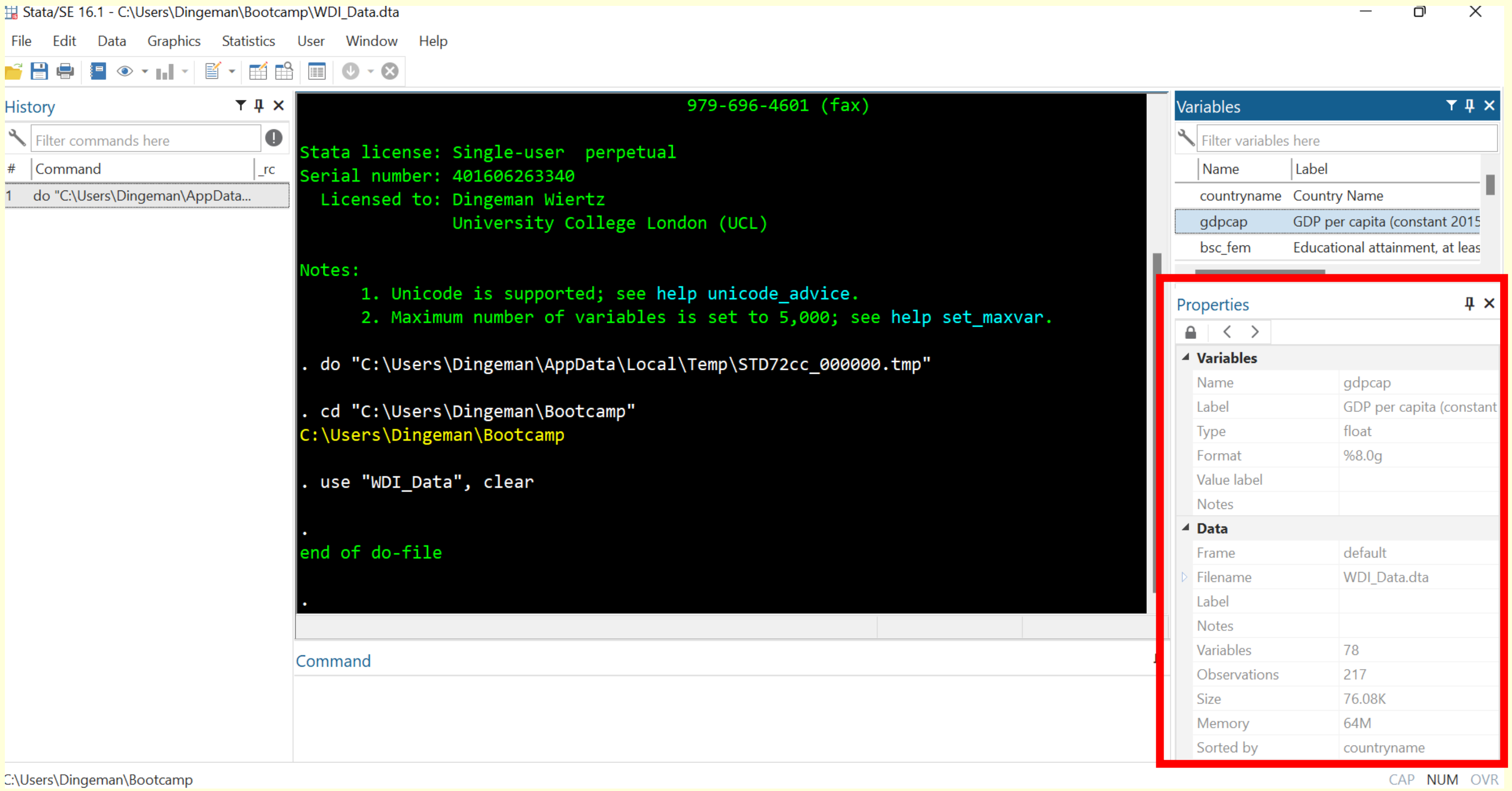
Variables	
Name	
Label	
Type	
Format	
Value label	
Notes	
Data	
Frame	default
Filename	WDI_Data.dta

C:\Users\Dingeman\Bootcamp

The <u>variables</u> window displays all variables in your dataset

- Single click on variable names to see details in properties window
- Double click to make variables appear in command window

Stata’s Interface: Properties Window



#	Command	_rc
1	do "C:\Users\Dingeman\AppData...	

```
                                 979-696-4601 (fax)

Stata license: Single-user  perpetual
Serial number: 401606263340
  Licensed to: Dingeman Wiertz
               University College London (UCL)

Notes:
      1. Unicode is supported; see help unicode_advice.
      2. Maximum number of variables is set to 5,000; see help set_maxvar.

. do "C:\Users\Dingeman\AppData\Local\Temp\STD72cc_000000.tmp"

. cd "C:\Users\Dingeman\Bootcamp"
C:\Users\Dingeman\Bootcamp

. use "WDI_Data", clear

.
end of do-file

.
```

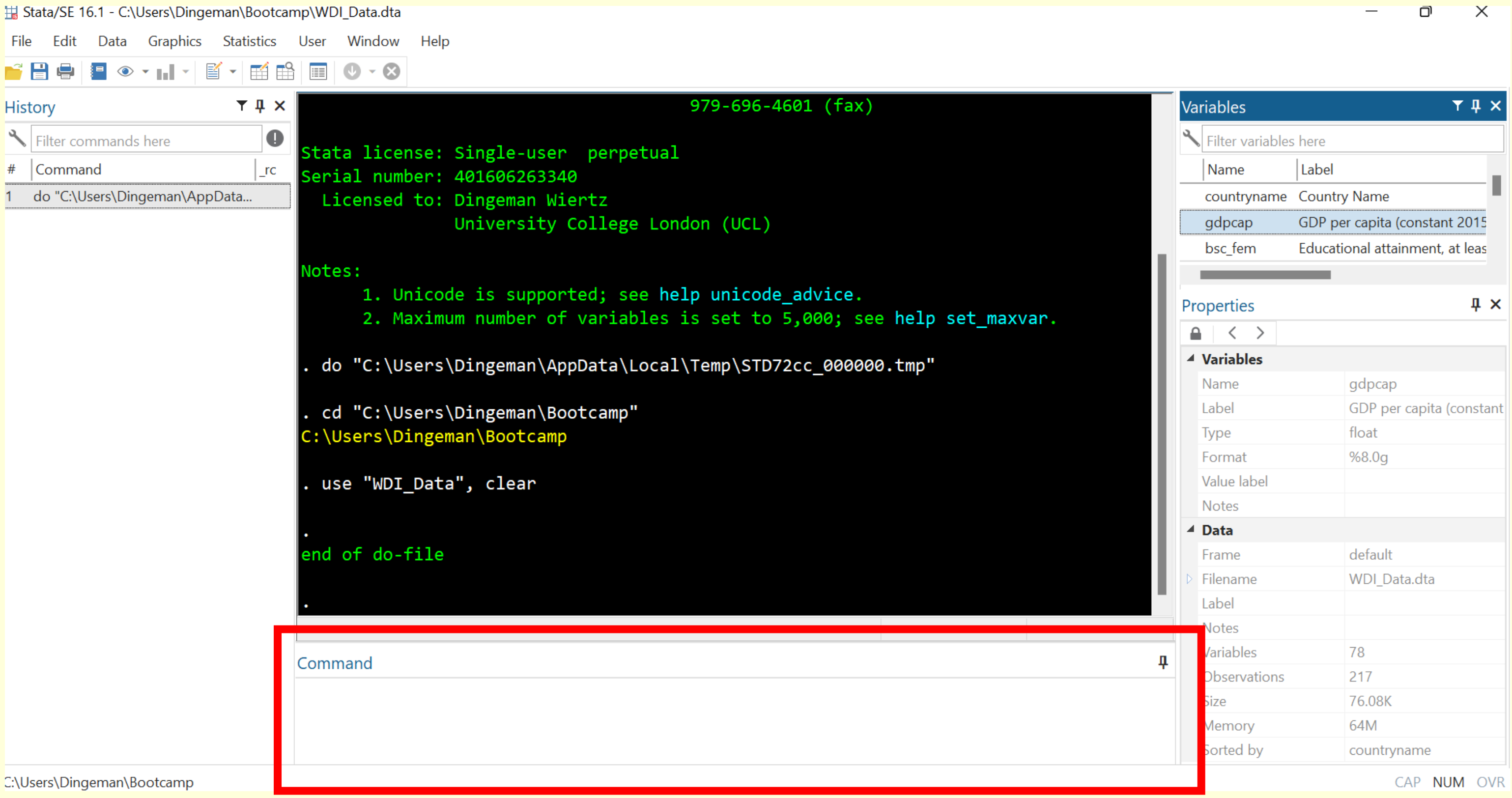
Name	Label
countryname	Country Name
gdpcap	GDP per capita (constant 2015
bsc_fem	Educational attainment, at leas

Variables	
Name	gdpcap
Label	GDP per capita (constant
Type	float
Format	%8.0g
Value label	
Notes	
Data	
Frame	default
Filename	WDI_Data.dta
Label	
Notes	
Variables	78
Observations	217
Size	76.08K
Memory	64M
Sorted by	countryname

C:\Users\Dingeman\Bootcamp

The properties window displays details about selected variables as well as the entire dataset (e.g., number of observations, sort order)

Stata’s Interface: Command Window



#	Command	_rc
1	do "C:\Users\Dingeman\AppData...	

```
                                        979-696-4601 (fax)

Stata license: Single-user  perpetual
Serial number: 401606263340
  Licensed to: Dingeman Wiertz
               University College London (UCL)

Notes:
      1. Unicode is supported; see help unicode_advice.
      2. Maximum number of variables is set to 5,000; see help set_maxvar.

. do "C:\Users\Dingeman\AppData\Local\Temp\STD72cc_000000.tmp"

. cd "C:\Users\Dingeman\Bootcamp"
C:\Users\Dingeman\Bootcamp

. use "WDI_Data", clear

.
end of do-file

.
```

Name	Label
countryname	Country Name
gdpcap	GDP per capita (constant 2015
bsc_fem	Educational attainment, at leas

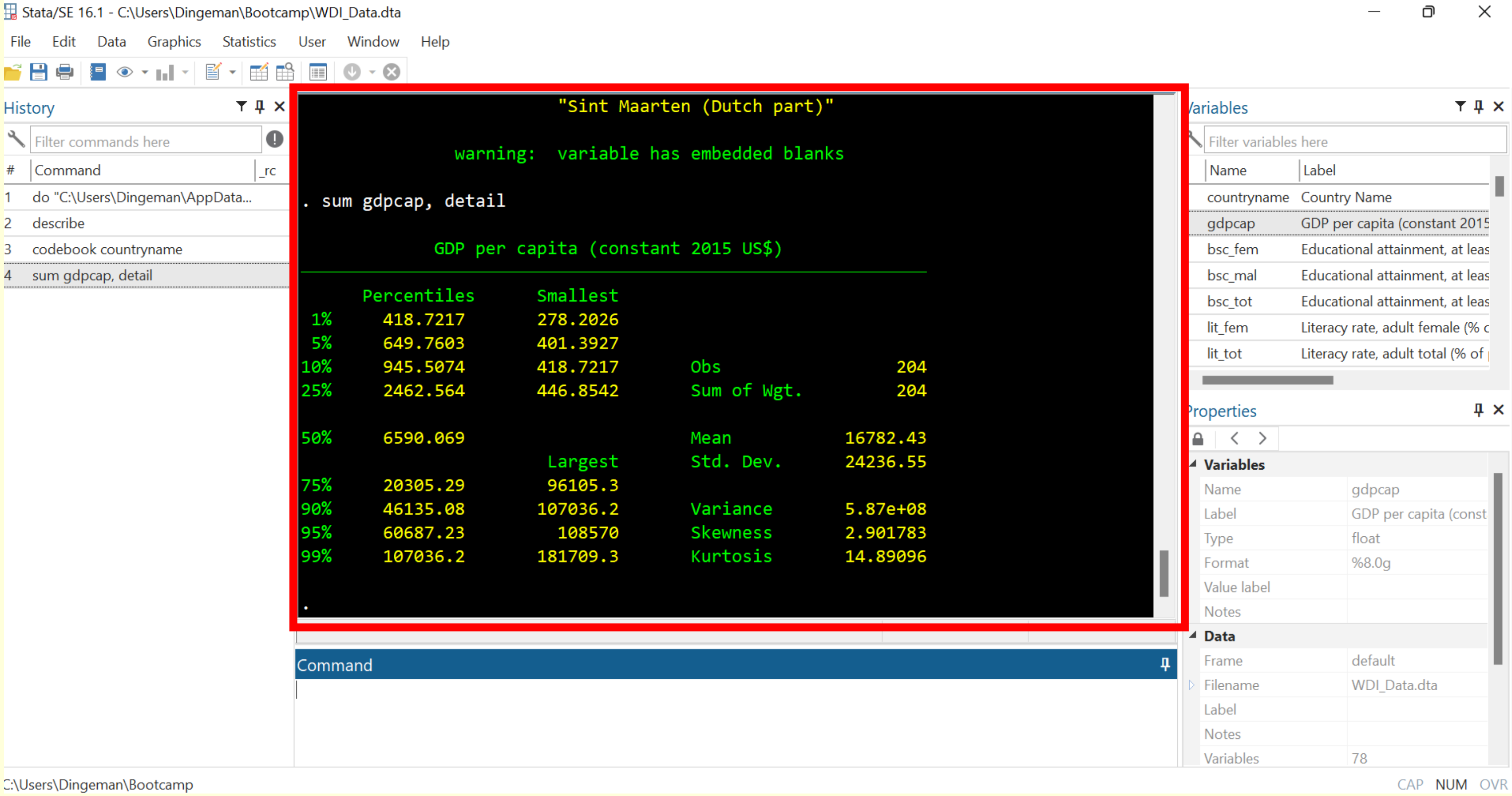
Variables	
Name	gdpcap
Label	GDP per capita (constant
Type	float
Format	%8.0g
Value label	
Notes	
Data	
Frame	default
Filename	WDI_Data.dta
Label	
Notes	
Variables	78
Observations	217
Size	76.08K
Memory	64M
Sorted by	countryname

C:\Users\Dingeman\Bootcamp

The command window is for entering and executing commands

- But it is better to use do-files (same applies to drop-down menus)

Stata’s Interface: Results Window



#	Command	_rc
1	do "C:\Users\Dingeman\AppData...	
2	describe	
3	codebook countryname	
4	sum gdpcap, detail	

```
                    "Sint Maarten (Dutch part)"

                 warning:  variable has embedded blanks

. sum gdpcap, detail

                 GDP per capita (constant 2015 US$)
-------------------------------------------------------------
      Percentiles      Smallest
 1%     418.7217       278.2026
 5%     649.7603       401.3927
10%     945.5074       418.7217       Obs                 204
25%     2462.564       446.8542       Sum of Wgt.         204

50%     6590.069                      Mean           16782.43
                        Largest       Std. Dev.      24236.55
75%     20305.29        96105.3
90%     46135.08       107036.2       Variance       5.87e+08
95%     60687.23         108570       Skewness       2.901783
99%     107036.2       181709.3       Kurtosis       14.89096

.
```

Name	Label
countryname	Country Name
gdpcap	GDP per capita (constant 2015
bsc_fem	Educational attainment, at leas
bsc_mal	Educational attainment, at leas
bsc_tot	Educational attainment, at leas
lit_fem	Literacy rate, adult female (% c
lit_tot	Literacy rate, adult total (% of

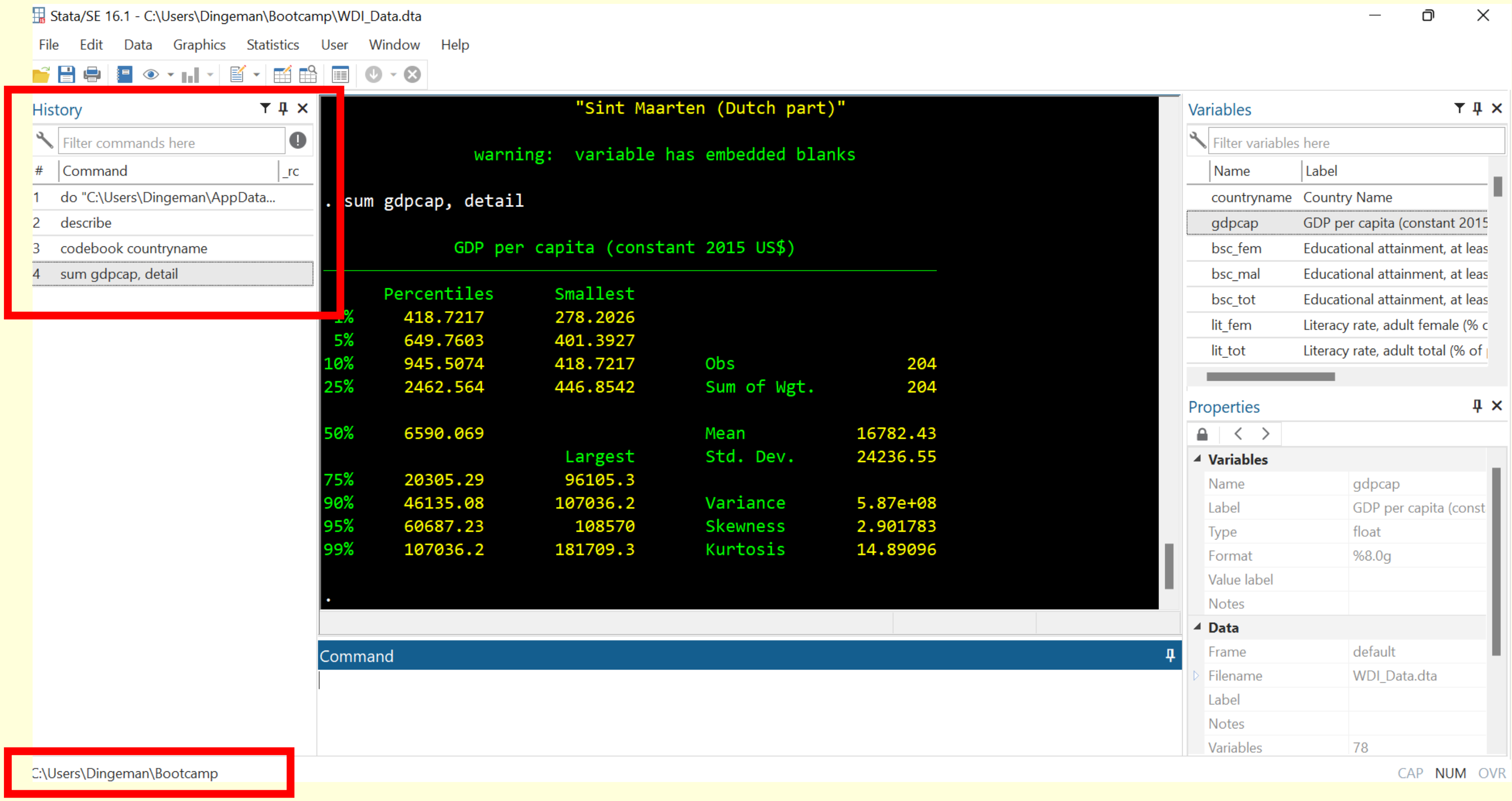
Variables	
Name	gdpcap
Label	GDP per capita (const
Type	float
Format	%8.0g
Value label	
Notes	
Data	
Frame	default
Filename	WDI_Data.dta
Label	
Notes	
Variables	78

:\Users\Dingeman\Bootcamp

The results window displays all output of your commands

- To change colour scheme: Select Edit from drop-down menu → Preferences → General preferences → Results → Classic

Stata's Interface: Command History and Current Working Directory



Stata/SE 16.1 - C:\Users\Dingeman\Bootcamp\WDI_Data.dta

#	Command	_rc
1	do "C:\Users\Dingeman\AppData...	
2	describe	
3	codebook countryname	
4	sum gdpcap, detail	

```
                         "Sint Maarten (Dutch part)"

                    warning:  variable has embedded blanks

. sum gdpcap, detail

                  GDP per capita (constant 2015 US$)
-------------------------------------------------------------
      Percentiles      Smallest
 1%    418.7217       278.2026
 5%    649.7603       401.3927
10%    945.5074       418.7217       Obs                 204
25%    2462.564       446.8542       Sum of Wgt.         204

50%    6590.069                      Mean           16782.43
                        Largest      Std. Dev.      24236.55
75%    20305.29        96105.3
90%    46135.08       107036.2       Variance       5.87e+08
95%    60687.23         108570       Skewness       2.901783
99%    107036.2       181709.3       Kurtosis       14.89096

.
```

Name	Label
countryname	Country Name
gdpcap	GDP per capita (constant 2015
bsc_fem	Educational attainment, at leas
bsc_mal	Educational attainment, at leas
bsc_tot	Educational attainment, at leas
lit_fem	Literacy rate, adult female (% c
lit_tot	Literacy rate, adult total (% of

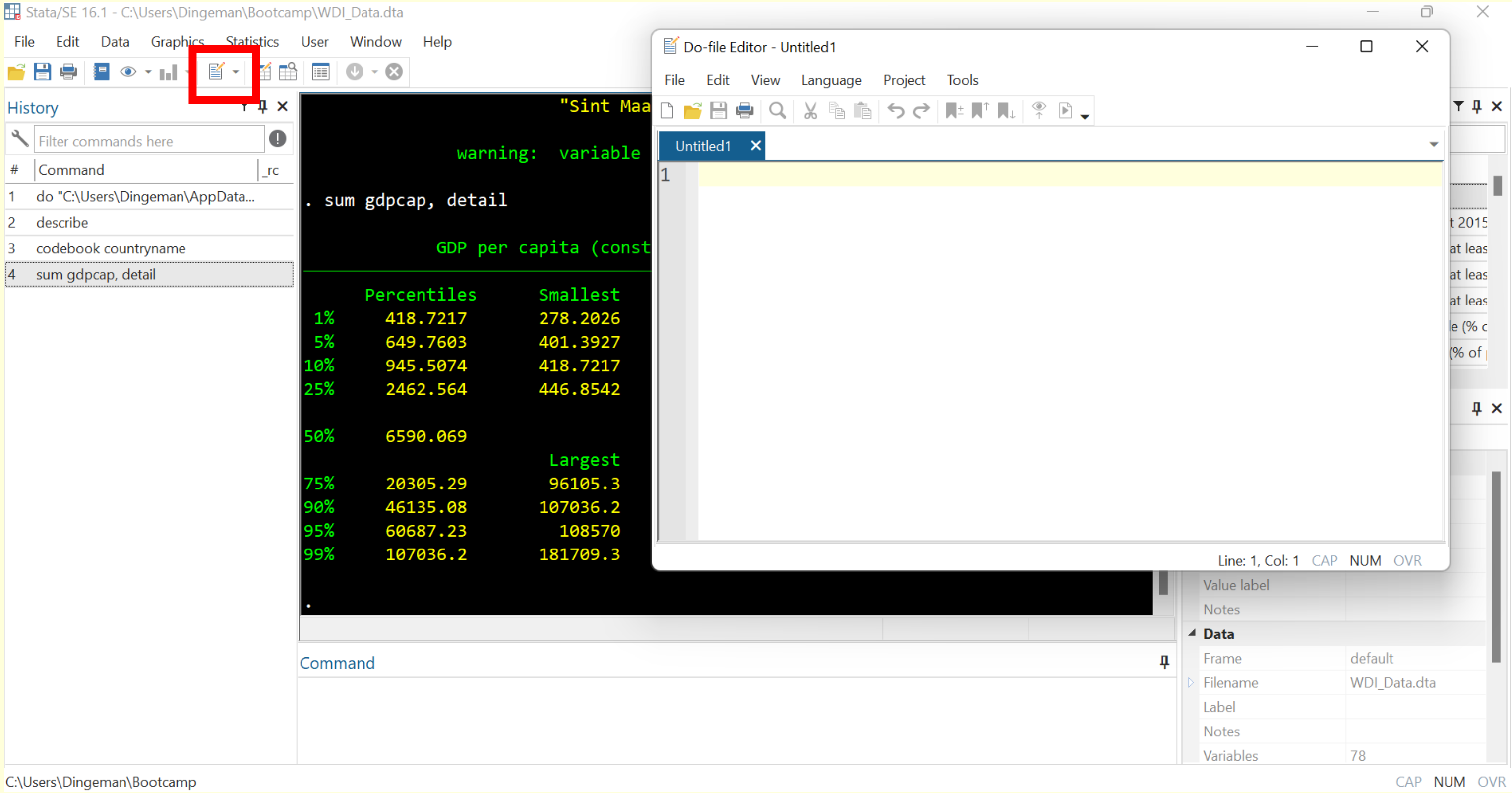
Variables	
Name	gdpcap
Label	GDP per capita (const
Type	float
Format	%8.0g
Value label	
Notes	
Data	
Frame	default
Filename	WDI_Data.dta
Label	
Notes	
Variables	78

C:\Users\Dingeman\Bootcamp

The command history window lists previously run commands

At the bottom you can see the current working directory: the folder where any files will be loaded from and saved to

Stata’s Interface: Opening a Do File



#	Command	_rc
1	do "C:\Users\Dingeman\AppData...	
2	describe	
3	codebook countryname	
4	sum gdpcap, detail	

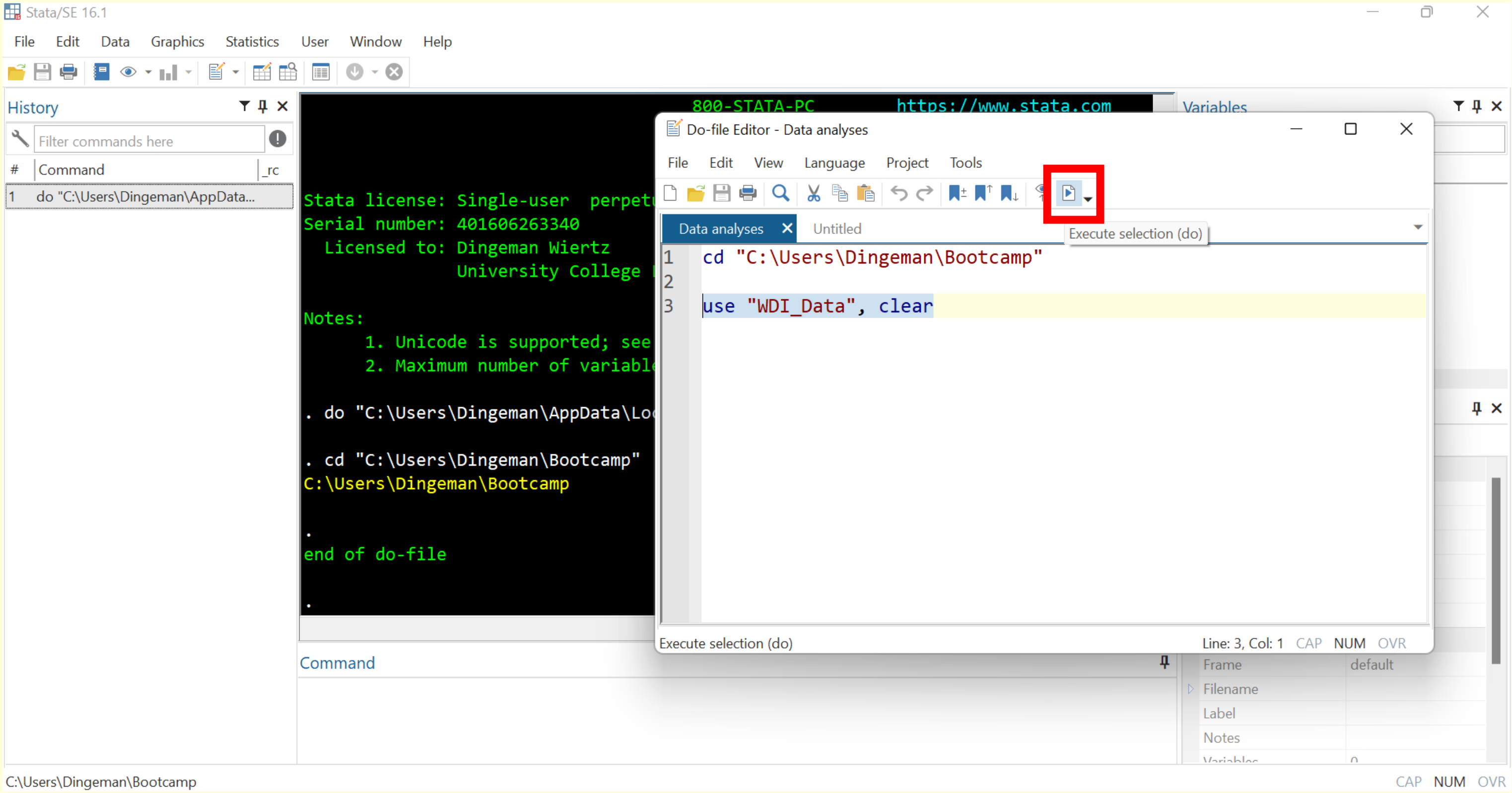
```
                                 "Sint Maa

           warning:  variable

. sum gdpcap, detail

                         GDP per capita (const
-------------------------------------------------
      Percentiles      Smallest
 1%     418.7217       278.2026
 5%     649.7603       401.3927
10%     945.5074       418.7217
25%     2462.564       446.8542

50%     6590.069
                        Largest
75%     20305.29        96105.3
90%     46135.08       107036.2
95%     60687.23         108570
99%     107036.2       181709.3

.
```

Frame: default
Filename: WDI_Data.dta
Variables: 78

Do files are text files where you can store commands for reuse

- Huge payoffs for reproducibility, debugging, adapting commands

Running Commands from a Do File



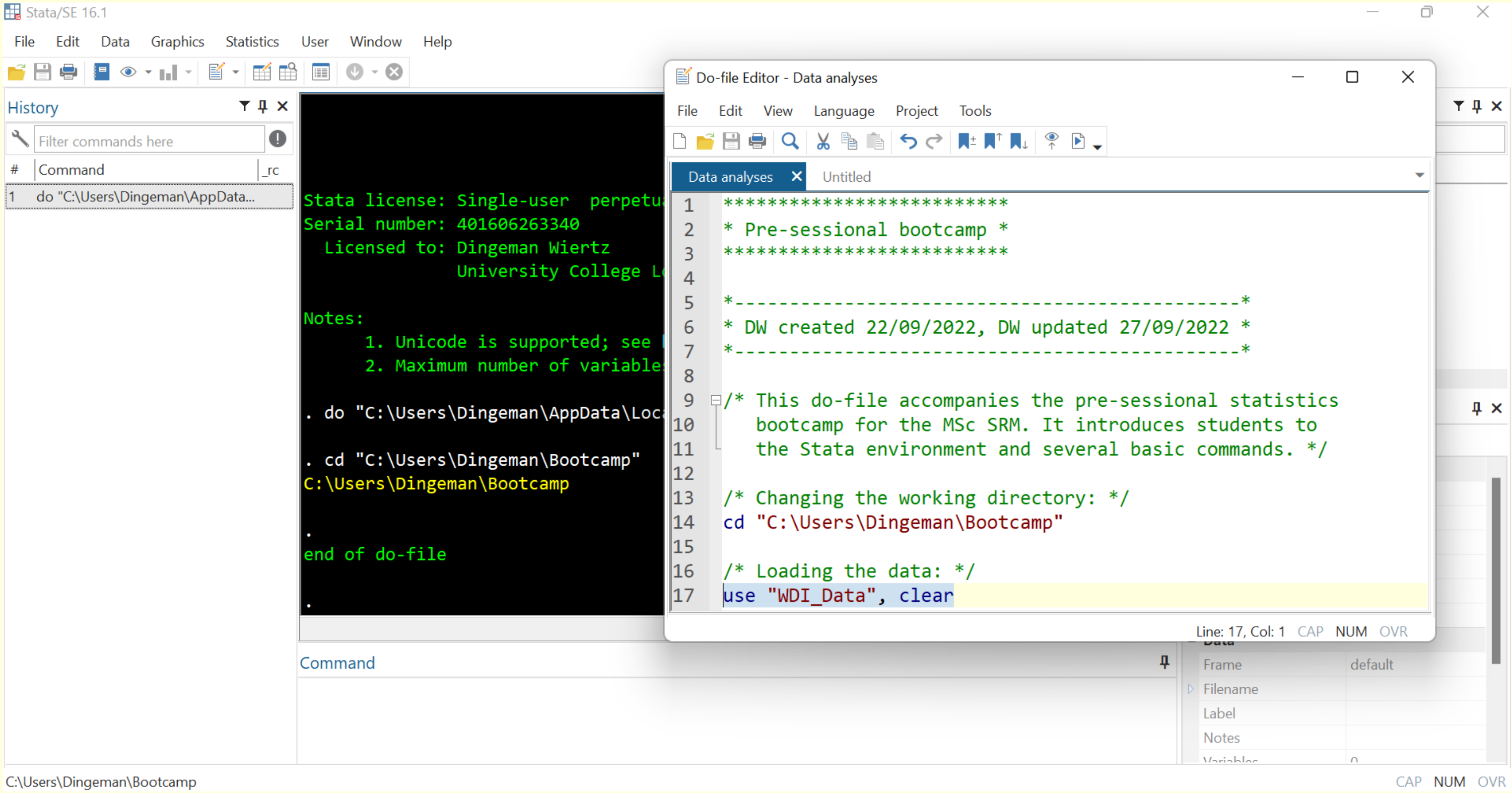
Stata/SE 16.1
File Edit Data Graphics Statistics User Window Help
History
Filter commands here
# Command _rc
1 do "C:\Users\Dingeman\AppData...
Do-file Editor - Data analyses
File Edit View Language Project Tools
Data analyses
Untitled
Execute selection (do)
Line: 3, Col: 1 CAP NUM OVR
Command
Variables
Frame default
Filename
Label
Notes
C:\Users\Dingeman\Bootcamp
CAP NUM OVR

```
                                   800-STATA-PC        https://www.stata.com

Stata license: Single-user  perpetu
Serial number: 401606263340
  Licensed to: Dingeman Wiertz
               University College 

Notes:
      1. Unicode is supported; see
      2. Maximum number of variabl

. do "C:\Users\Dingeman\AppData\Lo

. cd "C:\Users\Dingeman\Bootcamp"
C:\Users\Dingeman\Bootcamp

.
end of do-file

.
```

```
cd "C:\Users\Dingeman\Bootcamp"

use "WDI_Data", clear
```

After entering a command, you select it, and then click the “execute” button or press “Ctrl+D”

Do File Dos and Don’ts



#	Command	_rc
1	do "C:\Users\Dingeman\AppData...	

```
Stata license: Single-user  perpetu
Serial number: 401606263340
  Licensed to: Dingeman Wiertz
               University College L

Notes:
      1. Unicode is supported; see
      2. Maximum number of variable

. do "C:\Users\Dingeman\AppData\Loc

. cd "C:\Users\Dingeman\Bootcamp"
C:\Users\Dingeman\Bootcamp

.
end of do-file

.
```

```
****************************
* Pre-sessional bootcamp *
****************************

*-----------------------------------------------*
* DW created 22/09/2022, DW updated 27/09/2022 *
*-----------------------------------------------*

/* This do-file accompanies the pre-sessional statistics
   bootcamp for the MSc SRM. It introduces students to
   the Stata environment and several basic commands. */

/* Changing the working directory: */
cd "C:\Users\Dingeman\Bootcamp"

/* Loading the data: */
use "WDI_Data", clear
```

Line: 17, Col: 1

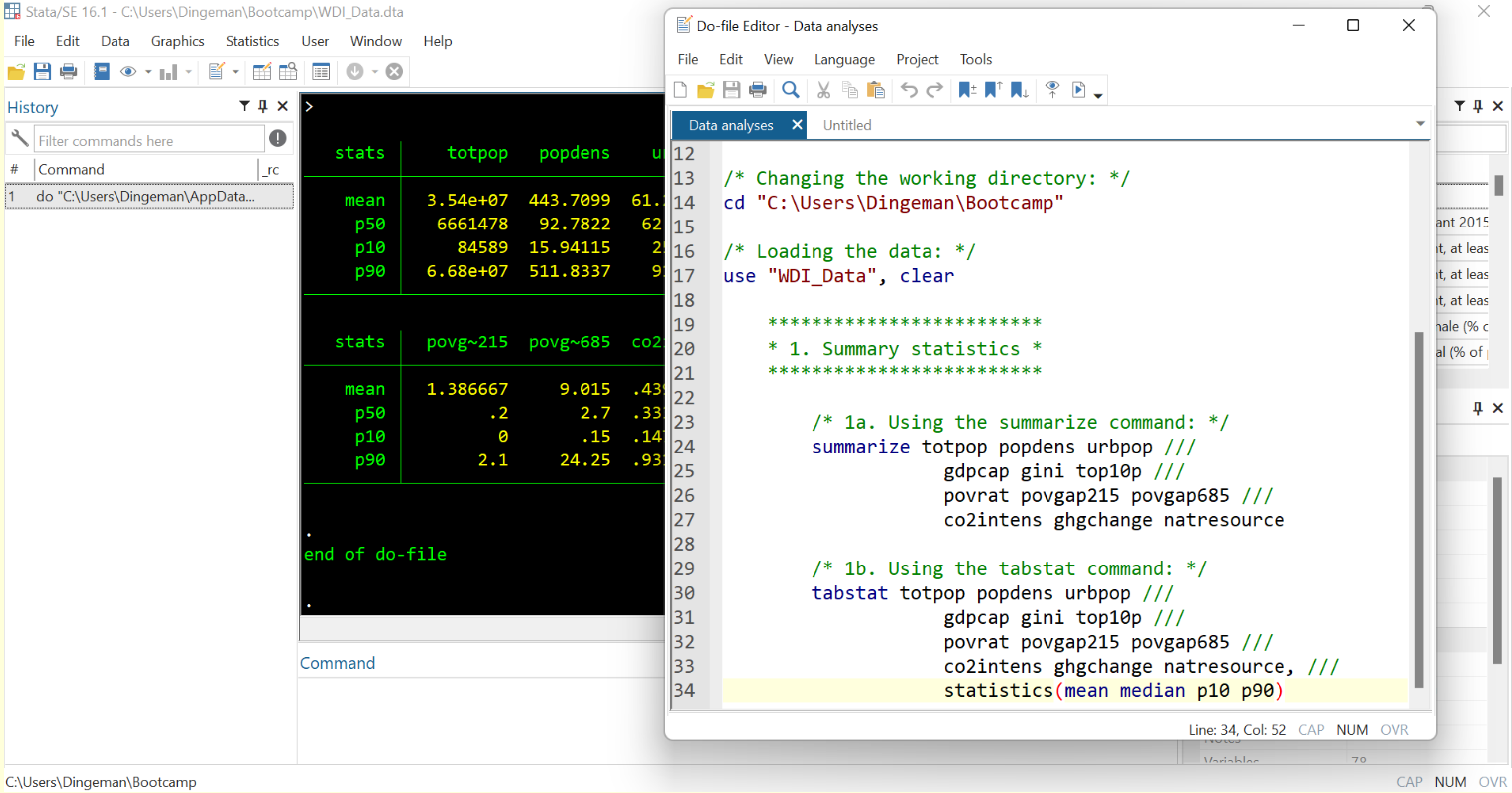
Frame	default
Filename	
Label	
Notes	
Variables	0

C:\Users\Dingeman\Bootcamp

1. Use annotations to facilitate replicability (incl. for future self!):

   Use * for single-line comments and /* */ for multiple lines

Do File Dos and Don’ts

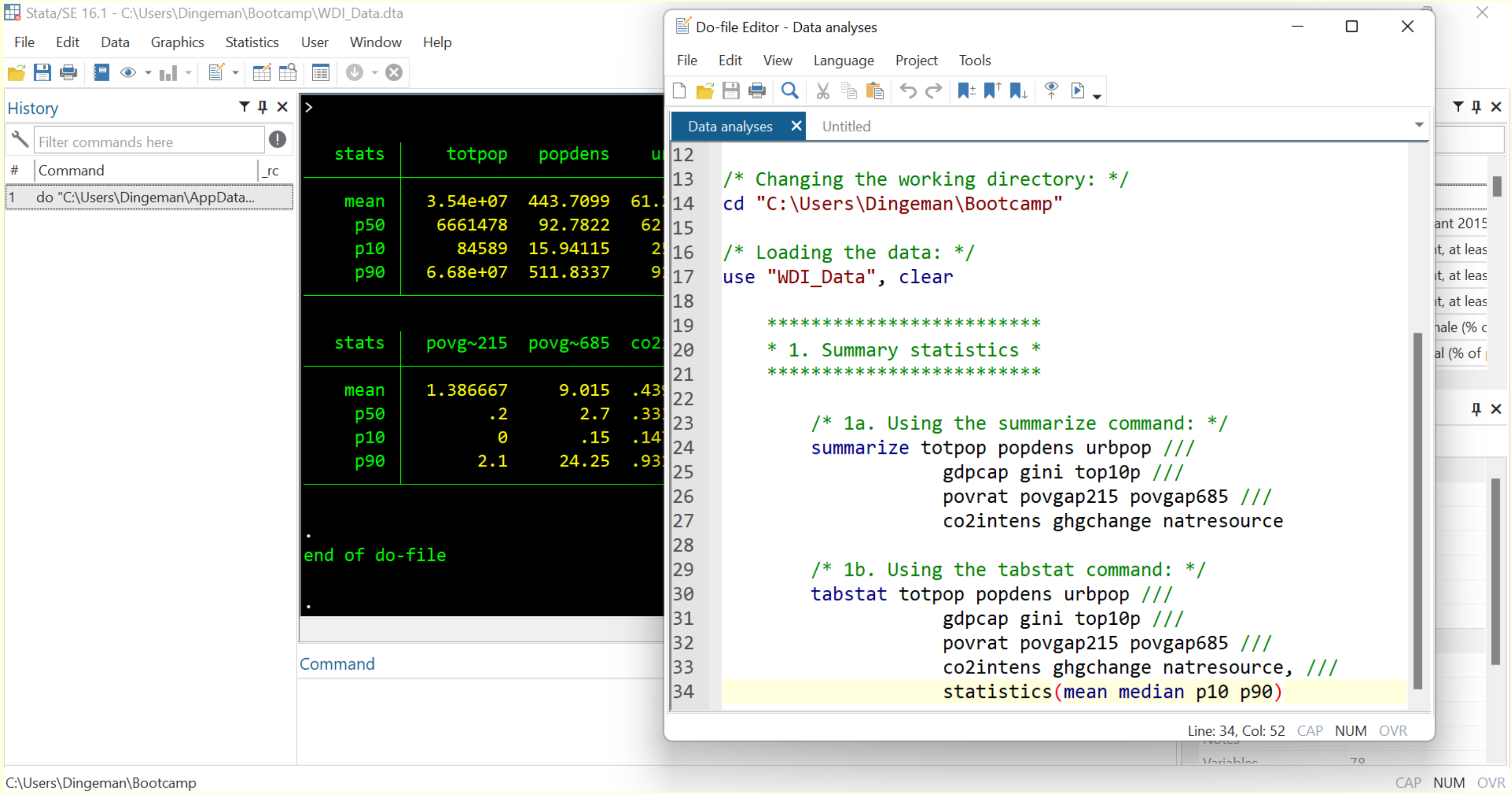

Stata/SE 16.1 - C:\Users\Dingeman\Bootcamp\WDI_Data.dta
File Edit Data Graphics Statistics User Window Help
History
Filter commands here
# Command _rc
1 do "C:\Users\Dingeman\AppData...
Command
C:\Users\Dingeman\Bootcamp
CAP NUM OVR
Do-file Editor - Data analyses
File Edit View Language Project Tools
Data analyses
Untitled
Line: 34, Col: 52 CAP NUM OVR

```
>

   stats |    totpop   popdens     u
---------+---------------------------
    mean |  3.54e+07  443.7099   61.
     p50 |   6661478   92.7822    62
     p10 |     84589  15.94115     2
     p90 |  6.68e+07  511.8337     9
---------+---------------------------

   stats |  povg~215  povg~685   co2
---------+---------------------------
    mean |  1.386667     9.015   .43
     p50 |        .2       2.7   .33
     p10 |         0       .15   .14
     p90 |       2.1     24.25   .93
---------------------------------------

.
end of do-file

.
```

```

/* Changing the working directory: */
cd "C:\Users\Dingeman\Bootcamp"

/* Loading the data: */
use "WDI_Data", clear

    *************************
    * 1. Summary statistics *
    *************************

        /* 1a. Using the summarize command: */
        summarize totpop popdens urbpop ///
                    gdpcap gini top10p ///
                    povrat povgap215 povgap685 ///
                    co2intens ghgchange natresource

        /* 1b. Using the tabstat command: */
        tabstat totpop popdens urbpop ///
                    gdpcap gini top10p ///
                    povrat povgap215 povgap685 ///
                    co2intens ghgchange natresource, ///
                    statistics(mean median p10 p90)
```

2. Break down code into clearly labelled sections / subsections
3. Use tab indentations to making things easy to read

Do File Dos and Don’ts



Stata/SE 16.1 - C:\Users\Dingeman\Bootcamp\WDI_Data.dta
File Edit Data Graphics Statistics User Window Help
History
Filter commands here
# Command _rc
1 do "C:\Users\Dingeman\AppData...
Command
C:\Users\Dingeman\Bootcamp
Do-file Editor - Data analyses
File Edit View Language Project Tools
Data analyses
Untitled
Line: 34, Col: 52 CAP NUM OVR

```
>

   stats |    totpop   popdens     u
---------+----------------------------
    mean |  3.54e+07  443.7099  61.
     p50 |   6661478   92.7822   62
     p10 |     84589  15.94115    2
     p90 |  6.68e+07  511.8337    9
---------+----------------------------

   stats |  povg~215  povg~685  co2
---------+----------------------------
    mean |  1.386667     9.015  .43
     p50 |        .2       2.7  .33
     p10 |         0       .15  .14
     p90 |       2.1     24.25  .93
--------------------------------------

.
end of do-file

.
```

```
/* Changing the working directory: */
cd "C:\Users\Dingeman\Bootcamp"

/* Loading the data: */
use "WDI_Data", clear

    *************************
    * 1. Summary statistics *
    *************************

        /* 1a. Using the summarize command: */
        summarize totpop popdens urbpop ///
                    gdpcap gini top10p ///
                    povrat povgap215 povgap685 ///
                    co2intens ghgchange natresource

        /* 1b. Using the tabstat command: */
        tabstat totpop popdens urbpop ///
                    gdpcap gini top10p ///
                    povrat povgap215 povgap685 ///
                    co2intens ghgchange natresource, ///
                    statistics(mean median p10 p90)
```

4. Don’t put too much information on a single line:

   Use /// to continue your command on the next line and write “top-to-bottom” instead of “left-to-right”

Now Have a Look Around Yourself…

Useful commands to get started:

- `cd`: sets your working directory
  - Where any data will be loaded from and saved to
  - E.g., `cd "C:\Users\Dingeman\Bootcamp"`

- `use`: opens a stored dataset
  - Dataset needs to be in .dta format
  - Add `clear` option if you already had another dataset open
  - E.g., `use "WDI_Data", clear`

Now Have a Look Around Yourself…

Useful commands to get started:

- `describe`: provides a description of variables
  - ➢ If you specify no variables, the command will be applied to all
  - ➢ E.g., `describe forest co2intens` or `describe`

- `codebook`: examines labels, data etc. to produce a codebook
  - ➢ If you specify no variables, command will be applied to all
  - ➢ E.g., `codebook countrycode femparl` or `codebook`

Now Have a Look Around Yourself…

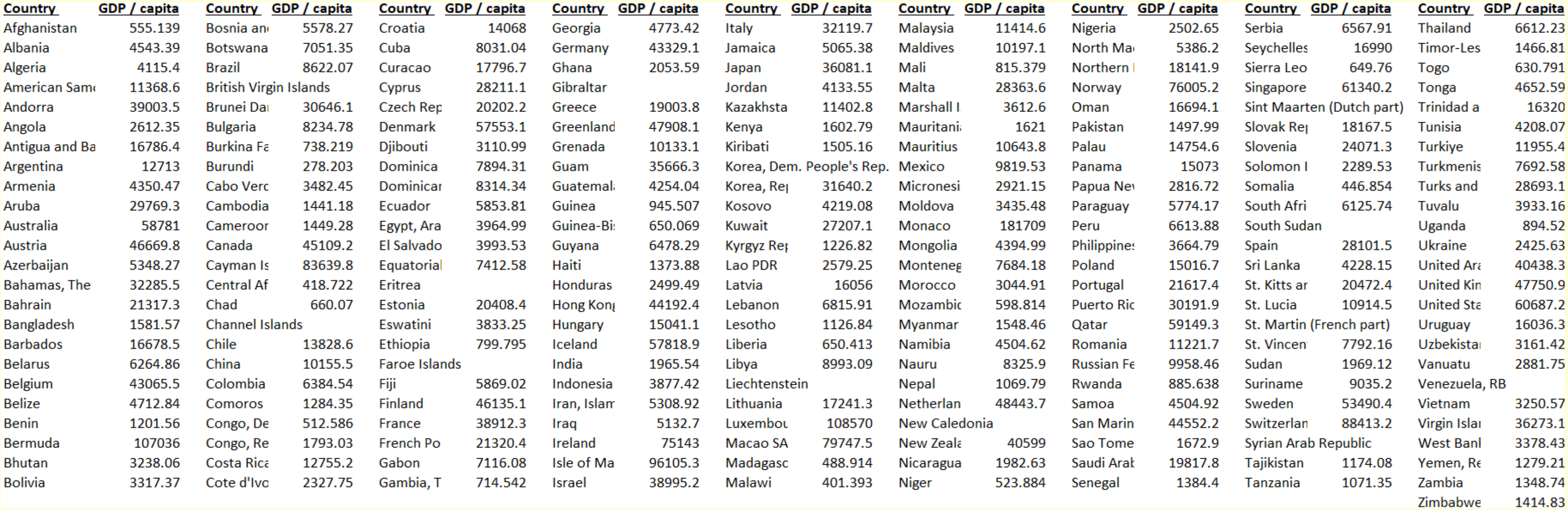
Some questions to guide your explorations:

- How many variables are there in this dataset?
- How many countries are identified in this dataset?
- For which continent do we observe the most countries and for which continent the fewest?
- What are the labels of the variables `femmanage` and `armyexp`?
- How many missing values do `femmanage` and `armyexp` have?
- What are the means for `femmanage` and `armyexp`?
- Which countries have the highest scores for `femmanage` and `armyexp`?

Describing Data

Describing data is necessary because there is usually too much of it, so it does not make any sense by itself

- E.g, what can you say about the world income distribution based on this table?



Country	GDP / capita	Country	GDP / capita	Country	GDP / capita	Country	GDP / capita	Country	GDP / capita	Country	GDP / capita	Country	GDP / capita	Country	GDP / capita	Country	GDP / capita
Afghanistan	555.139	Bosnia an	5578.27	Croatia	14068	Georgia	4773.42	Italy	32119.7	Malaysia	11414.6	Nigeria	2502.65	Serbia	6567.91	Thailand	6612.23
Albania	4543.39	Botswana	7051.35	Cuba	8031.04	Germany	43329.1	Jamaica	5065.38	Maldives	10197.1	North Ma	5386.2	Seychelles	16990	Timor-Les	1466.81
Algeria	4115.4	Brazil	8622.07	Curacao	17796.7	Ghana	2053.59	Japan	36081.1	Mali	815.379	Northern I	18141.9	Sierra Leo	649.76	Togo	630.791
American Sam	11368.6	British Virgin Islands		Cyprus	28211.1	Gibraltar		Jordan	4133.55	Malta	28363.6	Norway	76005.2	Singapore	61340.2	Tonga	4652.59
Andorra	39003.5	Brunei Da	30646.1	Czech Rep	20202.2	Greece	19003.8	Kazakhsta	11402.8	Marshall I	3612.6	Oman	16694.1	Sint Maarten (Dutch part)		Trinidad a	16320
Angola	2612.35	Bulgaria	8234.78	Denmark	57553.1	Greenland	47908.1	Kenya	1602.79	Mauritani	1621	Pakistan	1497.99	Slovak Re	18167.5	Tunisia	4208.07
Antigua and Ba	16786.4	Burkina Fa	738.219	Djibouti	3110.99	Grenada	10133.1	Kiribati	1505.16	Mauritius	10643.8	Palau	14754.6	Slovenia	24071.3	Turkiye	11955.4
Argentina	12713	Burundi	278.203	Dominica	7894.31	Guam	35666.3	Korea, Dem. People's Rep.		Mexico	9819.53	Panama	15073	Solomon I	2289.53	Turkmenis	7692.58
Armenia	4350.47	Cabo Verc	3482.45	Dominicar	8314.34	Guatemal	4254.04	Korea, Re	31640.2	Micronesi	2921.15	Papua Ne	2816.72	Somalia	446.854	Turks and	28693.1
Aruba	29769.3	Cambodia	1441.18	Ecuador	5853.81	Guinea	945.507	Kosovo	4219.08	Moldova	3435.48	Paraguay	5774.17	South Afri	6125.74	Tuvalu	3933.16
Australia	58781	Cameroor	1449.28	Egypt, Ara	3964.99	Guinea-Bi	650.069	Kuwait	27207.1	Monaco	181709	Peru	6613.88	South Sudan		Uganda	894.52
Austria	46669.8	Canada	45109.2	El Salvado	3993.53	Guyana	6478.29	Kyrgyz Re	1226.82	Mongolia	4394.99	Philippine	3664.79	Spain	28101.5	Ukraine	2425.63
Azerbaijan	5348.27	Cayman Is	83639.8	Equatorial	7412.58	Haiti	1373.88	Lao PDR	2579.25	Monteneg	7684.18	Poland	15016.7	Sri Lanka	4228.15	United Ara	40438.3
Bahamas, The	32285.5	Central Af	418.722	Eritrea		Honduras	2499.49	Latvia	16056	Morocco	3044.91	Portugal	21617.4	St. Kitts ar	20472.4	United Kin	47750.9
Bahrain	21317.3	Chad	660.07	Estonia	20408.4	Hong Kon	44192.4	Lebanon	6815.91	Mozambic	598.814	Puerto Ric	30191.9	St. Lucia	10914.5	United Sta	60687.2
Bangladesh	1581.57	Channel Islands		Eswatini	3833.25	Hungary	15041.1	Lesotho	1126.84	Myanmar	1548.46	Qatar	59149.3	St. Martin (French part)		Uruguay	16036.3
Barbados	16678.5	Chile	13828.6	Ethiopia	799.795	Iceland	57818.9	Liberia	650.413	Namibia	4504.62	Romania	11221.7	St. Vincen	7792.16	Uzbekista	3161.42
Belarus	6264.86	China	10155.5	Faroe Islands		India	1965.54	Libya	8993.09	Nauru	8325.9	Russian Fe	9958.46	Sudan	1969.12	Vanuatu	2881.75
Belgium	43065.5	Colombia	6384.54	Fiji	5869.02	Indonesia	3877.42	Liechtenstein		Nepal	1069.79	Rwanda	885.638	Suriname	9035.2	Venezuela, RB	
Belize	4712.84	Comoros	1284.35	Finland	46135.1	Iran, Islam	5308.92	Lithuania	17241.3	Netherlan	48443.7	Samoa	4504.92	Sweden	53490.4	Vietnam	3250.57
Benin	1201.56	Congo, De	512.586	France	38912.3	Iraq	5132.7	Luxembou	108570	New Caledonia		San Marin	44552.2	Switzerlan	88413.2	Virgin Isla	36273.1
Bermuda	107036	Congo, Re	1793.03	French Po	21320.4	Ireland	75143	Macao SA	79747.5	New Zeala	40599	Sao Tome	1672.9	Syrian Arab Republic		West Banl	3378.43
Bhutan	3238.06	Costa Rica	12755.2	Gabon	7116.08	Isle of Ma	96105.3	Madagasc	488.914	Nicaragua	1982.63	Saudi Aral	19817.8	Tajikistan	1174.08	Yemen, Re	1279.21
Bolivia	3317.37	Cote d'Ivo	2327.75	Gambia, T	714.542	Israel	38995.2	Malawi	401.393	Niger	523.884	Senegal	1384.4	Tanzania	1071.35	Zambia	1348.74
																Zimbabwe	1414.83

Describing Data

Describing data is necessary because there is usually too much of it, so it does not make any sense by itself

We thus have to look for ways to summarize central tendencies, variation, and relationships that exist in the data

There are many different ways to do this

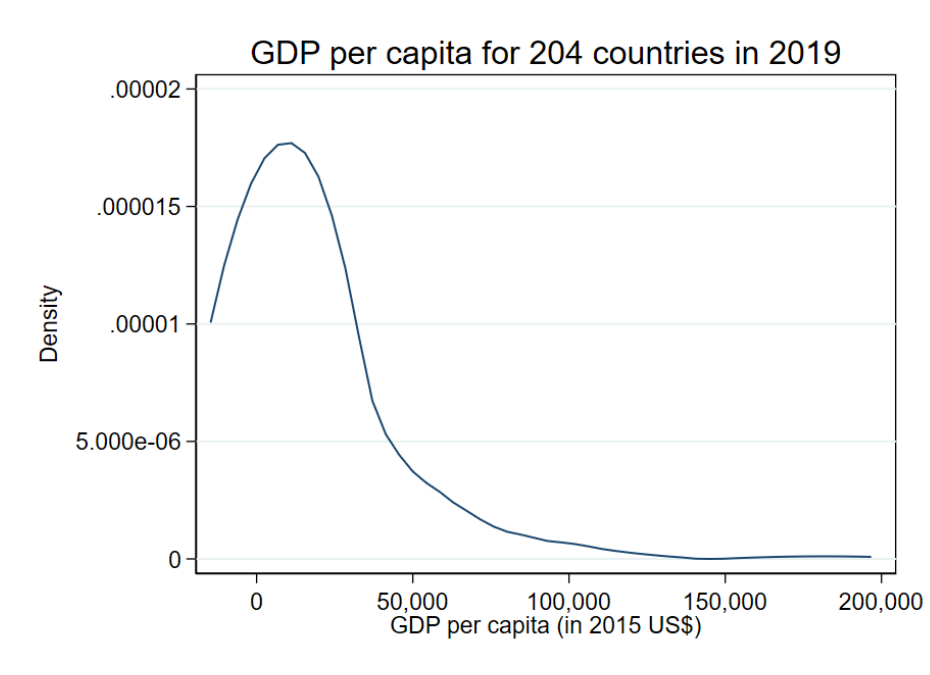
- Numerical descriptions
- Visual depictions

Distributions

Variables can be characterized by their **frequency distribution**:

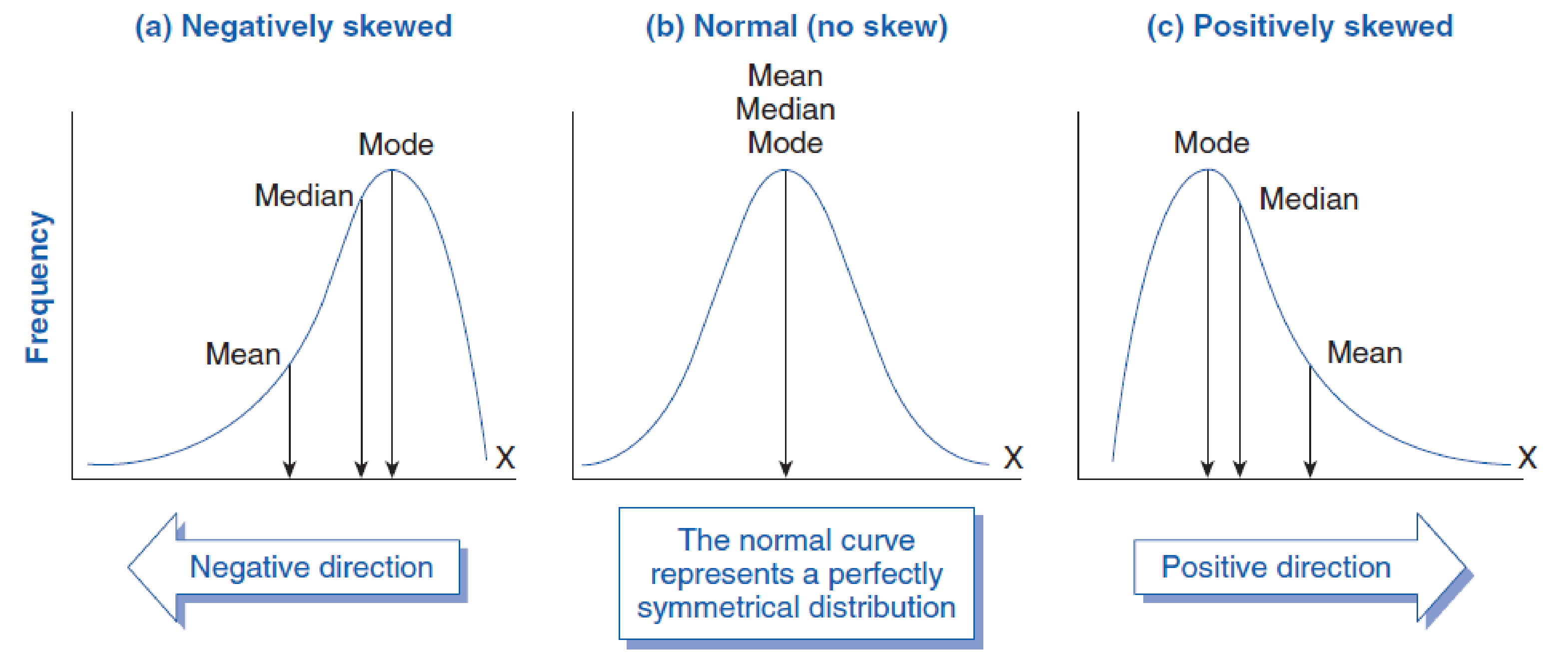
The distribution of the (relative) frequencies of their values

- In a sample or in the population
- E.g., we can graph the world income distribution:



Distributions

Distributions can take on many different shapes; some examples:



Measures of Central Tendency

Common measures to denote central tendencies of distributions:

1. **Mean**: conventional average calculated by adding all values for all units and dividing by the number of units
   - $Mean = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{N}$ with units i and number of units N
   - May give a distorted impression if there are outliers

2. **Median**: value that falls in the middle if we order all units by their value on the variable

3. **Mode**: most frequently occurring value across all units

Measures of Dispersion / Spread

1. **Variance**: average of the squared differences between each observed value on a variable and its mean
   - $Variance = s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N}$
   - Why the square? To treat + and – differences alike

2. **Standard deviation**: average departure of the observed values on a variable from its mean
   - $Standard\ deviation = s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N}}$
   - It is the square root of the variance; it “reverts” the square-taking in the variance calculation, to bring the statistic back to the original scale of the variable

A Note on Notation

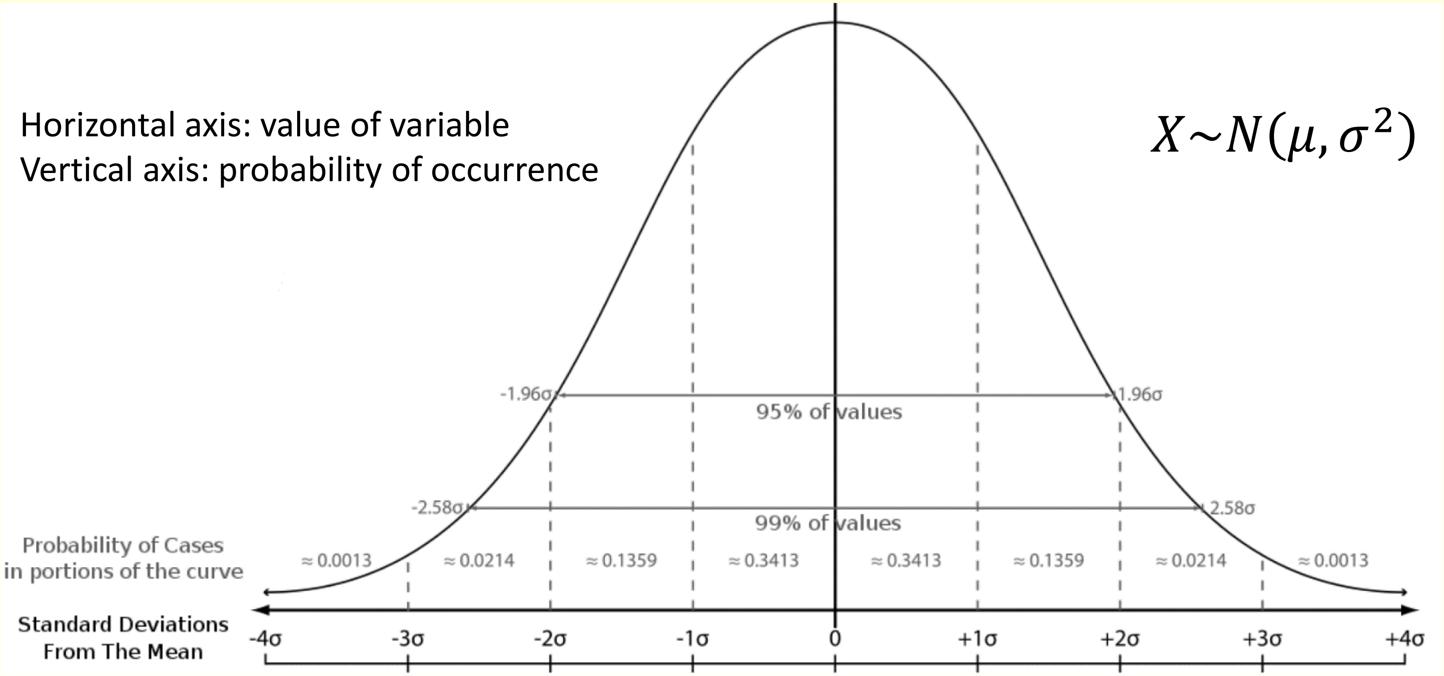
Typically, we use Roman letters for sample statistics and Greek letters for population statistics:

- Sample mean = $\bar{x}$, population mean = μ
- Sample variance = s^2, population variance = σ^2
- Sample standard error = s, population standard deviation = σ
- In a regression context:
  - Coefficient estimate based on sample = b
  - True underlying coefficient in the population = β

Recall: the sample is what we observe, the population is what we want to make inferences about

The Normal Distribution

If a variable X is normally distributed with mean μ and variance σ^2:



Horizontal axis: value of variable
Vertical axis: probability of occurrence

$X \sim N(\mu, \sigma^2)$

- The normal distribution is symmetric around μ and bell-shaped
- 95% of all cases lie between $X = \mu - 1.96\sigma$ and $X = \mu + 1.96\sigma$

Standard Normal Distribution

To calculate probabilities for a normal variable, we must **standardize** it by subtracting its mean and dividing by its standard deviation:

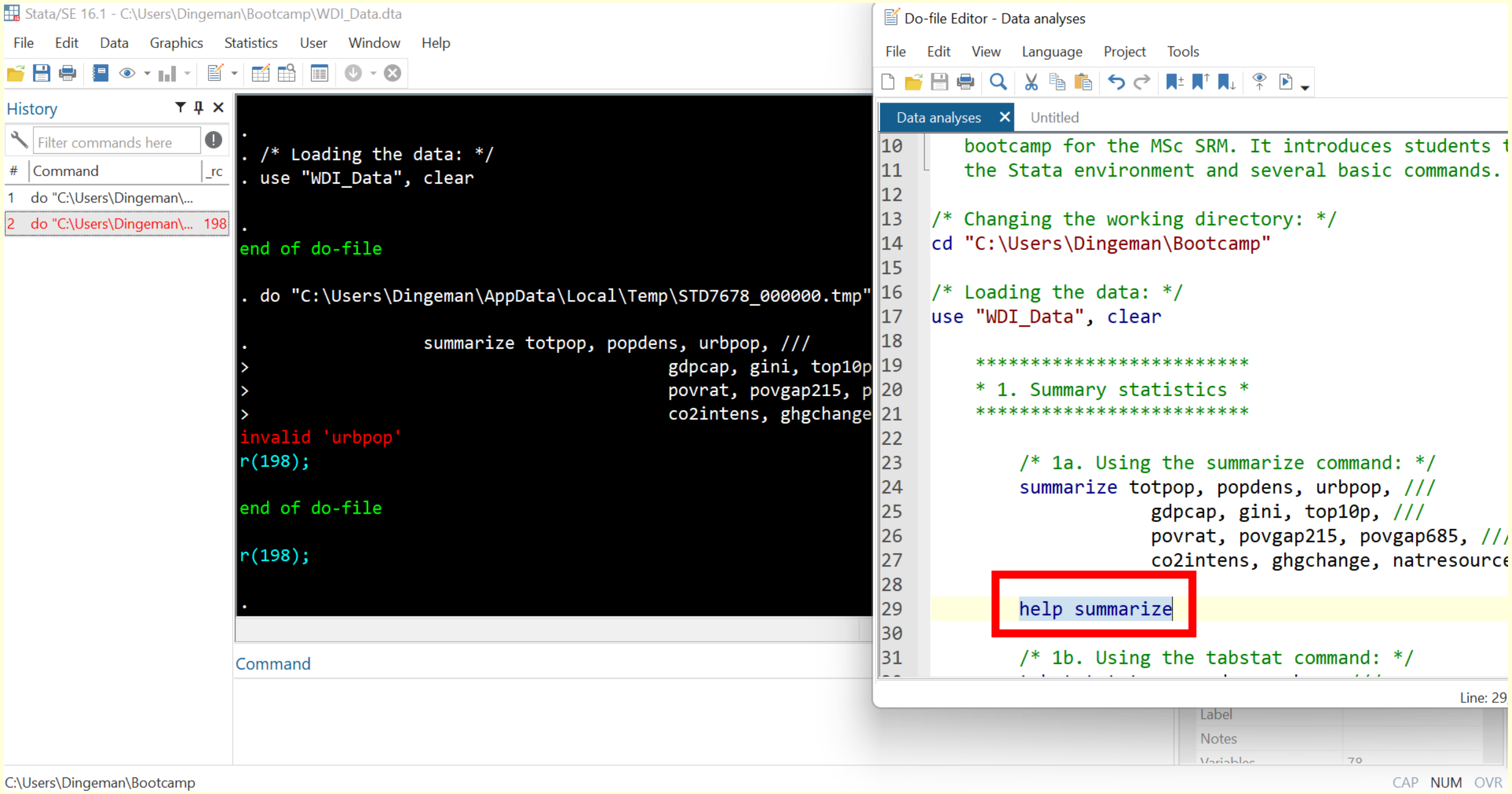
$$Z = \frac{X - \mu_x}{\sigma_x}$$

The thus obtained **Z-scores** indicate the number of standard deviations a value is from the mean of a distribution

The **standard normal distribution** is the normal distribution with mean 0 and variance 1

- Z-scores have a standard normal distribution: $Z{\sim}N(0,1)$
- Z-scores and are extensively used in hypothesis testing

Stata Help Files Are Your Friend


Stata/SE 16.1 - C:\Users\Dingeman\Bootcamp\WDI_Data.dta
File Edit Data Graphics Statistics User Window Help
History
Filter commands here
Command
Do-file Editor - Data analyses
File Edit View Language Project Tools
Data analyses
Untitled
Label
Notes
Variables
78
CAP NUM OVR

#	Command	_rc
1	do "C:\Users\Dingeman\...	
2	do "C:\Users\Dingeman\...	198

```
.
. /* Loading the data: */
. use "WDI_Data", clear

.
end of do-file

. do "C:\Users\Dingeman\AppData\Local\Temp\STD7678_000000.tmp"

.                 summarize totpop, popdens, urbpop, ///
>                             gdpcap, gini, top10p
>                             povrat, povgap215, p
>                             co2intens, ghgchange
invalid 'urbpop'
r(198);

end of do-file

r(198);

.
```

```
   bootcamp for the MSc SRM. It introduces students t
   the Stata environment and several basic commands.

/* Changing the working directory: */
cd "C:\Users\Dingeman\Bootcamp"

/* Loading the data: */
use "WDI_Data", clear

    ************************
    * 1. Summary statistics *
    ************************

        /* 1a. Using the summarize command: */
        summarize totpop, popdens, urbpop, ///
                    gdpcap, gini, top10p, ///
                    povrat, povgap215, povgap685, //
                    co2intens, ghgchange, natresource

        help summarize

        /* 1b. Using the tabstat command: */
```

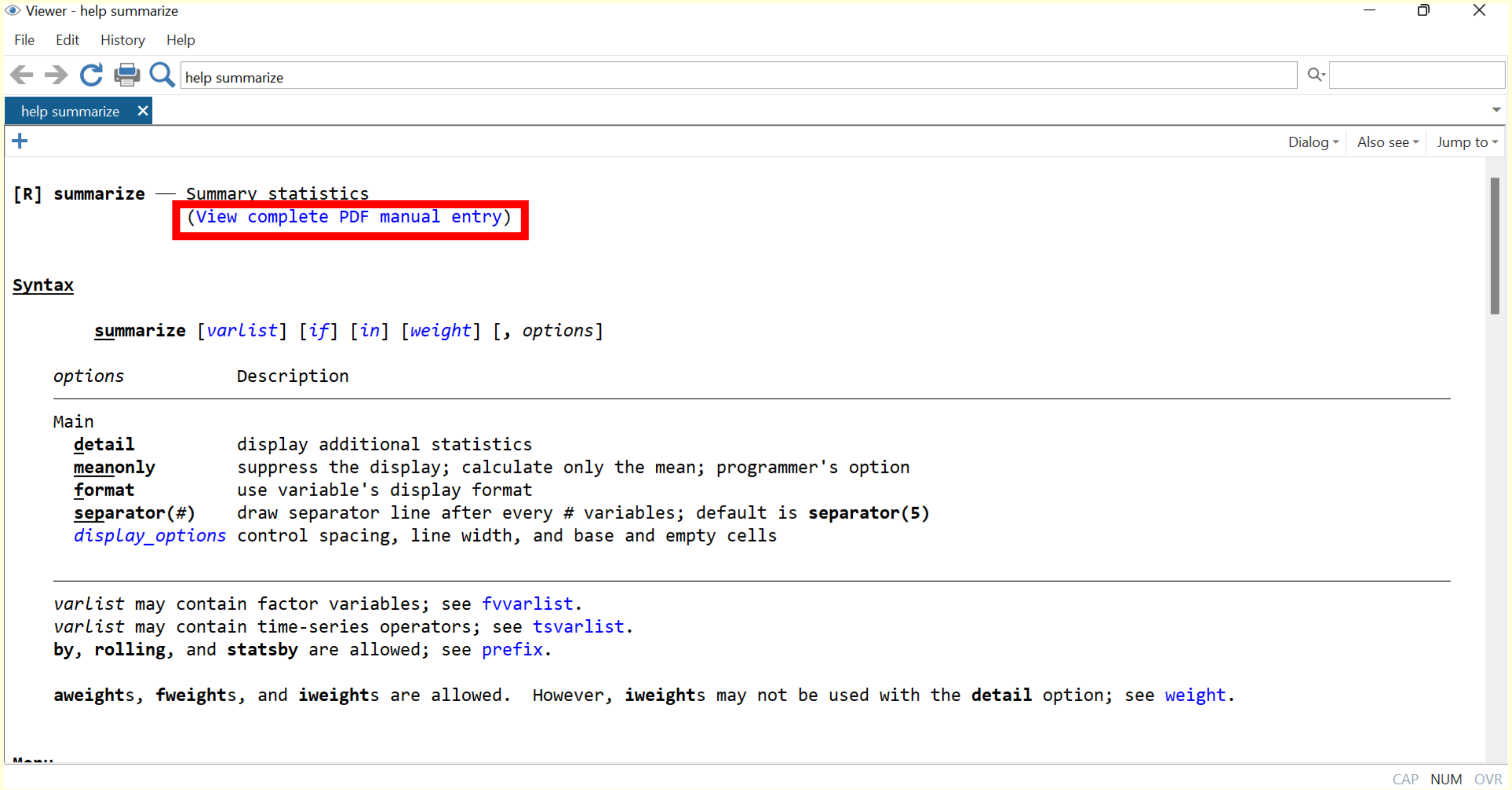
Line: 29

C:\Users\Dingeman\Bootcamp

If you forgot how a command works or receive an error message, access Stata’s help files by running: `help` …

(inserting the relevant command name)

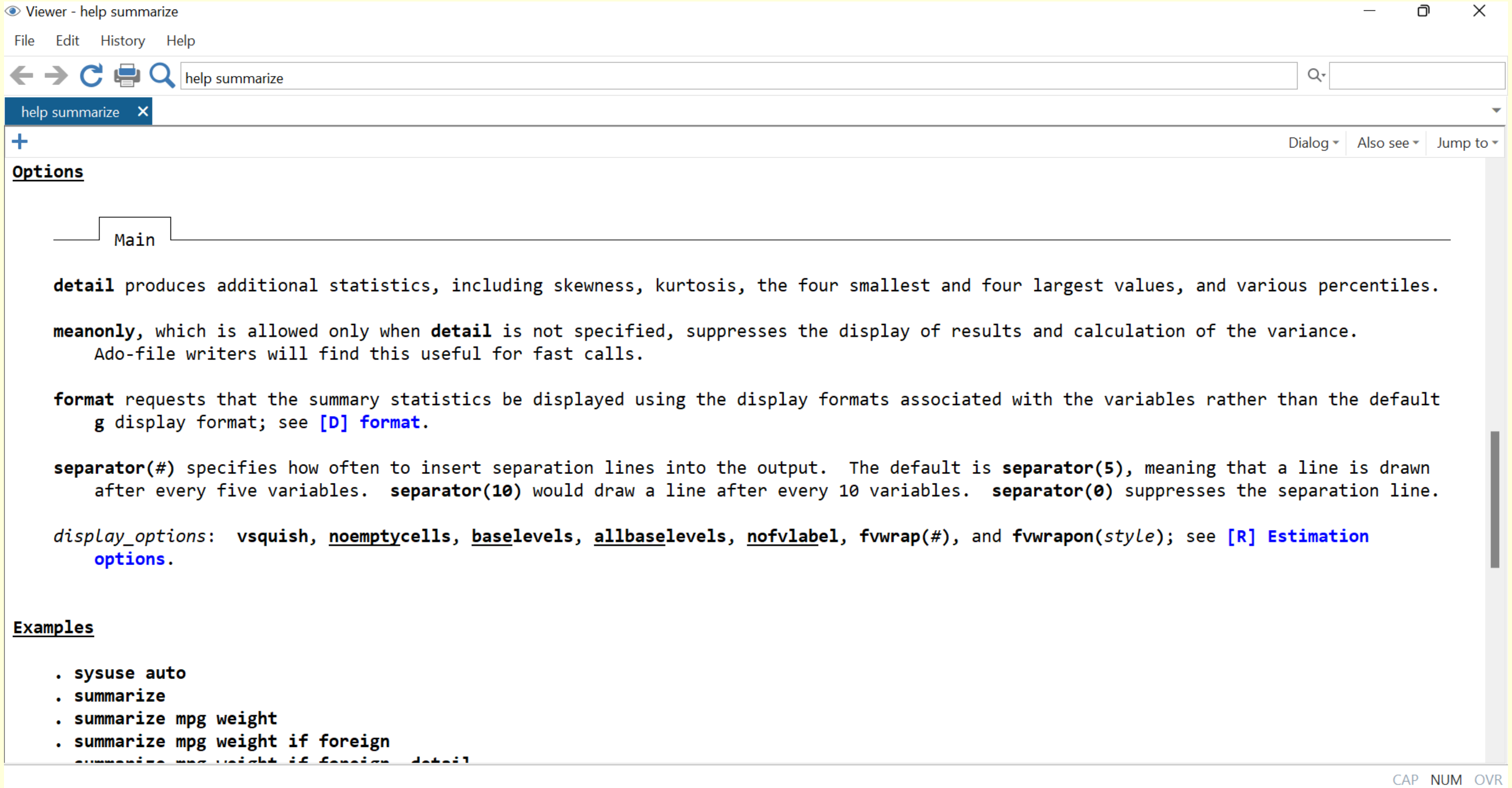
Stata Help Files Are Your Friend



```
[R] summarize — Summary statistics
                (View complete PDF manual entry)


Syntax

        summarize [varlist] [if] [in] [weight] [, options]

    options            Description
    ------------------------------------------------------------------------
    Main
      detail           display additional statistics
      meanonly         suppress the display; calculate only the mean; programmer's option
      format           use variable's display format
      separator(#)     draw separator line after every # variables; default is separator(5)
      display_options  control spacing, line width, and base and empty cells
    ------------------------------------------------------------------------

    varlist may contain factor variables; see fvvarlist.
    varlist may contain time-series operators; see tsvarlist.
    by, rolling, and statsby are allowed; see prefix.

    aweights, fweights, and iweights are allowed.  However, iweights may not be used with the detail option; see weight.
```

- Syntax section describes uses of command + how to specify them
- See also the link to the Stata manual (with extended help)

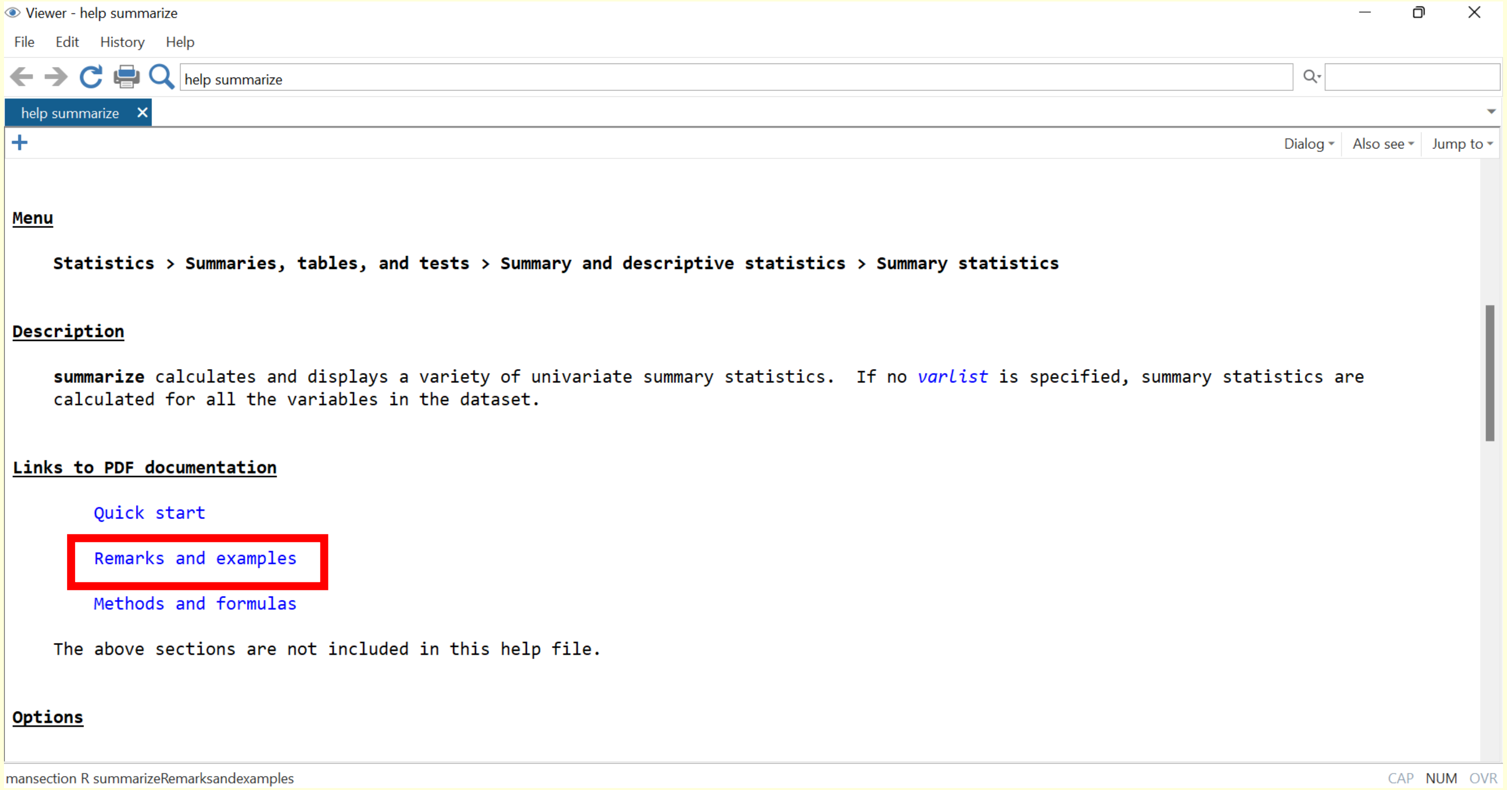
Stata Help Files Are Your Friend



```
Options

        Main

    detail produces additional statistics, including skewness, kurtosis, the four smallest and four largest values, and various percentiles.

    meanonly, which is allowed only when detail is not specified, suppresses the display of results and calculation of the variance.
        Ado-file writers will find this useful for fast calls.

    format requests that the summary statistics be displayed using the display formats associated with the variables rather than the default
        g display format; see [D] format.

    separator(#) specifies how often to insert separation lines into the output.  The default is separator(5), meaning that a line is drawn
        after every five variables.  separator(10) would draw a line after every 10 variables.  separator(0) suppresses the separation line.

    display_options:  vsquish, noemptycells, baselevels, allbaselevels, nofvlabel, fvwrap(#), and fvwrapon(style); see [R] Estimation
        options.


Examples

    . sysuse auto
    . summarize
    . summarize mpg weight
    . summarize mpg weight if foreign
    . summarize mpg weight if foreign, detail
```

- Options section gives details about the options of the command
- Examples section gives examples (extended ones in manual)

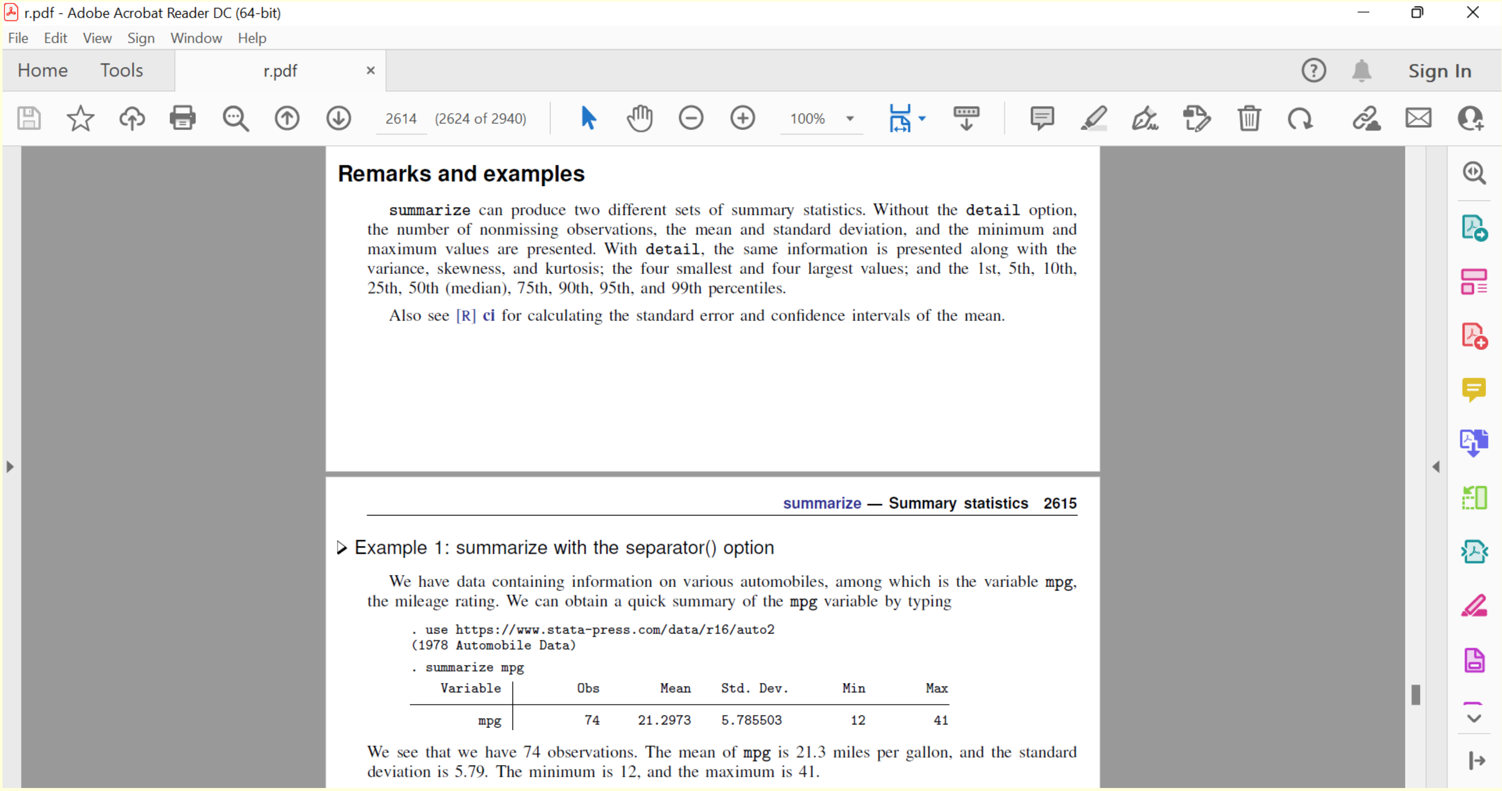
Stata Help Files Are Your Friend



```
Menu

    Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Summary statistics


Description

    summarize calculates and displays a variety of univariate summary statistics.  If no varlist is specified, summary statistics are
    calculated for all the variables in the dataset.


Links to PDF documentation

        Quick start

        Remarks and examples

        Methods and formulas

    The above sections are not included in this help file.


Options
```

- Description section provides extended description of command
- See also further links to Stata manual

Stata Help Files Are Your Friend


r.pdf - Adobe Acrobat Reader DC (64-bit)
File Edit View Sign Window Help
Home
Tools
r.pdf
Sign In
2614 (2624 of 2940)
100%

Remarks and examples

summarize can produce two different sets of summary statistics. Without the **detail** option, the number of nonmissing observations, the mean and standard deviation, and the minimum and maximum values are presented. With **detail**, the same information is presented along with the variance, skewness, and kurtosis; the four smallest and four largest values; and the 1st, 5th, 10th, 25th, 50th (median), 75th, 90th, 95th, and 99th percentiles.

Also see [R] **ci** for calculating the standard error and confidence intervals of the mean.

summarize — **Summary statistics** 2615

▷ Example 1: summarize with the separator() option

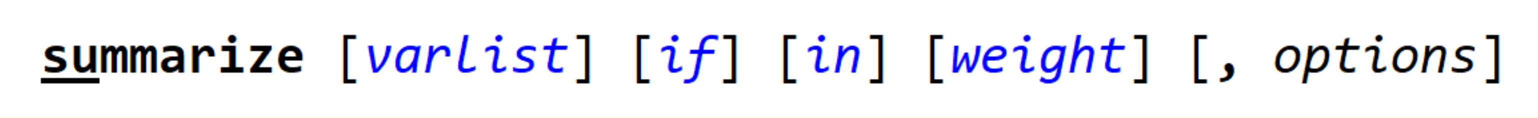
We have data containing information on various automobiles, among which is the variable **mpg**, the mileage rating. We can obtain a quick summary of the **mpg** variable by typing

```
. use https://www.stata-press.com/data/r16/auto2
(1978 Automobile Data)
. summarize mpg
    Variable         Obs        Mean    Std. Dev.       Min        Max

         mpg          74     21.2973    5.785503         12         41
```

We see that we have 74 observations. The mean of **mpg** is 21.3 miles per gallon, and the standard deviation is 5.79. The minimum is 12, and the maximum is 41.

- The Stata manual is even more helpful than the help files
- Other “friends”: Google and the Stata list (statlist.org)

General Stata Command Syntax

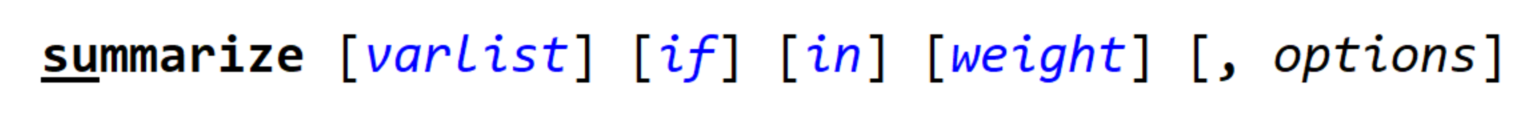


```
summarize [varlist] [if] [in] [weight] [, options]
```

Stata commands mirror everyday commands in their structure:

- They often start with a verb: “Bring me…”
- They then list an object: “… a pint of milk…”
- They may add a condition: “… if it is still before noon…”
- They may specify further details about what needs to be done: “… and quickly please” or “… but I only want semi-skimmed”

General Stata Command Syntax



```
summarize [varlist] [if] [in] [weight] [, options]
```

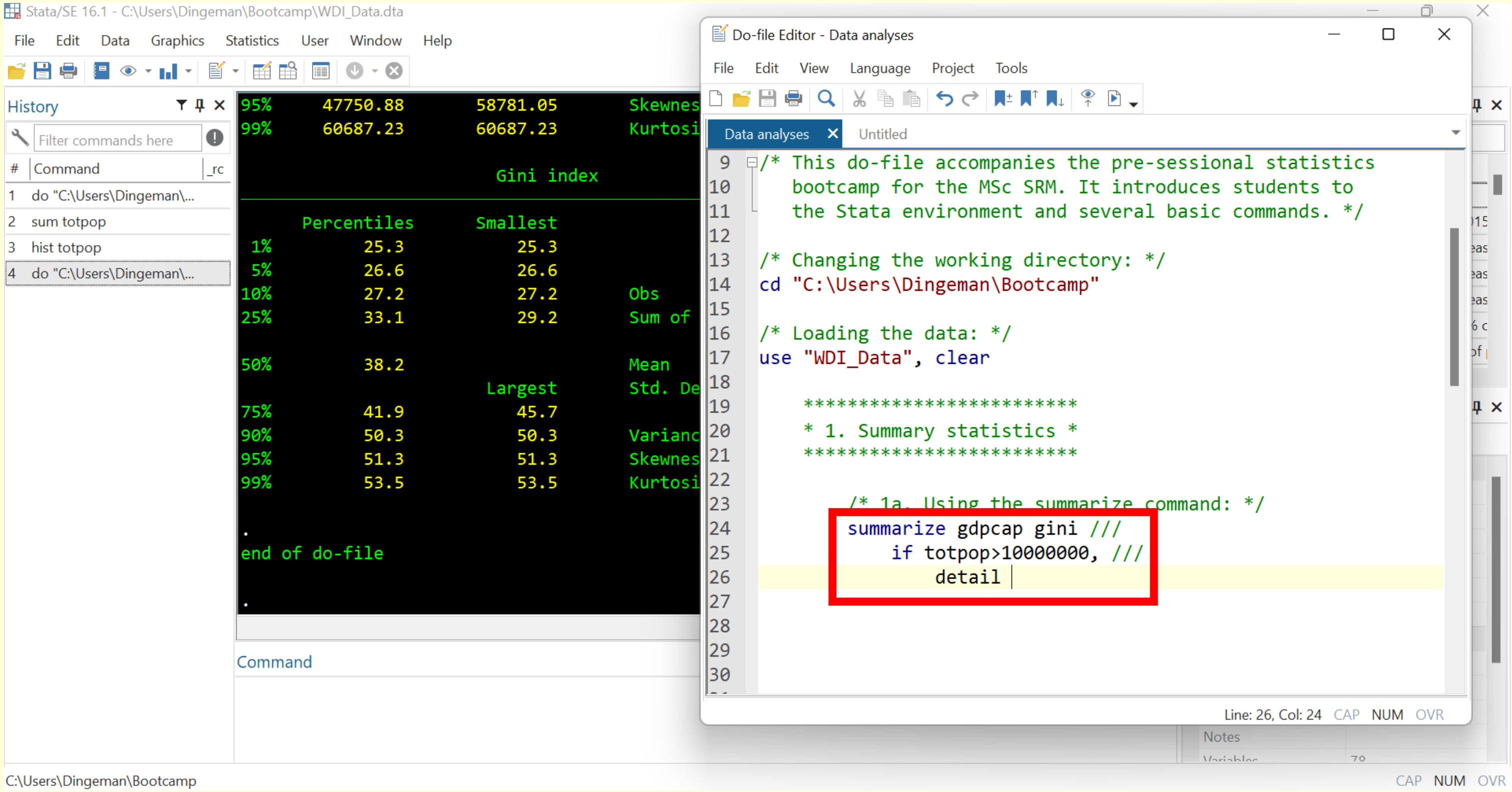
In nearly all cases, Stata syntax consists of four parts:

1. **Command**: What action do you want to see performed?
2. **Names of variables, files, objects**: On what objects is the command to be performed (“varlist”)
3. **Qualifier(s) on observations**: Which observations are to be taken into account (and how)? (“if”, “in”, “weight”)
4. **Options**: What special things should be done in the execution?

Part 1 is always there and Part 2 is most often also present

General Stata Command Syntax

```
summarize [varlist] [if] [in] [weight] [, options]
```



Stata/SE 16.1 - C:\Users\Dingeman\Bootcamp\WDI_Data.dta
File Edit Data Graphics Statistics User Window Help
History
Filter commands here
# Command _rc
Command
C:\Users\Dingeman\Bootcamp
CAP NUM OVR
Do-file Editor - Data analyses
File Edit View Language Project Tools
Data analyses
Untitled
Line: 26, Col: 24 CAP NUM OVR
Notes
Variables

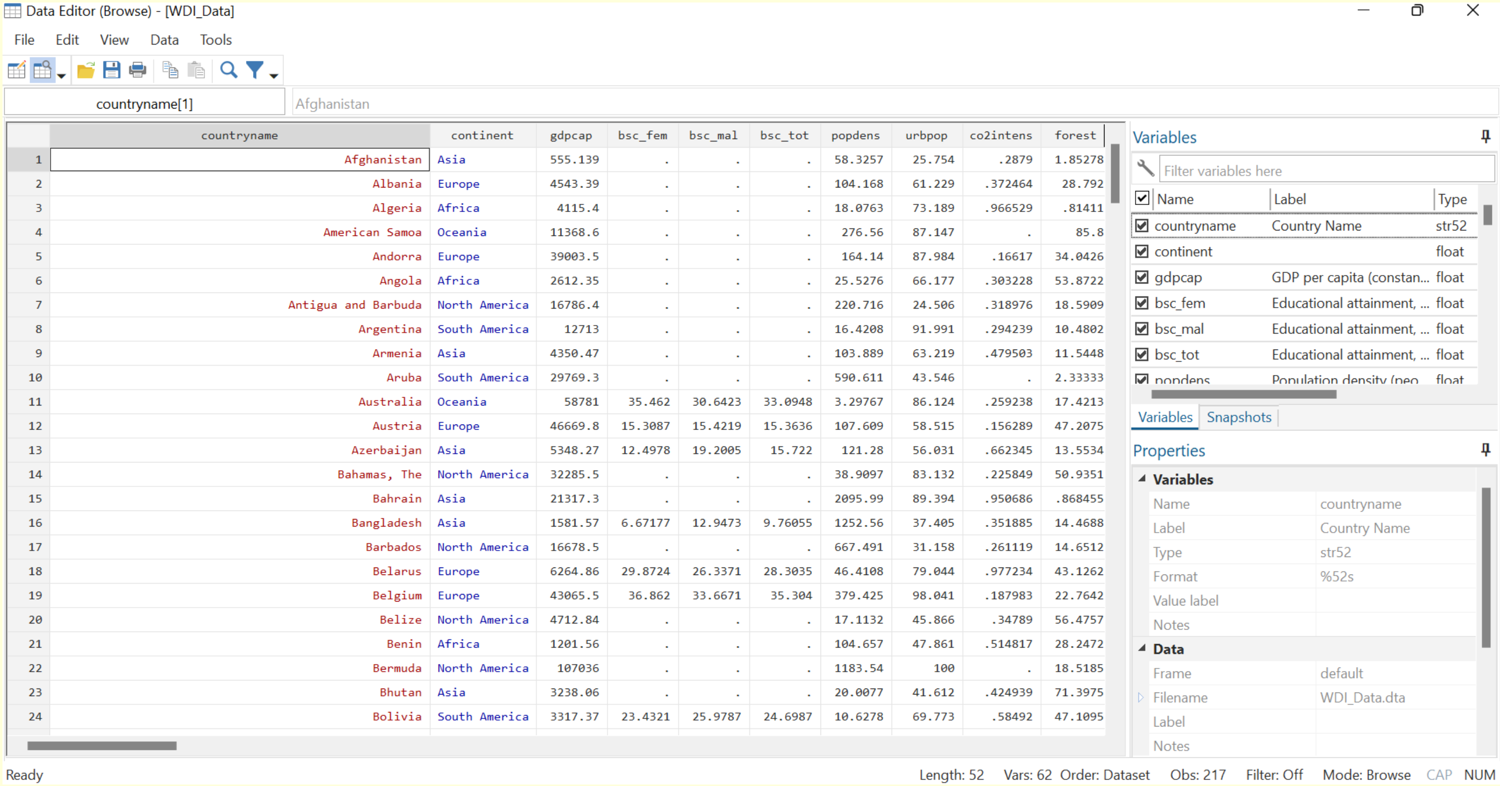
```
1  do "C:\Users\Dingeman\...
2  sum totpop
3  hist totpop
4  do "C:\Users\Dingeman\...
```

```
95%      47750.88       58781.05       Skewnes
99%      60687.23       60687.23       Kurtosi

                         Gini index

      Percentiles      Smallest
 1%         25.3           25.3
 5%         26.6           26.6
10%         27.2           27.2       Obs
25%         33.1           29.2       Sum of

50%         38.2                      Mean
                        Largest       Std. De
75%         41.9           45.7
90%         50.3           50.3       Varianc
95%         51.3           51.3       Skewnes
99%         53.5           53.5       Kurtosi

.
end of do-file

.
```

```
/* This do-file accompanies the pre-sessional statistics
   bootcamp for the MSc SRM. It introduces students to
   the Stata environment and several basic commands. */

/* Changing the working directory: */
cd "C:\Users\Dingeman\Bootcamp"

/* Loading the data: */
use "WDI_Data", clear

    ************************
    * 1. Summary statistics *
    ************************

        /* 1a. Using the summarize command: */
        summarize gdpcap gini ///
            if totpop>10000000, ///
                detail
```

Viewing Your Dataset

To view all your data, simply run the command `browse`:



countryname[1] Afghanistan

	countryname	continent	gdpcap	bsc_fem	bsc_mal	bsc_tot	popdens	urbpop	co2intens	forest
1	Afghanistan	Asia	555.139	.	.	.	58.3257	25.754	.2879	1.85278
2	Albania	Europe	4543.39	.	.	.	104.168	61.229	.372464	28.792
3	Algeria	Africa	4115.4	.	.	.	18.0763	73.189	.966529	.81411
4	American Samoa	Oceania	11368.6	.	.	.	276.56	87.147	.	85.8
5	Andorra	Europe	39003.5	.	.	.	164.14	87.984	.16617	34.0426
6	Angola	Africa	2612.35	.	.	.	25.5276	66.177	.303228	53.8722
7	Antigua and Barbuda	North America	16786.4	.	.	.	220.716	24.506	.318976	18.5909
8	Argentina	South America	12713	.	.	.	16.4208	91.991	.294239	10.4802
9	Armenia	Asia	4350.47	.	.	.	103.889	63.219	.479503	11.5448
10	Aruba	South America	29769.3	.	.	.	590.611	43.546	.	2.33333
11	Australia	Oceania	58781	35.462	30.6423	33.0948	3.29767	86.124	.259238	17.4213
12	Austria	Europe	46669.8	15.3087	15.4219	15.3636	107.609	58.515	.156289	47.2075
13	Azerbaijan	Asia	5348.27	12.4978	19.2005	15.722	121.28	56.031	.662345	13.5534
14	Bahamas, The	North America	32285.5	.	.	.	38.9097	83.132	.225849	50.9351
15	Bahrain	Asia	21317.3	.	.	.	2095.99	89.394	.950686	.868455
16	Bangladesh	Asia	1581.57	6.67177	12.9473	9.76055	1252.56	37.405	.351885	14.4688
17	Barbados	North America	16678.5	.	.	.	667.491	31.158	.261119	14.6512
18	Belarus	Europe	6264.86	29.8724	26.3371	28.3035	46.4108	79.044	.977234	43.1262
19	Belgium	Europe	43065.5	36.862	33.6671	35.304	379.425	98.041	.187983	22.7642
20	Belize	North America	4712.84	.	.	.	17.1132	45.866	.34789	56.4757
21	Benin	Africa	1201.56	.	.	.	104.657	47.861	.514817	28.2472
22	Bermuda	North America	107036	.	.	.	1183.54	100	.	18.5185
23	Bhutan	Asia	3238.06	.	.	.	20.0077	41.612	.424939	71.3975
24	Bolivia	South America	3317.37	23.4321	25.9787	24.6987	10.6278	69.773	.58492	47.1095

Variables

Filter variables here

☑ Name	Label	Type
☑ countryname	Country Name	str52
☑ continent		float
☑ gdpcap	GDP per capita (constan...	float
☑ bsc_fem	Educational attainment, ...	float
☑ bsc_mal	Educational attainment, ...	float
☑ bsc_tot	Educational attainment, ...	float
☑ popdens	Population density (peo...	float

Variables Snapshots

Properties

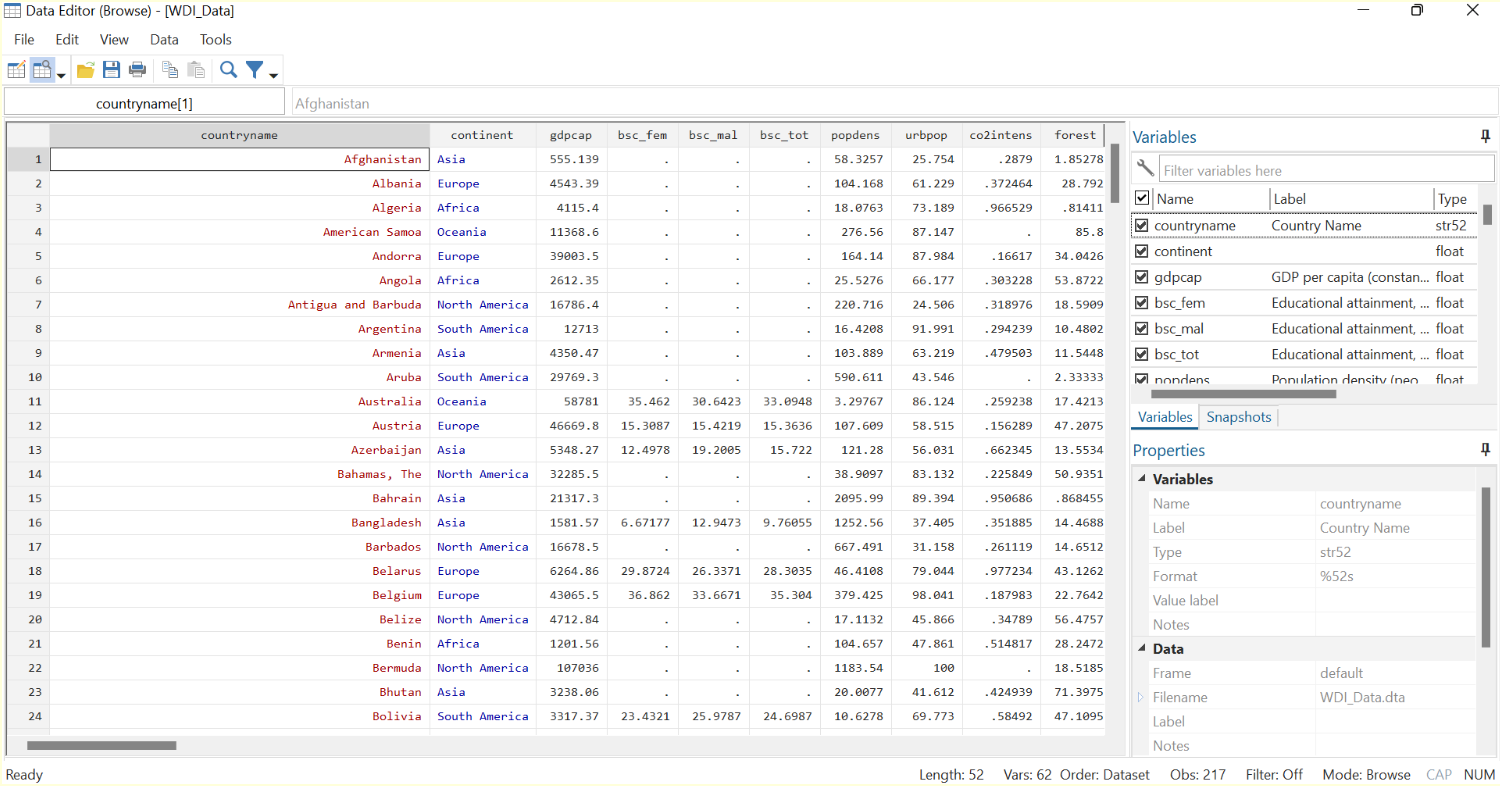
◢ Variables	
Name	countryname
Label	Country Name
Type	str52
Format	%52s
Value label	
Notes	
◢ Data	
Frame	default
▷ Filename	WDI_Data.dta
Label	
Notes	

Ready Length: 52 Vars: 62 Order: Dataset Obs: 217 Filter: Off Mode: Browse CAP NUM

- ➢ Black columns are **numeric variables**, red columns **string variables**, blue columns numeric variables with string labels

Viewing Your Dataset

To view all your data, simply run the command `browse`:



countryname[1] Afghanistan

	countryname	continent	gdpcap	bsc_fem	bsc_mal	bsc_tot	popdens	urbpop	co2intens	forest
1	Afghanistan	Asia	555.139	.	.	.	58.3257	25.754	.2879	1.85278
2	Albania	Europe	4543.39	.	.	.	104.168	61.229	.372464	28.792
3	Algeria	Africa	4115.4	.	.	.	18.0763	73.189	.966529	.81411
4	American Samoa	Oceania	11368.6	.	.	.	276.56	87.147	.	85.8
5	Andorra	Europe	39003.5	.	.	.	164.14	87.984	.16617	34.0426
6	Angola	Africa	2612.35	.	.	.	25.5276	66.177	.303228	53.8722
7	Antigua and Barbuda	North America	16786.4	.	.	.	220.716	24.506	.318976	18.5909
8	Argentina	South America	12713	.	.	.	16.4208	91.991	.294239	10.4802
9	Armenia	Asia	4350.47	.	.	.	103.889	63.219	.479503	11.5448
10	Aruba	South America	29769.3	.	.	.	590.611	43.546	.	2.33333
11	Australia	Oceania	58781	35.462	30.6423	33.0948	3.29767	86.124	.259238	17.4213
12	Austria	Europe	46669.8	15.3087	15.4219	15.3636	107.609	58.515	.156289	47.2075
13	Azerbaijan	Asia	5348.27	12.4978	19.2005	15.722	121.28	56.031	.662345	13.5534
14	Bahamas, The	North America	32285.5	.	.	.	38.9097	83.132	.225849	50.9351
15	Bahrain	Asia	21317.3	.	.	.	2095.99	89.394	.950686	.868455
16	Bangladesh	Asia	1581.57	6.67177	12.9473	9.76055	1252.56	37.405	.351885	14.4688
17	Barbados	North America	16678.5	.	.	.	667.491	31.158	.261119	14.6512
18	Belarus	Europe	6264.86	29.8724	26.3371	28.3035	46.4108	79.044	.977234	43.1262
19	Belgium	Europe	43065.5	36.862	33.6671	35.304	379.425	98.041	.187983	22.7642
20	Belize	North America	4712.84	.	.	.	17.1132	45.866	.34789	56.4757
21	Benin	Africa	1201.56	.	.	.	104.657	47.861	.514817	28.2472
22	Bermuda	North America	107036	.	.	.	1183.54	100	.	18.5185
23	Bhutan	Asia	3238.06	.	.	.	20.0077	41.612	.424939	71.3975
24	Bolivia	South America	3317.37	23.4321	25.9787	24.6987	10.6278	69.773	.58492	47.1095

Variables

Filter variables here

☑ Name	Label	Type
☑ countryname	Country Name	str52
☑ continent		float
☑ gdpcap	GDP per capita (constan...	float
☑ bsc_fem	Educational attainment, ...	float
☑ bsc_mal	Educational attainment, ...	float
☑ bsc_tot	Educational attainment, ...	float
☑ popdens	Population density (peo...	float

Properties

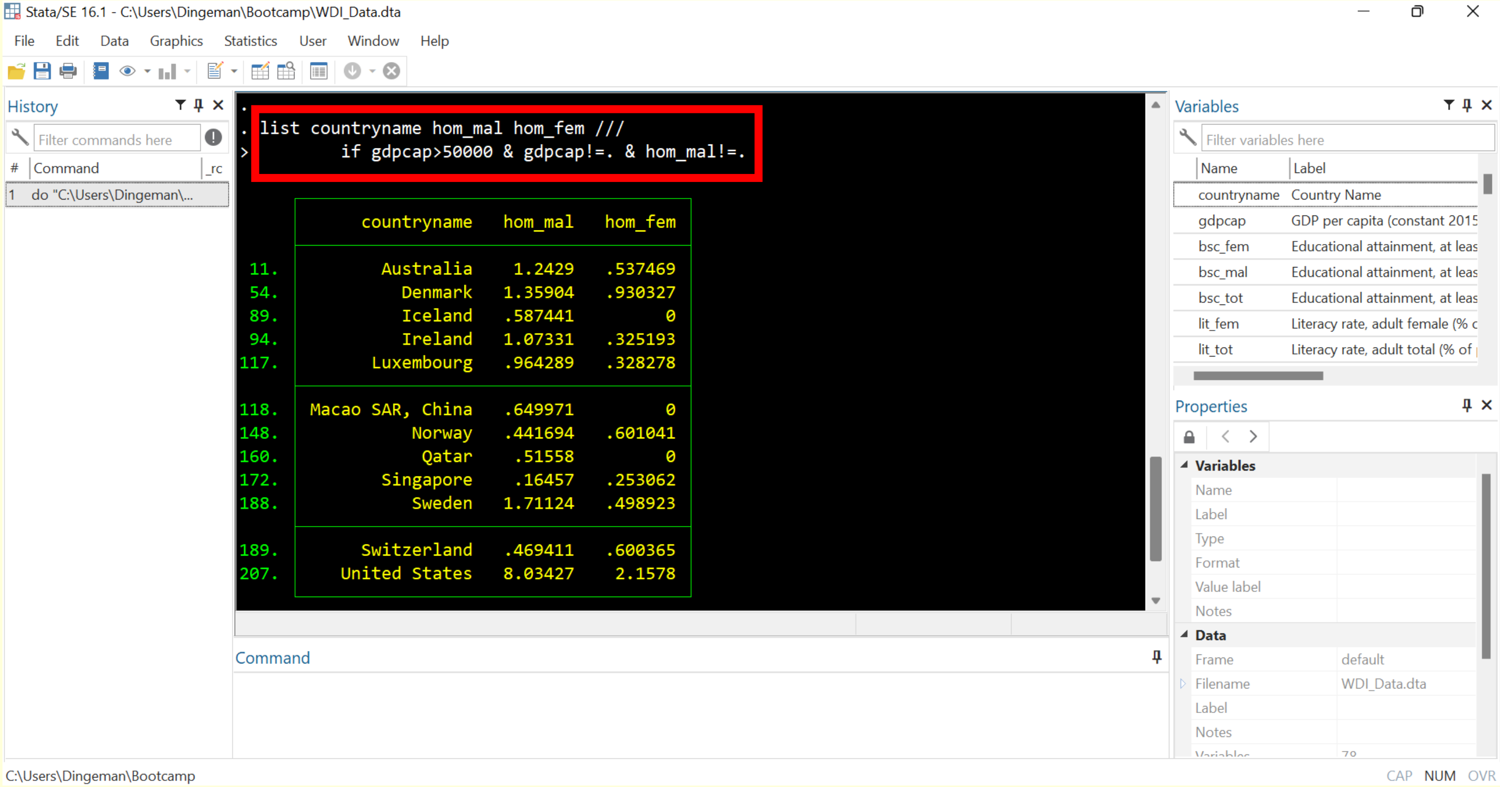
Variables	
Name	countryname
Label	Country Name
Type	str52
Format	%52s
Value label	
Notes	
Data	
Frame	default
Filename	WDI_Data.dta
Label	
Notes	

Ready Length: 52 Vars: 62 Order: Dataset Obs: 217 Filter: Off Mode: Browse CAP NUM

- For numeric variables, missing values are recorded as "."
- For string variables, missing values are recorded as " "

Viewing Your Dataset

You can also view (subsets of) your data using the `list` command:



History:

1 do "C:\Users\Dingeman\...

```
. 
. list countryname hom_mal hom_fem ///
>         if gdpcap>50000 & gdpcap!=. & hom_mal!=.

             countryname    hom_mal    hom_fem

 11.           Australia     1.2429    .537469
 54.             Denmark    1.35904    .930327
 89.             Iceland    .587441          0
 94.             Ireland    1.07331    .325193
117.          Luxembourg    .964289    .328278

118.    Macao SAR, China    .649971          0
148.              Norway    .441694    .601041
160.               Qatar     .51558          0
172.           Singapore     .16457    .253062
188.              Sweden    1.71124    .498923

189.         Switzerland    .469411    .600365
207.       United States    8.03427     2.1578
```

Name	Label
countryname	Country Name
gdpcap	GDP per capita (constant 2015
bsc_fem	Educational attainment, at leas
bsc_mal	Educational attainment, at leas
bsc_tot	Educational attainment, at leas
lit_fem	Literacy rate, adult female (% c
lit_tot	Literacy rate, adult total (% of

Variables	
Name	
Label	
Type	
Format	
Value label	
Notes	
Data	
Frame	default
Filename	WDI_Data.dta
Label	
Notes	

- In addition to or instead of “if”, you can also use “in”: e.g., `list … in 1/20` selects the first 20 cases of the dataset

Stata’s Logical and Relational Operators

In “if” statements, you often need the following operators:

Operator	Meaning
==	is equal to
!=	is not equal to
>	greater than
<	smaller than
>=	greater or equal to
<=	smaller or equal to
&	and
|	or

Over to You Again…

Task: Pick a few variables (perhaps first run `describe`) and describe their distributions:

- For numerical descriptions, use `summarize`
- For visual descriptions, use `histogram`

Further instructions:

- Run `help` *`commandname`* if you run into problems
- Try to find the median of the variables
- Try to change the look of the histograms
- Consider low vs. high income countries separately
- Consider Europe and another continent separately

Random Sampling and Sampling Distribution

- Sampling theory plays a huge role in specifying the assumptions upon which our statistical inferences rely

- We draw inferences from the **sample** about the **population**

- Most statistical methods rely on **averages** or **weighted averages** of a sample of data

- The values of the variables in our sample constitute **random variables**, and so any average we take of those values will also constitute a random variable

What If We Don’t Have a Simple Random Sample?

- **Simple random sample**: every unit has an equal probability of being selected
  - I.e., selecting units from a population without replacement

- **Stratified random sample**: the population is divided into strata (e.g., based on sex, ethnicity, region) and a simple random sample is drawn from each stratum

- **Convenience sample**: chosen in a way convenient to the researcher, not at random from the population of interest
  - E.g., snowball sampling, or only selecting UCL students when one is really interested in the entire UK population

The Sampling Distribution of the Sample Mean

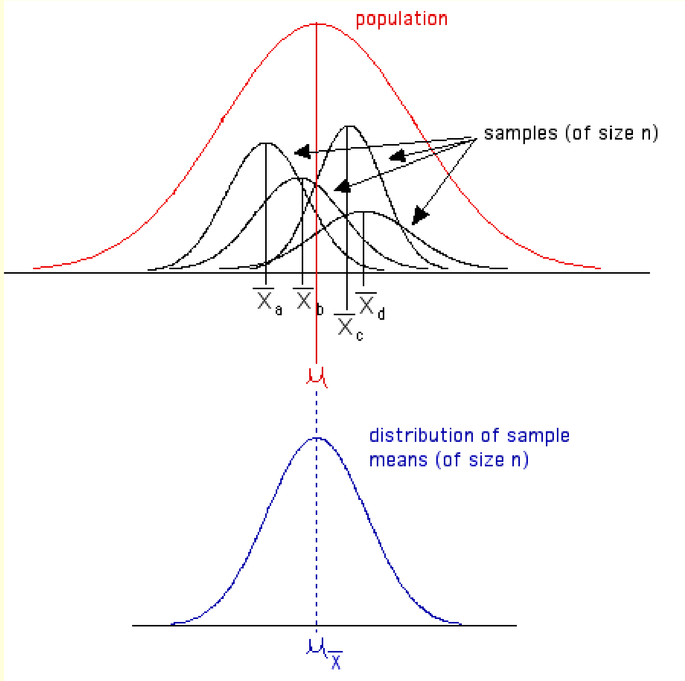
- In a **simple random sample**, n objects are drawn at random from a population and each object is equally likely to be drawn

- The sample mean is: $\bar{Y} = \frac{Y_1+Y_2+\cdots+Y_n}{n} = \frac{\sum_{i=1}^{N} Y_i}{n}$

- The value of $\bar{Y}$ will differ from one random sample to the next

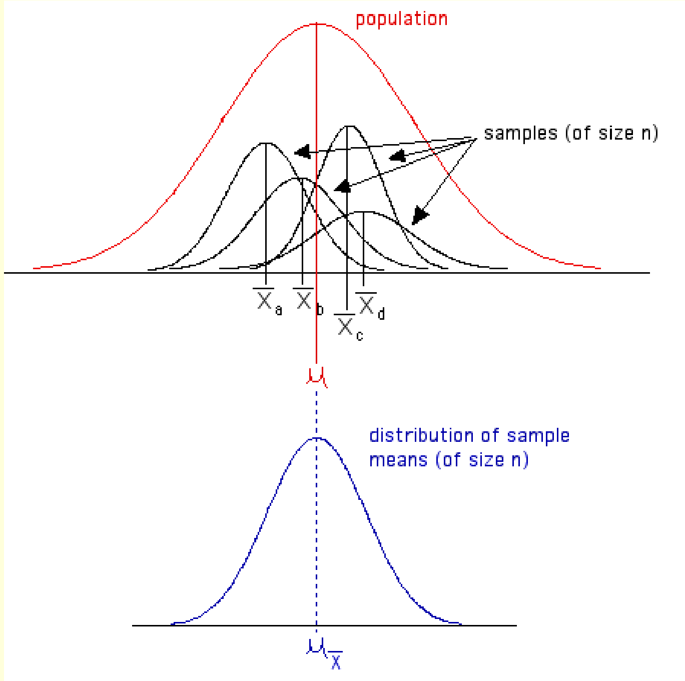
- As $\bar{Y}$ is random (recall a variable derived from a random variable is also a random variable itself), it has a probability distribution, which is called the **sampling distribution** of $\bar{Y}$
  - Example: we repeatedly draw a sample from this classroom to determine the average age of all students present

The Sampling Distribution of the Sample Mean



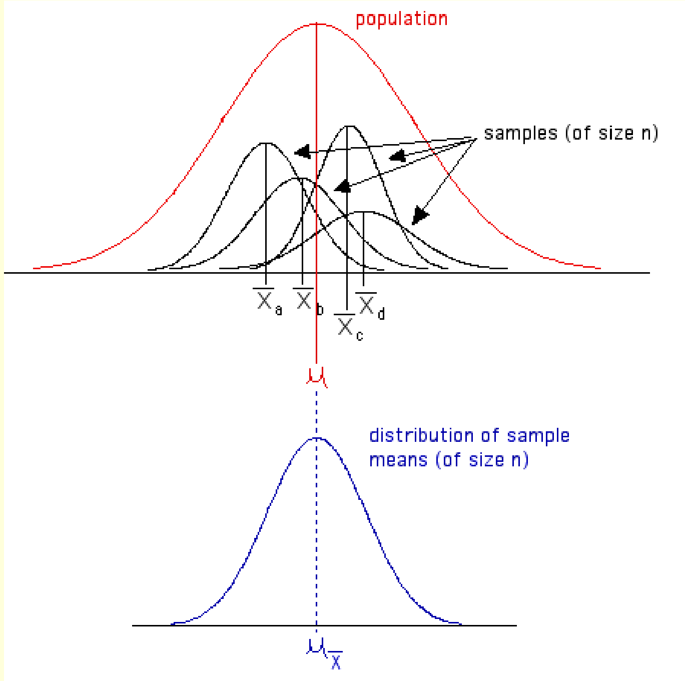
- The sampling distribution gives an idea how much the values of the estimate (sample mean) might vary from sample to sample
  - Example: How much does the average age vary across repeated samples that we draw from this classroom?

The Sampling Distribution of the Sample Mean



- The sampling distribution tells us something about the uncertainty in the estimate obtained from the one sample we actually observe
  - How predictive is the average age in our 1st classroom sample w.r.t. the underlying average age in the total classroom?

The Sampling Distribution of the Sample Mean



- There exist sampling distributions for any **estimand** (i.e., the thing that we try to estimate based on our sample)
  - E.g., also for a regression coefficient β that captures the linear association between student age and exam results

Hypothesis Testing

- A **theory** is a tentative conjecture about the causes of some phenomenon of interest; once a theory has been developed, we can restate it into one or more testable hypotheses
  - E.g., girls read more books than boys and thus have better reading skills
- A **hypothesis** is a theory-based statement about a relationship that we expect to observe
  - E.g., girls achieve higher scores than boys on reading tests
- Theories and hypotheses both pertain to the population
  - E.g., we expect girls to be better readers across <u>all students</u>

Hypothesis Testing

- A **hypothesis** is a theory-based statement about a relationship that we expect to observe
  - E.g., girls achieve higher scores than boys on reading tests

- For every hypothesis there is a corresponding **null hypothesis**
  - About what we would expect if our theory is incorrect
  - E.g., there is no association between Y and X in the population
  - In our example: girls are <u>not</u> better readers than boys

The Logic of Hypothesis Testing

- We have a coin we suspect may be biased in favour of heads
  - How would we decide whether it is indeed biased?

- We try tossing it 10 times (i.e., we draw a sample), and we find that it comes up heads 8 times

- We formulate our null hypothesis that the coin is fair, and ask:

  *What is the **probability** that we would see 8 or more heads in ten tosses if this coin is indeed fair?*

  - H_0 = coin is fair,
  - H_1 = coin is biased in favour of heads

The Logic of Hypothesis Testing

- We essentially ask: What is **P(8 or more heads|H_0)**?

- If this probability turns out very small, we will "reject the null"
  - I.e., we will conclude that the coin is not fair but biased

- Now, the probability of 8 heads in 10 tosses of a fair coin is:
  - B(8; 10; 0.5) = $(\frac{10!}{8!})(\frac{1}{2^{10}})$ = 0.044
  - Likewise, we can calculate B(9; 10; 0.5) and B(10; 10; 0.5)
  - Added together, **P(8 or more heads|H_0) ≈ 0.055**

The Logic of Hypothesis Testing

- So, P(8 or more heads|H_0) ≈ 0.055
  - We call this value the **p-value**: the probability of observing something at least as extreme as we do observe, if H_0 were true
- We **reject the null** when the p-value is small
  - In that case, there is a low probability that the claim made by the null is true in the population of interest
- How low should the p-value be?
  - This depends on the chosen level of statistical significance (**α**), which captures how tolerant we are of incorrectly rejecting H_0
  - In the social sciences, **α** typically equals 0.05, 0.01, or 0.001

Hypothesis Testing: Central Limit Theorem

- To calculate the p-value, we rely on information about the sampling distribution of the parameter of interest (e.g., the sample mean $\bar{x}$ or the regression coefficient b_x)

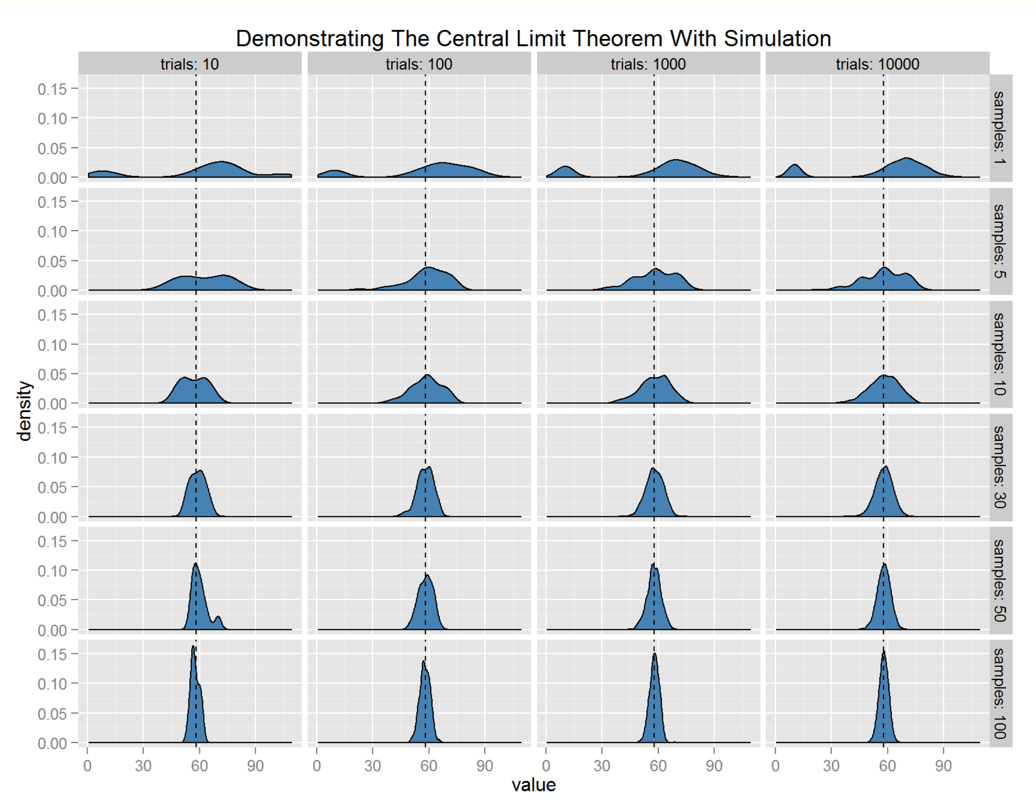
- In particular, we apply the **central limit theorem**:

If sufficiently large random samples are repeatedly drawn from a population, then the sampling distribution of the sample means will be approximately normally distributed

Sample sizes of N ≥ 30 are typically large enough

Hypothesis Testing: Central Limit Theorem

If sufficiently large random samples are repeatedly drawn from a population, then the sampling distribution of the sample means will be approximately normally distributed → An illustration:



Hypothesis Testing: Central Limit Theorem

- Implication central limit theorem:
    - If we take repeated samples from a population, the means from those samples will fall in a **normal distribution** around the population mean

- This principle is used in hypothesis testing to calculate the p-value:
    - Applying this principle, we can determine the probability of observing what we do observe in our sample if the null hypothesis about the underlying population was true

Hypothesis Testing: Type of Errors

In hypothesis testing, we can make two types of errors:

- Type-I error (**false positive**): rejecting H_0 when it is true
  - Covid test says you have got the virus even though you don’t
- Type-II error (**false negative**): not rejecting H_0 when it is false
  - Covid test says you don’t have the virus even though you do

- There is a trade-off between avoiding Type-I and Type-II errors

Hypothesis Testing: Type of Errors

In hypothesis testing, we can make two types of errors:

- Type-I error (**false positive**): rejecting H_0 when it is true
- Type-II error (**false negative**): not rejecting H_0 when it is false

- Hypothesis testing rules are constructed to make the probability of committing a Type-I error (false positive) fairly small
  - We define the significance level of a test (**α**) as the probability of committing a Type-I error
  - E.g., **α** = 0.05 implies that, with repeated sampling, we will incorrectly reject the null hypothesis 5% of the time
    - In the Covid test example: in 5% of cases, the test tells us that someone has Covid even though they don’t

Hypothesis Testing: Test Statistics

- A **test statistic** is a single number that quantifies the difference between the observed data in our sample, and the expected data under the null hypothesis
  - E.g., it compares a sample mean to the hypothesised population mean, or a sample regression coefficient to the assumed population regression coefficient

- Examples of test statistics: χ^2, F, T, Z
  - The test statistic that should be used in a given test depends on the type of analysis

T-test

For tests about means and OLS coefficients, we use **T-tests**:

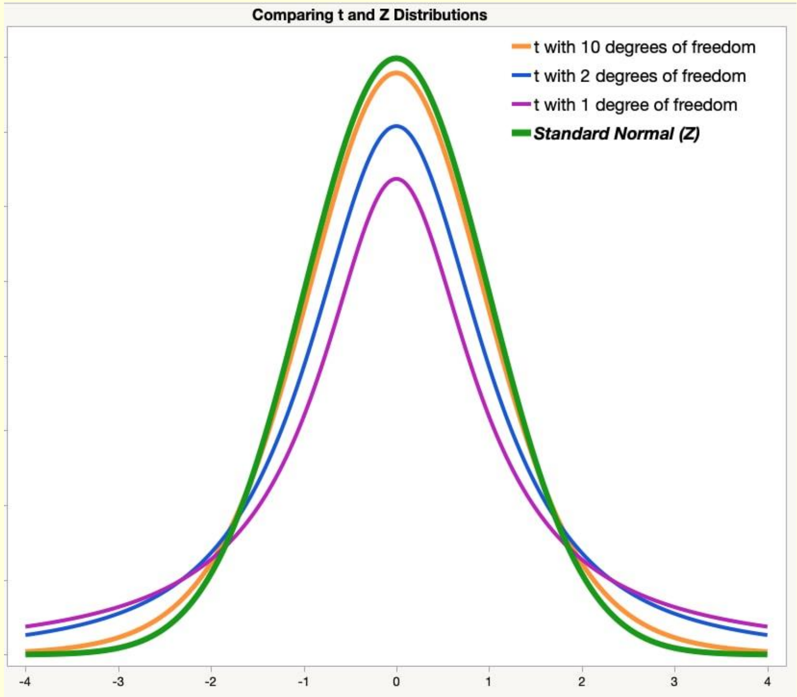
- Because we must estimate the standard deviation by using the standard error, we must use the t-distribution rather than the standard normal (Z) distribution
- T-distribution controls for the fact that the estimate of the standard deviation is going to be a bit off for smaller samples
- T-statistic is similar to Z-statistic, except it uses the standard error of the parameter of interest as denominator:

$$\boldsymbol{t = \frac{Sample\ mean - Population\ mean}{Sample\ standard\ error}}$$

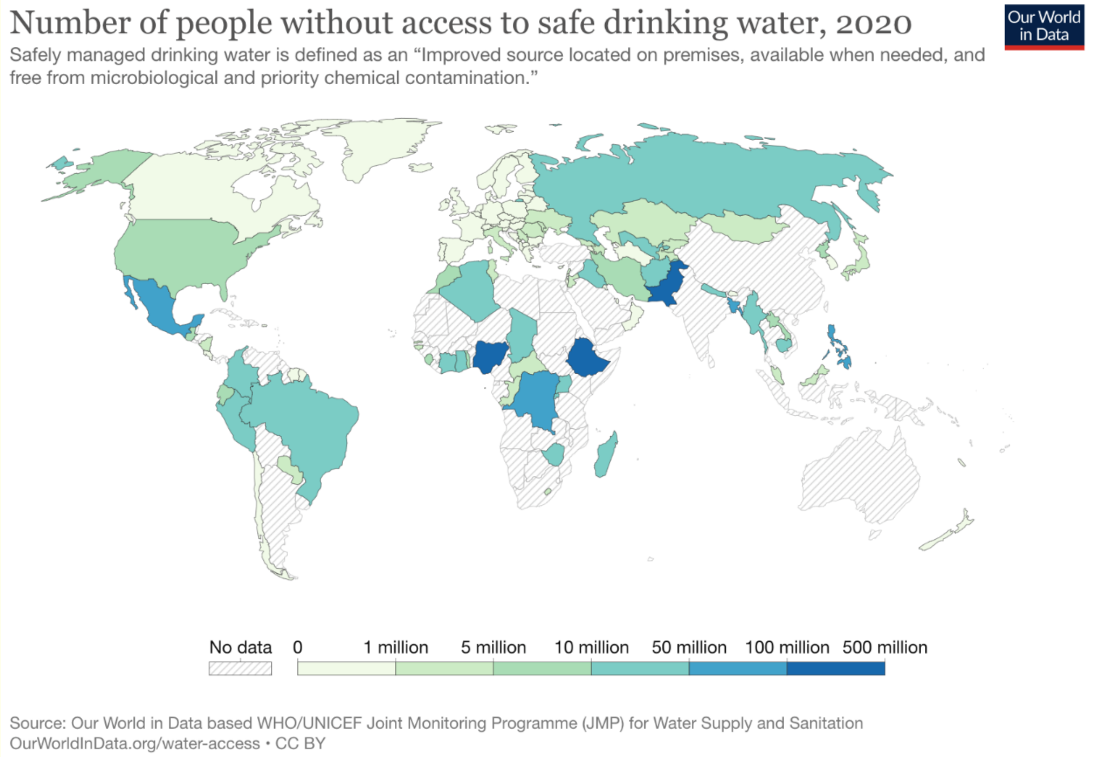
T-test

For tests about means and OLS coefficients, we use **T-tests**:

- T-distribution looks like a standard normal distribution but it has fatter tails (and degrees of freedom attached to it)
- T-distribution will be flatter in smaller samples, but it approximates the Z-distribution as sample size increases



Hypothesis Testing: Stata Example



Number of people without access to safe drinking water, 2020

Safely managed drinking water is defined as an "Improved source located on premises, available when needed, and free from microbiological and priority chemical contamination."

Source: Our World in Data based WHO/UNICEF Joint Monitoring Programme (JMP) for Water Supply and Sanitation
OurWorldInData.org/water-access • CC BY

Testable statement:

People who live in higher-income countries and lower-income countries **do not have equal access** to *safe drinking water*

Hypothesis Testing: Stata Example

Testable statement:

People who live in higher-income countries and lower-income countries **do not have equal access** to *safe drinking water*

This statement concerns a difference of population means

- If $\boldsymbol{\mu_1}$ = population mean higher-income countries, $\boldsymbol{\mu_2}$ = population mean lower-income countries, and $\Delta = \boldsymbol{\mu_1} - \boldsymbol{\mu_2}$:
  - Then the null hypothesis reads $\boldsymbol{H_0}: \Delta = \mathbf{0}$
  - Although the two-sided alternative hypothesis $\boldsymbol{H_a}: \Delta \neq \mathbf{0}$ is considered here, you can also test one-sided alternative hypotheses, i.e., $\boldsymbol{H_a}: \Delta < \mathbf{0}$ or $\boldsymbol{H_a}: \Delta > \mathbf{0}$

Hypothesis Testing: Stata Example

Before going ahead with our test, we first need to group countries into higher-income and lower-income ones

- We can create a variable for this, which equals 0 for lower-income countries and 1 for higher-income countries
- We use the `generate` and `replace` commands for this:

```
generate high_income = 0

replace high_income = 1 if gdpcap > 12695
replace high_income = . if gdpcap == .
```

- `generate` creates completely new variables, `replace` changes the values of existing variables

Hypothesis Testing: Stata Example

Before going ahead with our test, we first need to group countries into higher-income and lower-income ones

- We can create a variable for this, which equals 0 for lower-income countries and 1 for higher-income countries
- It is always important to check your data manipulations; here we can use `tabulate` and/or `summarize` for this:

```
tabulate gdpcap high_income
summarize gdpcap if high_income == 0
summarize gdpcap if high_income == 1
```

Hypothesis Testing: Stata Example

Before going ahead with our test, we first need to group countries into higher-income and lower-income ones

- We can create a variable for this, which equals 0 for lower-income countries and 1 for higher-income countries
- It is also smart to attach informative labels to new variables:

```
label variable high_income ///
     "Whether country is higher-income"

label define HIGH_INCOME ///
     0 "Lower-income country" ///
     1 "Higher-income country"

label values high_income HIGH_INCOME
```

Hypothesis Testing: Stata Example

Remember to check the Stata help files for more information on the set-up, syntax, and working of these commands, e.g.:

```
help generate
help tabulate
help label
```

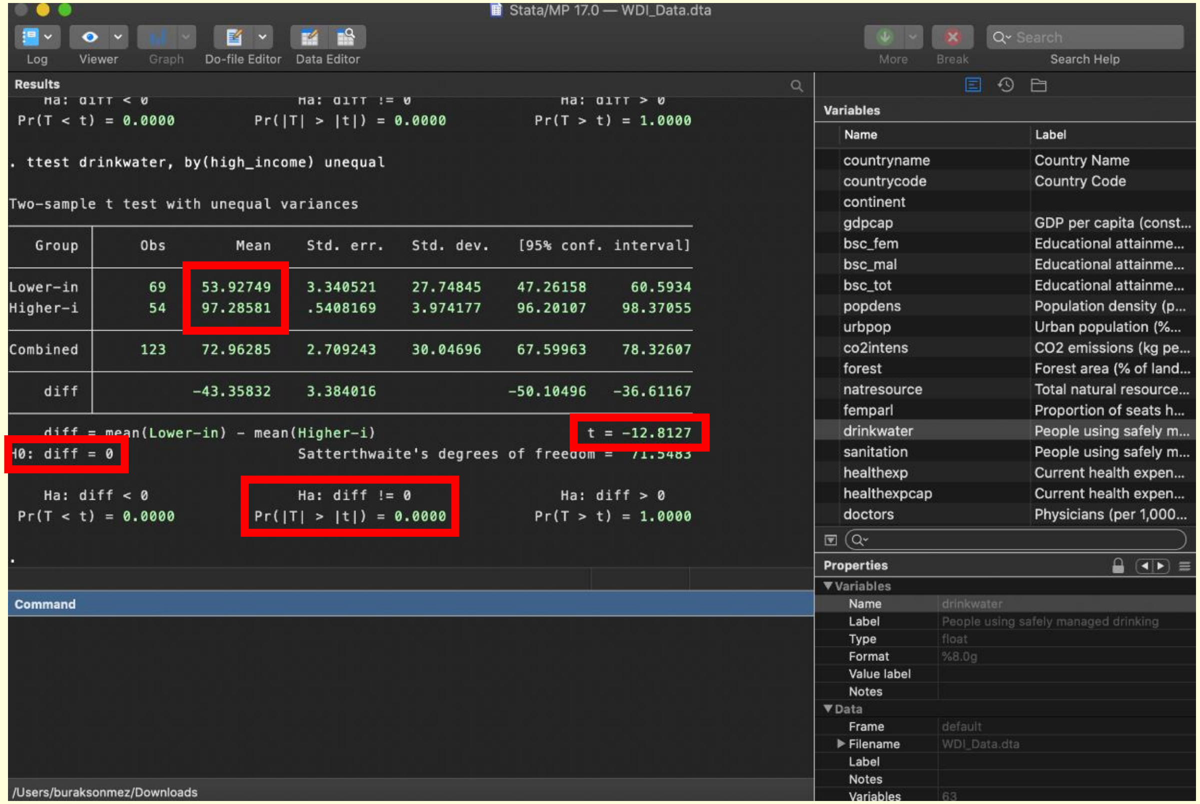
Bonus item:

Another way to construct the same grouping variable based on income level is with the `recode` command. Have a look at the help file for `recode` and see whether you can reproduce the “high_income” variable using `recode`.

Hypothesis Testing: Stata Example

We can now conduct the t-test by running:

- `ttest drinkwater, by(high_income) unequal`
- See `help ttest` for details about this command



```
    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000

. ttest drinkwater, by(high_income) unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
Lower-in |      69    53.92749    3.340521    27.74845    47.26158     60.5934
Higher-i |      54    97.28581    .5408169    3.974177    96.20107    98.37055
---------+--------------------------------------------------------------------
Combined |     123    72.96285    2.709243    30.04696    67.59963    78.32607
---------+--------------------------------------------------------------------
    diff |           -43.35832    3.384016               -50.10496   -36.61167
------------------------------------------------------------------------------
    diff = mean(Lower-in) - mean(Higher-i)                        t = -12.8127
H0: diff = 0                     Satterthwaite's degrees of freedom =  71.5483

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000

.
```

/Users/buraksonmez/Downloads

Variables

Name	Label
countryname	Country Name
countrycode	Country Code
continent	
gdpcap	GDP per capita (const...
bsc_fem	Educational attainme...
bsc_mal	Educational attainme...
bsc_tot	Educational attainme...
popdens	Population density (p...
urbpop	Urban population (%...
co2intens	CO2 emissions (kg pe...
forest	Forest area (% of land...
natresource	Total natural resource...
fempari	Proportion of seats h...
drinkwater	People using safely m...
sanitation	People using safely m...
healthexp	Current health expen...
healthexpcap	Current health expen...
doctors	Physicians (per 1,000...

Properties

▼Variables	
Name	drinkwater
Label	People using safely managed drinking
Type	float
Format	%8.0g
Value label	
Notes	
▼Data	
Frame	default
▶ Filename	WDI_Data.dta
Label	
Notes	
Variables	63

Hypothesis Testing: Stata Example

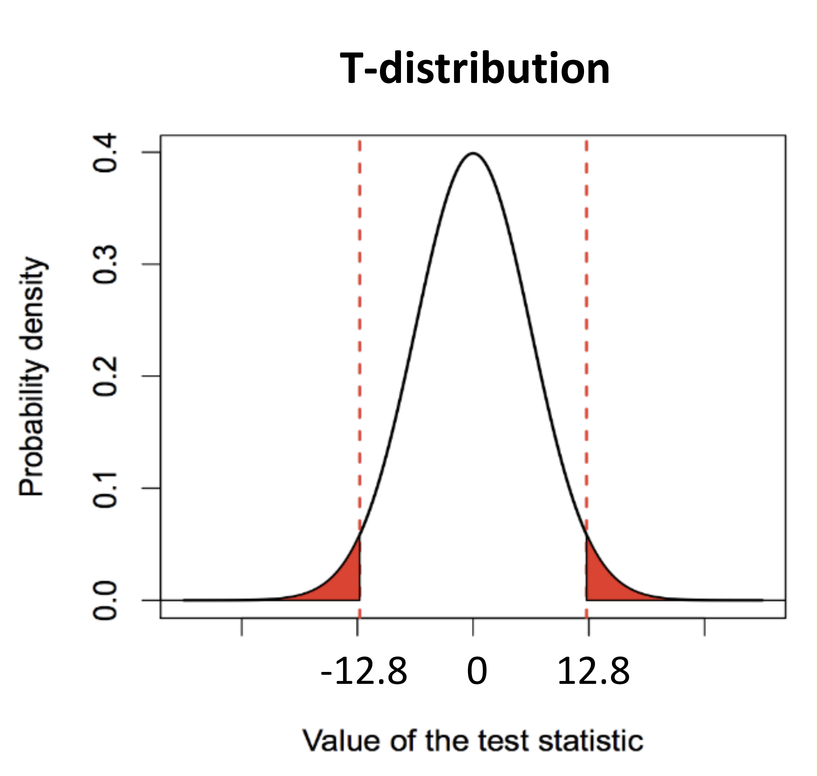
Key questions to answer:

- If in the population there was no association between a country's income level and access to safe drinking water, what are the chances of observing the group means that we do observe in our sample data?
- In other words, what is the probability of observing a group mean difference of at least 43 percentage points if in reality there is no difference between these two groups?
- The **t-value** for the observed group mean difference is -12.8, and the **p-value** (i.e., the probability of observing something at least as extreme by pure chance if H_0 is true) is <0.0001

Hypothesis Testing: Stata Example

A graphical illustration of what we are testing:

- What is the size of the red area in this graph?



Hypothesis Testing: Stata Example

Conclusions that we can draw from this test:

- Given that the p-value is smaller than **α** (whether we set **α** at 0.05, 0.01, or 0.001), we can **reject the null** that there is no association b/w a country’s income level and access to safe drinking water
- It is extremely unlikely that we observe what we do in our sample data if in reality there was no difference in access to safe drinking water between higher-income and lower-income countries
- In fact, if we set **α** at 0.05, the critical t-value equals 1.96
  - This means that a t-value larger than 1.96 or smaller than -1.96 would already have been enough to reject the null hypothesis

Uncertainty – Confidence Intervals

- We may also wish to know the range within which the population parameter of interest is likely to fall

- We can refer to **confidence intervals** for this, which quantify the uncertainty embedded in our inferences
  - We can identify confidence intervals at different levels (e.g., 90%, 95%, 99% confidence intervals)

- In general, confidence intervals are formulated as follows:

$$(estimate\ \ - k * SE\ , estimate + k\ \ * SE)$$

  - SE is the standard error of the estimate, and the value for "k" depends on the chosen confidence level (e.g., 1.96 for 95%)

Uncertainty – Confidence Intervals

- In general, confidence intervals are formulated as follows:

$$(estimate\ - k * SE\ , estimate + k\ \ * SE)$$

  - SE is the standard error of the estimate, and the value for “k” depends on the chosen confidence level (e.g., 1.96 for 95%)

- The width of a confidence interval depends on the standard error of the estimator, which partially depends on the sample size
  - Larger sample size ⟶ smaller standard error
  - Smaller standard error ⟶ narrower confidence interval
  - Narrower confidence interval = more precise estimate

Uncertainty – Confidence Intervals

- In our Stata example, the 95% confidence interval for the difference in access to safe drinking water for lower-income vs. higher-income countries is: (-50.1 ; 36.3)

- What a 95% confidence interval tells us:
  - With repeated sampling, 95% of the confidence intervals calculated based on these samples will contain the population parameter of interest

- What a 95% confidence interval does not tell us:
  - The population parameter of interest has a 95% probability of lying within the confidence interval that we have calculated based on this particular sample

Time for a Task…

Task: Conduct a similar t-test to compare two group means as in the example above, based on the WDI dataset

- See `help ttest` if you are unsure about the command
- As grouping variable, you can again use countries' level of income, but for example also their continent, or a newly created grouping based on any other variable
- Prior to conducting your test, form a hypothesis
- Reflect on your results with a peer, paying attention to the observed difference in this dataset, and the accompanying p-value, t-value, and 95% confidence interval

If you have any spare time, conduct a second test

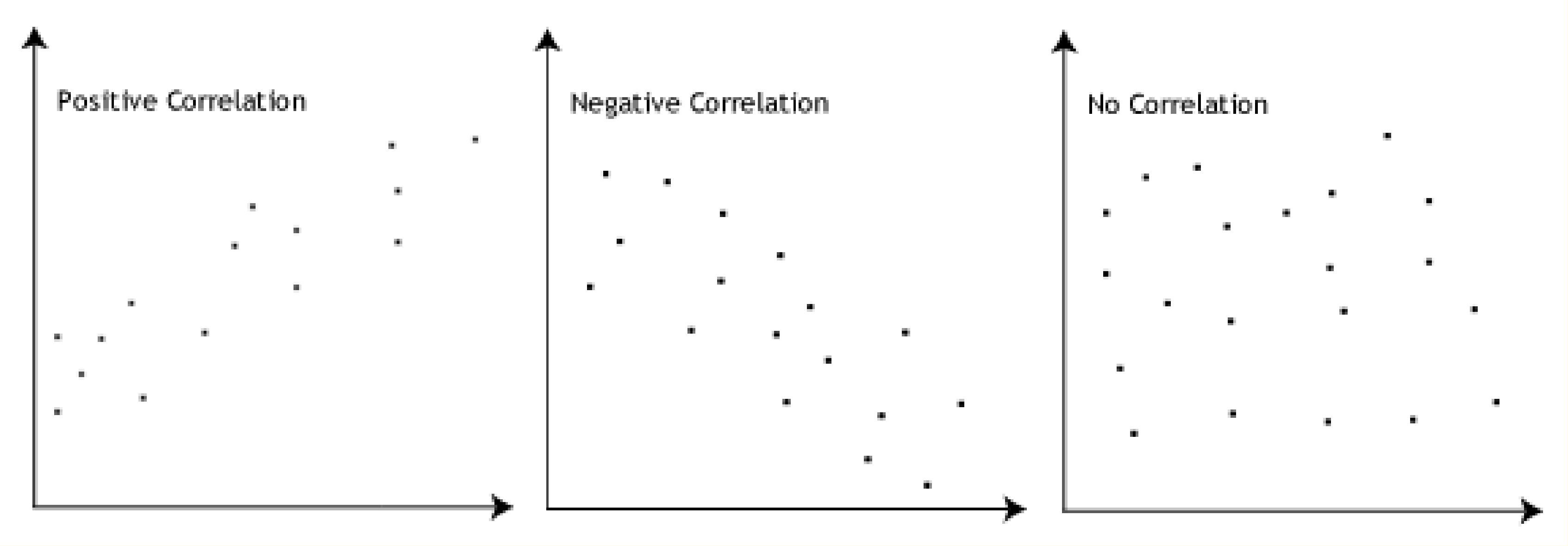
Understanding Associations

- In everyday language, the terms dependence, association and correlation are used interchangeably
  - However, statistically, association is synonymous with dependence but different from correlation

- **Association** is a general relationship: one variable provides information about another
  - Two variables are associated if there is a pattern in a scatter plot that is too strong to arise simply by chance

- **Correlation** measures a specific form of association: two variables are correlated when there is a linear trend

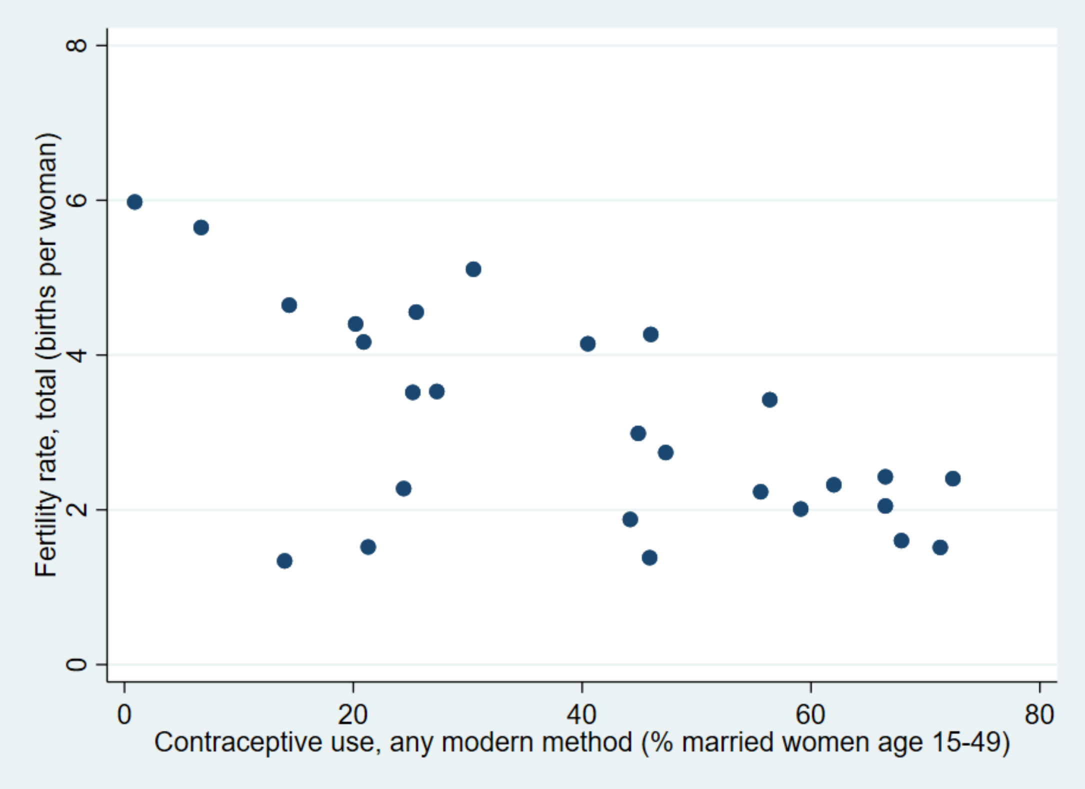
Understanding Associations



Understanding Associations

To create a scatter plot, use the `scatter` command

- The first variable listed will be plotted along the vertical axis
- The second variable will be plotted along the horizontal axis
- E.g., `scatter tfr contracep`

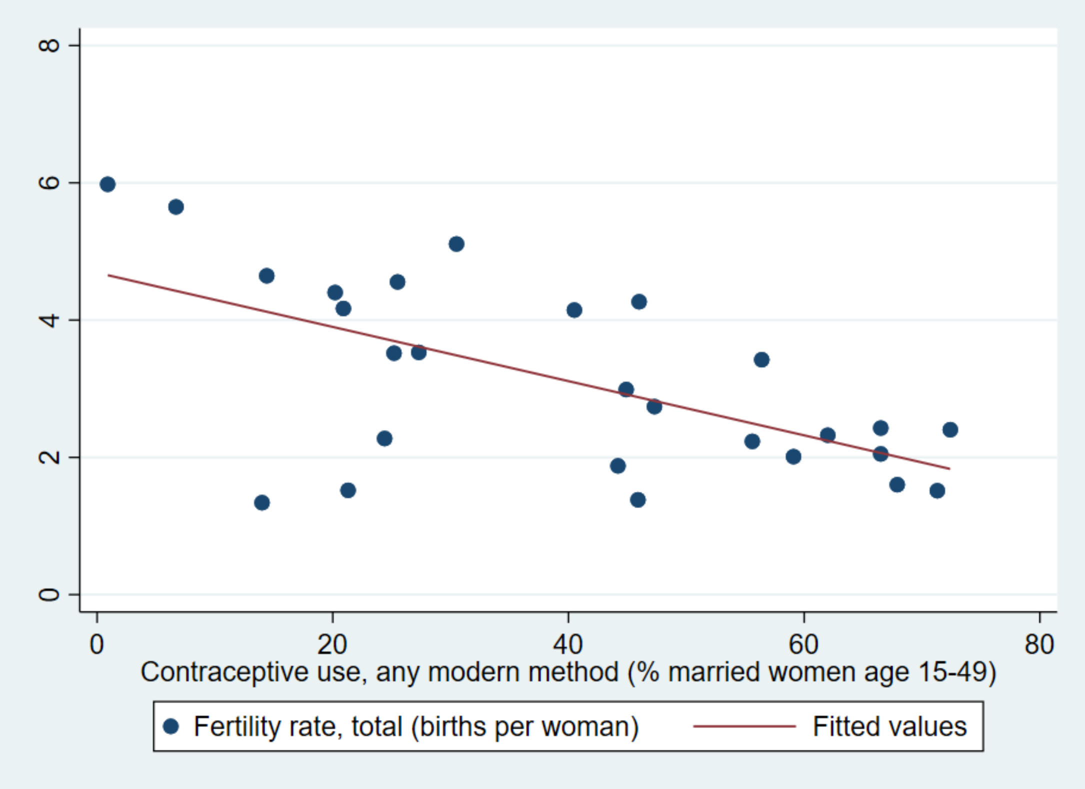


Understanding Associations

To create a scatter plot, use the `scatter` command

- We can add another plot after specifying `||`
- E.g., we can add a best-fitting line using `lfit`:

```
scatter tfr contracept || lfit tfr contracept
```



Covariance

- **Covariance** refers to the idea that the pattern of variation for one variable corresponds to the pattern of variation for another variable: the two variables "vary together"

- Statistically speaking, covariance is the multiplication of the deviations from the mean for the first variables and the deviations from the mean for the second variable:

$$cov(X,Y) = \frac{\sum_{i}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

- The covariance tells us the <u>direction</u> of an association: + or –
  - It does not tell us about the <u>strength</u> of the association (reference to underlying distributions of variables is missing)

Pearson’s Correlation

For continuous data, we can calculate **Pearson's correlation (r)**

- r measures strength & direction of association for linear trends
- r rescales the covariance to the underlying distributions of the variables involved; in formula format: $\mathbf{r} = \frac{\boldsymbol{cov(X,Y)}}{\boldsymbol{\sigma_X \sigma_Y}}$

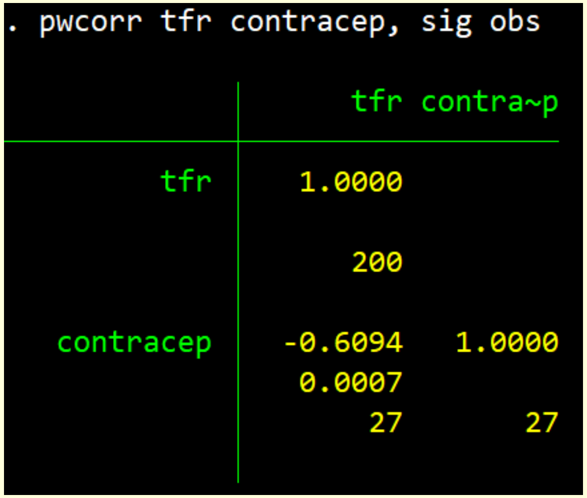
Dividing the covariance by the product of the standard deviations normalizes the covariance to a range from -1 to +1

- -1 = perfectly negative correlation (all points on decreasing line)
- +1 = perfectly positive correlation (all points on increasing line)
- 0 = no correlation (random cloud of points)
- Correlation is weaker closer to 0 and stronger closer to +/-1

Pearson’s Correlation

In Stata, we can use the pwcorr command to get Pearson’s r:

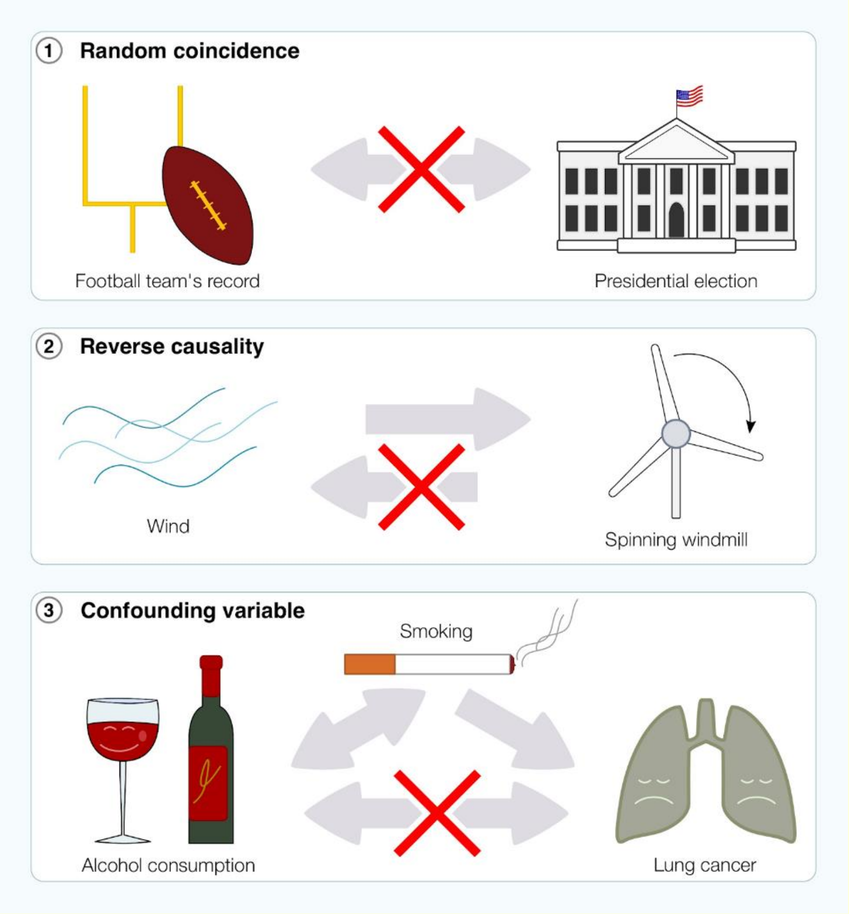
- E.g., `pwcorr tfr contracep`
- `pwcorr` stands for “pairwise correlation”
- We can add the options `obs` and `sig` to get the underlying number of observations and the p-value of the correlation
- You can also obtain correlation coefficients for more than 2 variables add a time; you just extend the variable list for this



```
. pwcorr tfr contracep, sig obs

             |      tfr contra~p
-------------+------------------
         tfr |   1.0000
             |
             |      200
             |
   contracep |  -0.6094   1.0000
             |   0.0007
             |       27       27
             |
```

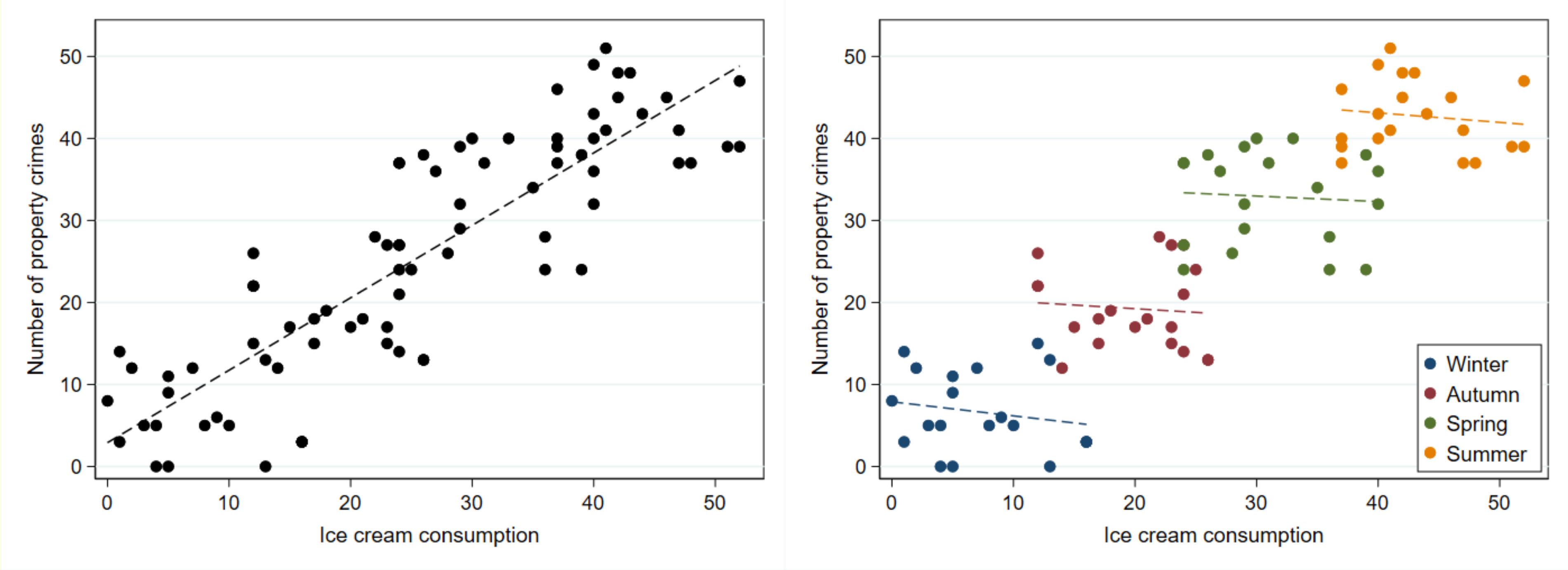
Spurious Correlations

Not all correlations reflect causal relationships:



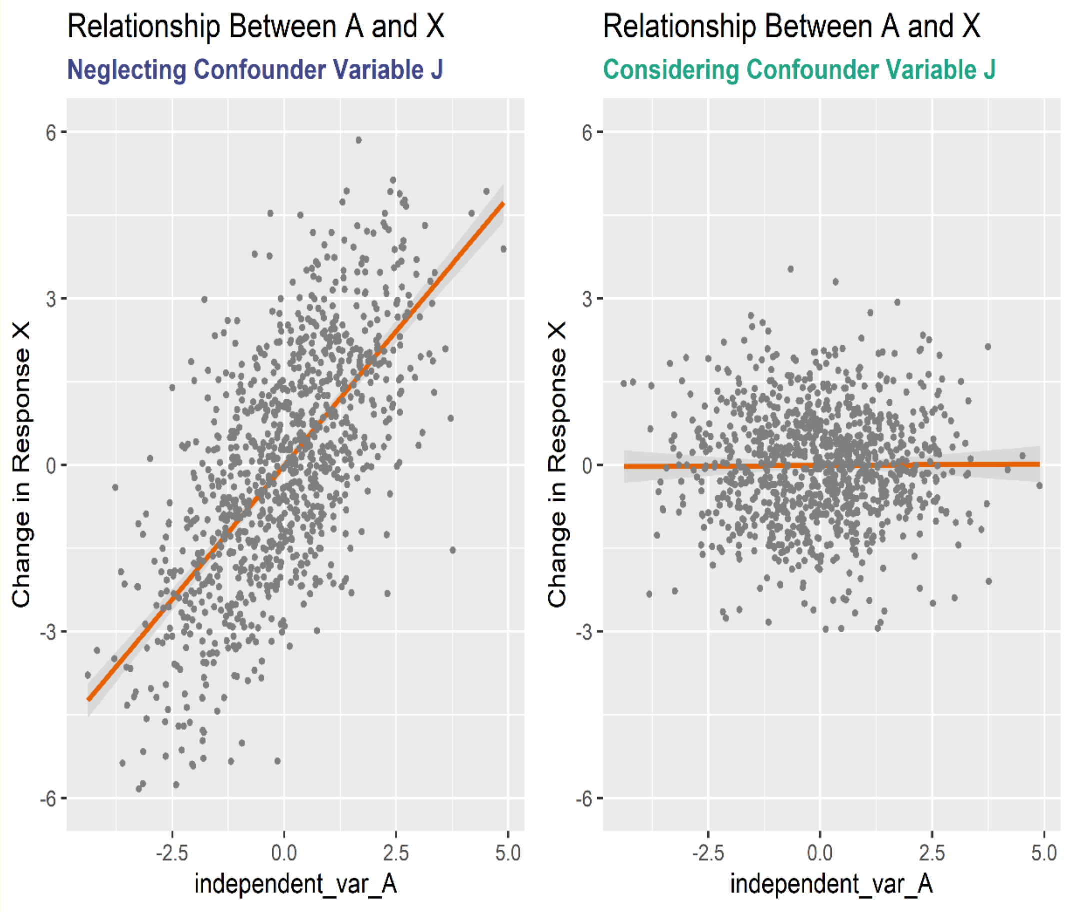
Spurious Correlations

Not all correlations reflect causal relationships:



Spurious Correlations

Not all correlations reflect causal relationships:



Final Exercise

Task: Please go back to the WDI dataset and try to identify a potential spurious correlation

- Report the correlation coefficient
- Draw a scatter plot of the association
- Explain why you think the relationship is not causal
- Investigate how the association varies across different types of countries
  - You can apply "if" conditions and create new variables to group countries together

Questions

- Please ask any remaining questions now or feel free to email us at b.sonmez@ucl.ac.uk or d.wiertz@ucl.ac.uk
- Feedback on this bootcamp is also very welcome
- Good luck with your MSc at UCL!