Introduction to Data Science

An Introduction to Statistical Concepts for Data Analysis

Professor Vernon Gayle

vernon.gayle@ed.ac.uk
@Profbigvern
github.com/vernongayle

AQMEN

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

THE UNIVERSITY of EDINBURGH

The Idealized Data Analysis Process

1. Planning stage
2. Data wrangling
3. Exploratory data analysis
4. Advanced data analysis (e.g. predictive model)
5. Documenting results (e.g. reports & visualisations)
6. Dissemination of results (e.g. presentation)
7. Recognition (e.g. promotion, Nobel Prize)

Over the years I have observed that there is probably more chance me winning a stage of the Tour de France and marrying a former model, than there is of one of my research projects moving unproblematically through these steps!

The Data Analysis Process in Reality

The Structure of the Workshop

- No philosophical discussion of the limitations of statistical methods
- No discussion of the limitations of data sources
- No discussion of data quality
- General data examples
- Plenty of anecdotes (stop me if there are too many)

My aims

- Convey some of my enthusiasm for the topic
- Engage (and possibly even entertain) participants
- Alleviate anxiety a little
- Encourage people to ask questions
- Leaving with a bit more knowledge
- Possibly motivate people to do more in future

The speed of presentations - tell me if it is too fast or too slow!!!

Part 1 Basic Concepts

Variables and Cases (probably revision)

- Variables
  - measures of concepts

- Cases
  - Distinctive entities
  - *People, firms, farms, hospitals, schools, local authorities, regions, nation states, animals, police stations/divisions, prisons*

The Variable by Case Matrix

ID	GENDER	AGE	DEGREE
001	0	21	0
002	1	22	1
003	1	25	1

Cases are distinctive entities

- People, firms, farms, hospitals, schools, local authorities, regions, nation states, animals, police stations/divisions, prisons

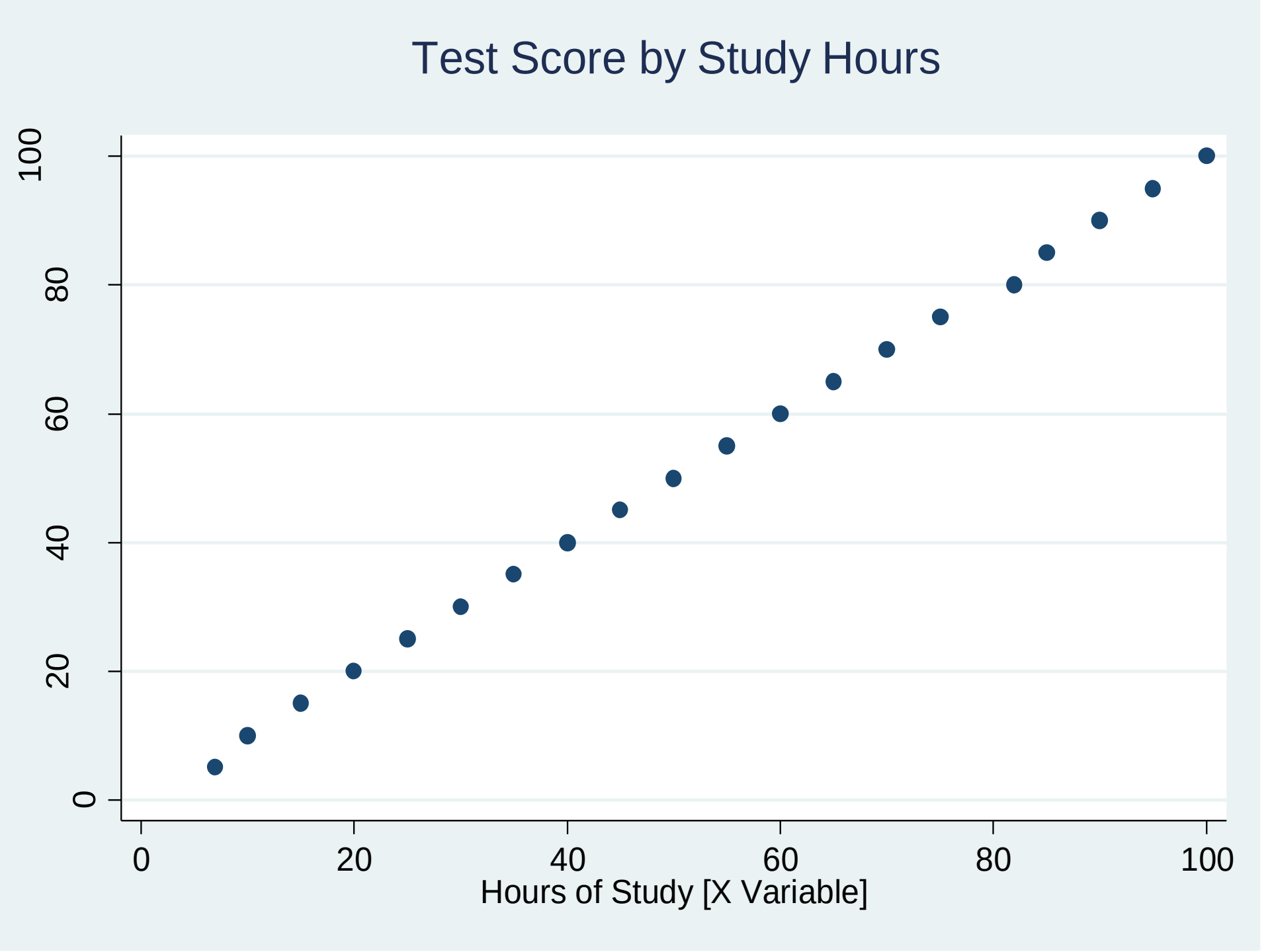
Variables

- Outcome variables
  - Y variables
    - Educational test score
    - Life expectancy (years)
    - Number of criminal convictions
    - Numerous health outcomes
    - Subjective wellbeing (SWB) measures

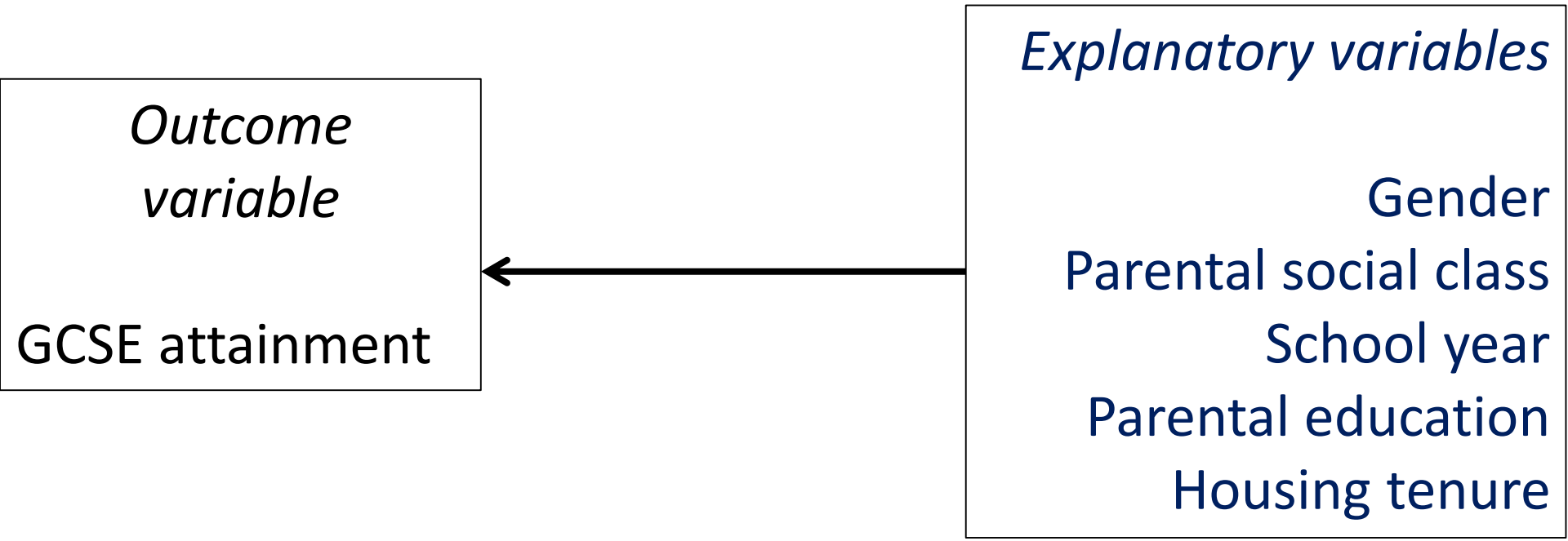
Variables

- Explanatory variables
  - X variables
  - These variables explain outcome variables

    - Hours of study
    - Gender
    - Ethnicity
    - Socioeconomic classifications
    - Age
    - Housing tenure (type)

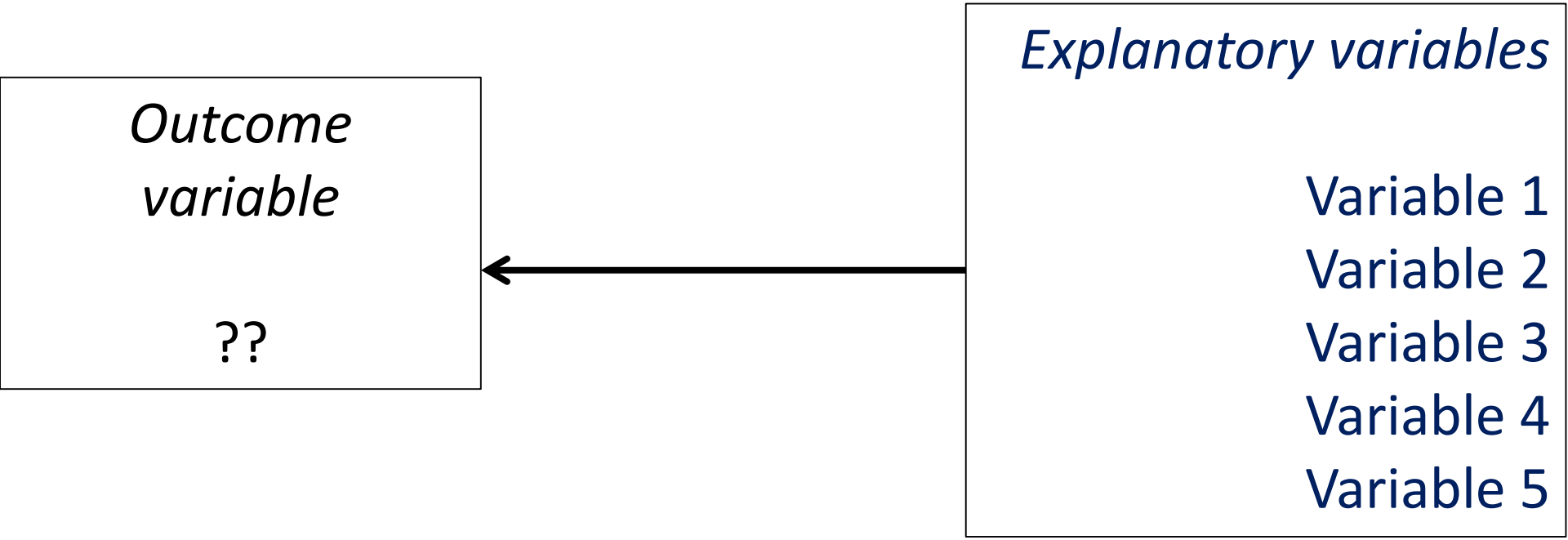

Test Score by Study Hours
100
80
60
40
20
0
0
20
40
60
80
100
Hours of Study [X Variable]

Example of variables in a study



Connelly, Murray and Gayle (2013) *Sociological Research Online*

Think of variables that might be related to your area…



- Population N
  - UK (Decennial) Census
  - All the police officers in Scotland

- Sample n
  - 10% of all of the police officers in Scotland

- Census (whole population)
- Social Survey (usually a sample)
- Administrative source might cover all or part of population

- **Univariate** – a single variable

- **Bivariate** – two variables
  - One outcome variable (Y) and one explanatory variable (X)

- **Multivariate** – three (or many more) variables
  - One outcome variable (Y) and many explanatory variables (X)
  - This is the ‘cheddar’ (see Urban Dictionary)
  - We are all about the multivariate bass not the bivariate treble (Meghan Trainor)

There are more advanced multivariate analyses which have multiple outcomes!

The approaches are beyond a two day course

Register for a part-time MSc or PhD with me and learn all about these approaches

Part 2 Levels of Measurement (types of variables)

Categorical variables

Gender

Male [ ]

Female [ ]

Professional Grade

Professor [ ]

Reader [ ]

Senior Lecturer [ ]

Lecturer [ ]

Continuous variables

Age in years

Number of years of service

Annual salary in pounds (£)

Number of days sick in the last three years

Levels of Measurement (types of variables)

This is presented in much more confusing terms in most statistics (and psychology research methods) text books however!

You can read more in any standard psychology statistics and research methods textbook or

http://psc.dss.ucdavis.edu/sommerb/sommerdemo/scaling/levels.htm

Part 3 Measures of Central Tendency (Averages) and Some Descriptive Statistics

The Mean

$$\bar{x} = \frac{\sum x_i}{n}$$

Mean height of policemen

$$\frac{\text{Sum of all heights of officers}}{\text{Number of officers}}$$

Is the mean a suitable summary for categorical variables?

The average number of legs for participants in the Marathon des Sables is 1.9?

The Mode

- A measure of central tendency for categorical data
- The most common category

Table 8 Police officer strength by ethnicity, police force area and gender, as at 31 March 2012

Full-time equivalents [1]

Police force	Female							
	White	Mixed	Black or Black British	Asian or Asian British	Chinese or Other ethnic group	Not stated	**Total female**	**Total all staff**
British Transport Police	344	10	9	8	1	20	**391**	**2,557**

The Median

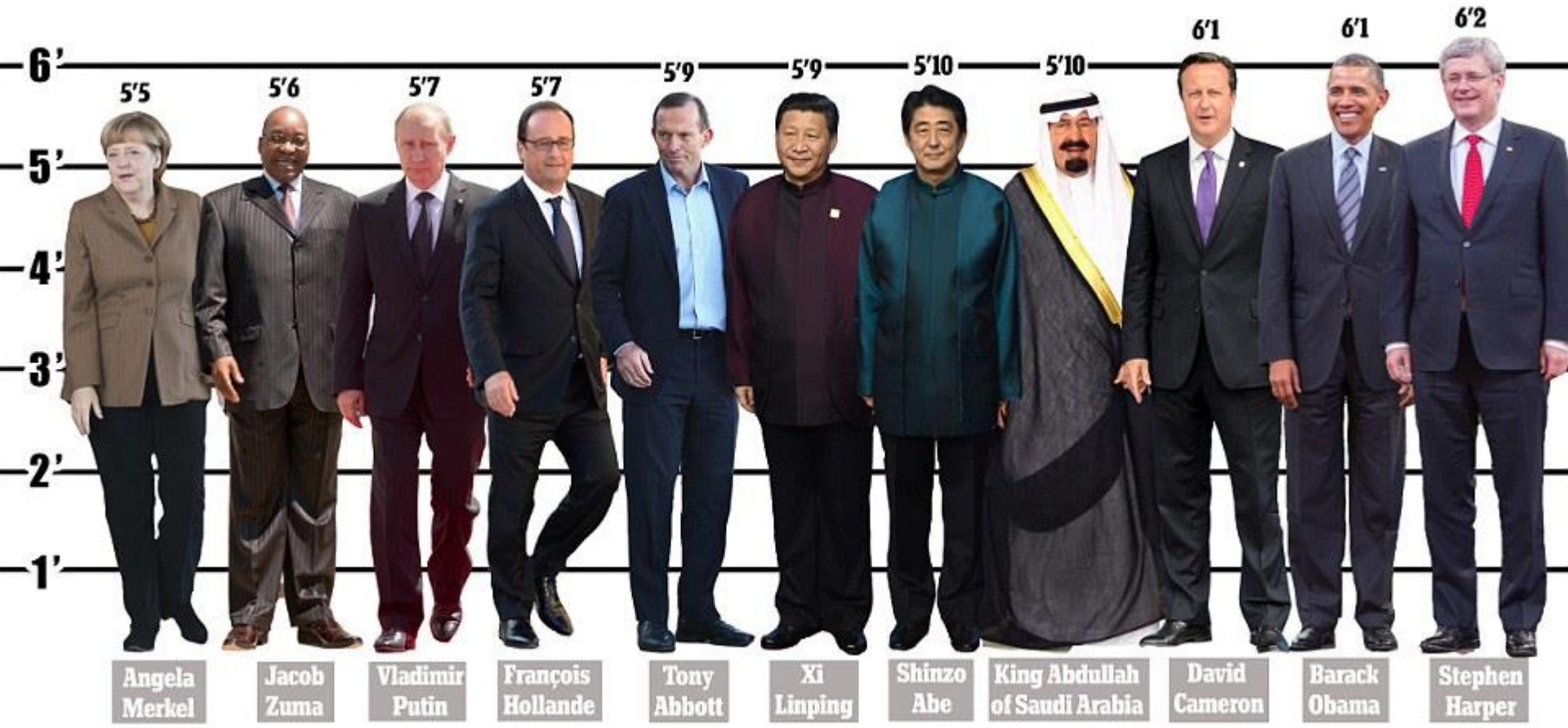
203 cm
193 cm
193 cm
191 cm
188 cm

---

185 cm **The 50th Percentile**

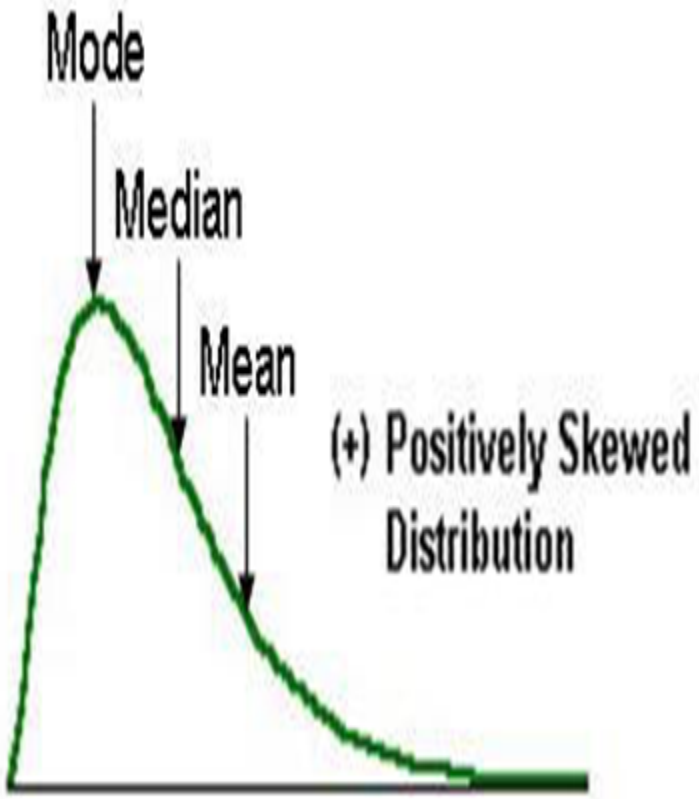
183 cm
180 cm
175 cm
173 cm
168 cm

The Median


6'
5'
4'
3'
2'
1'
5'5
5'6
5'7
5'7
5'9
5'9
5'10
5'10
6'1
6'1
6'2
Angela Merkel
Jacob Zuma
Vladimir Putin
François Hollande
Tony Abbott
Xi Linping
Shinzo Abe
King Abdullah of Saudi Arabia
David Cameron
Barack Obama
Stephen Harper

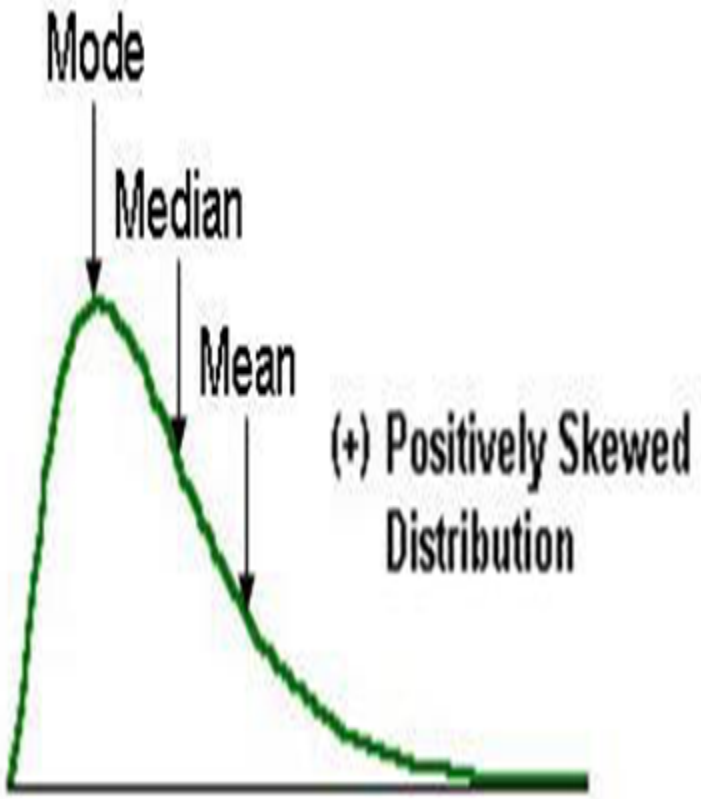
The Shape of a Distribution

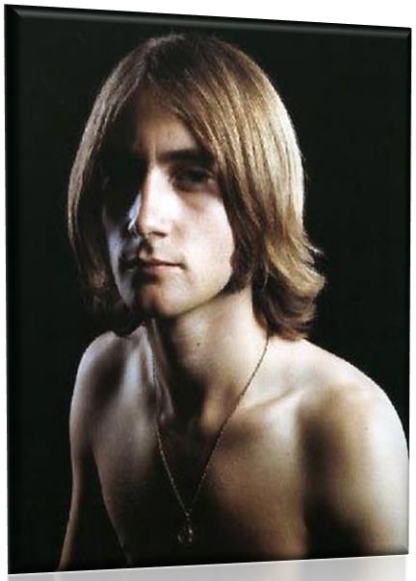
- For **positively skewed data**, the mean has a higher value than the median, and the median has a higher value than the mode.



Positive skew: Mode < Median < Mean

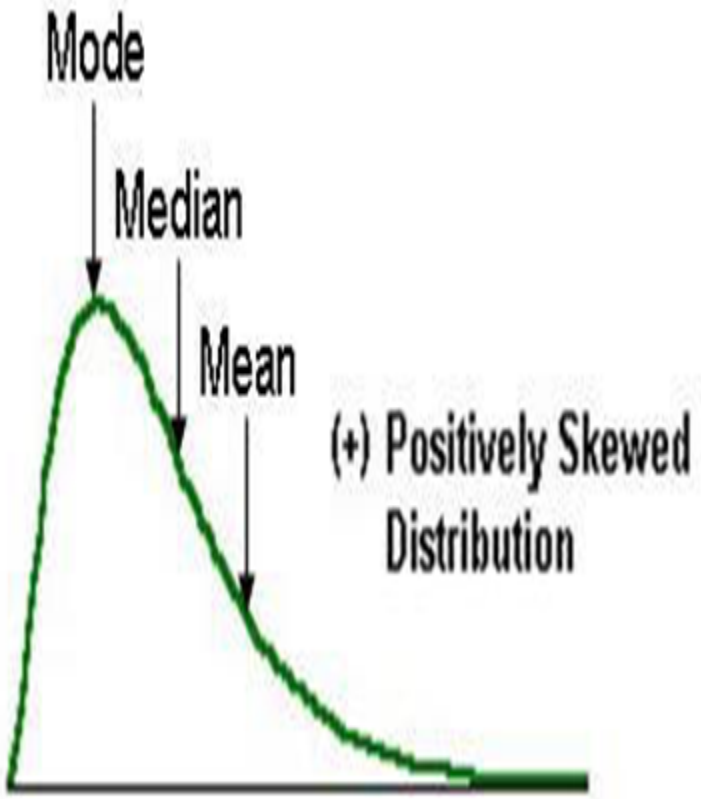
- For **positively skewed data**, the mean has a higher value than the median, and the median has a higher value than the mode.

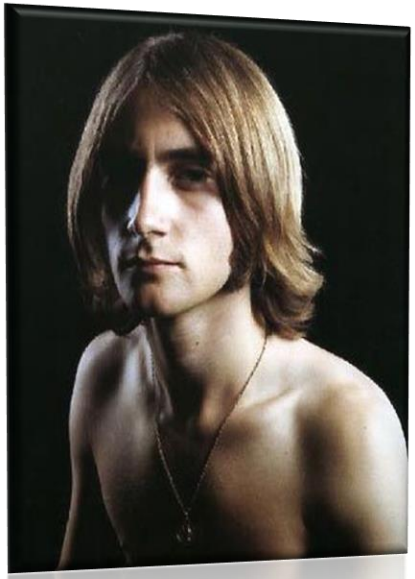




Positive skew: Mode < Median < Mean

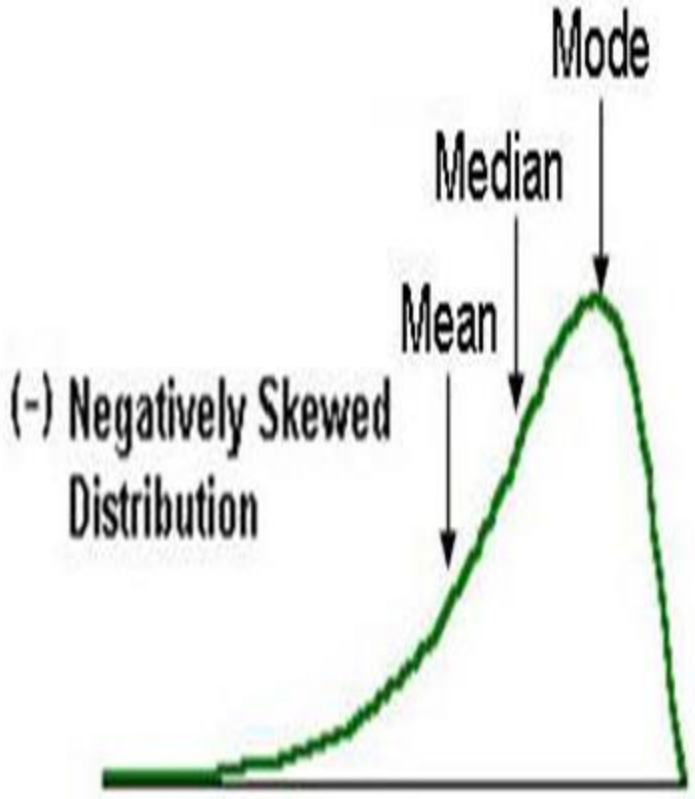
- For **positively skewed data**, the mean has a higher value than the median, and the median has a higher value than the mode.





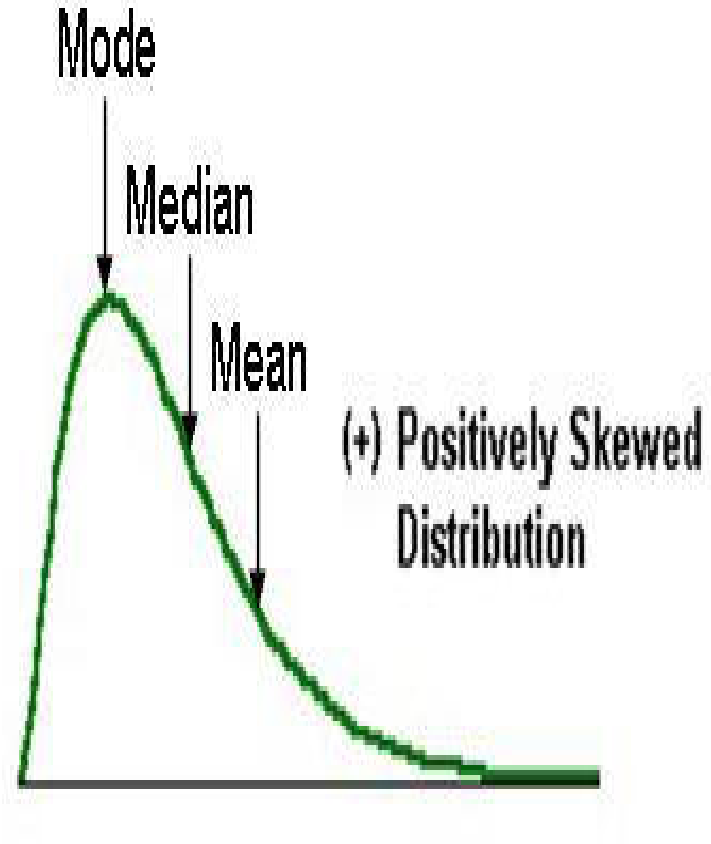
Positive skew: Mode < Median < Mean

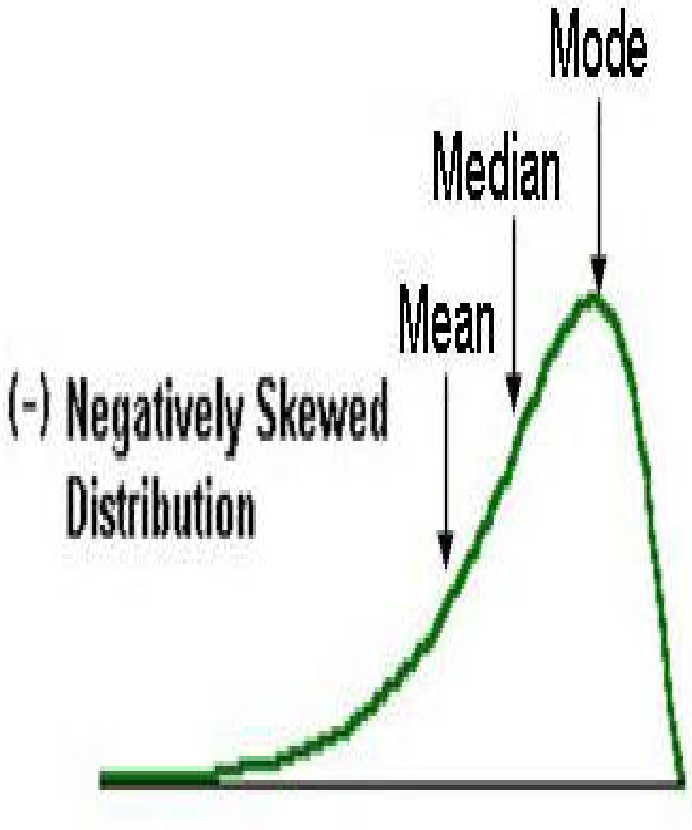
- For **negatively skewed data**, the mean has a lower value than the median, and the median has a lower value than the mode.



Negative skew: Mean < Median < Mode

- For **positively skewed data**, the mean has a higher value than the median, and the median has a higher value than the mode
- For **negatively skewed data**, the mean has a lower value than the median, and the median has a lower value than the mode.





Positive skew: Mode < Median < Mean

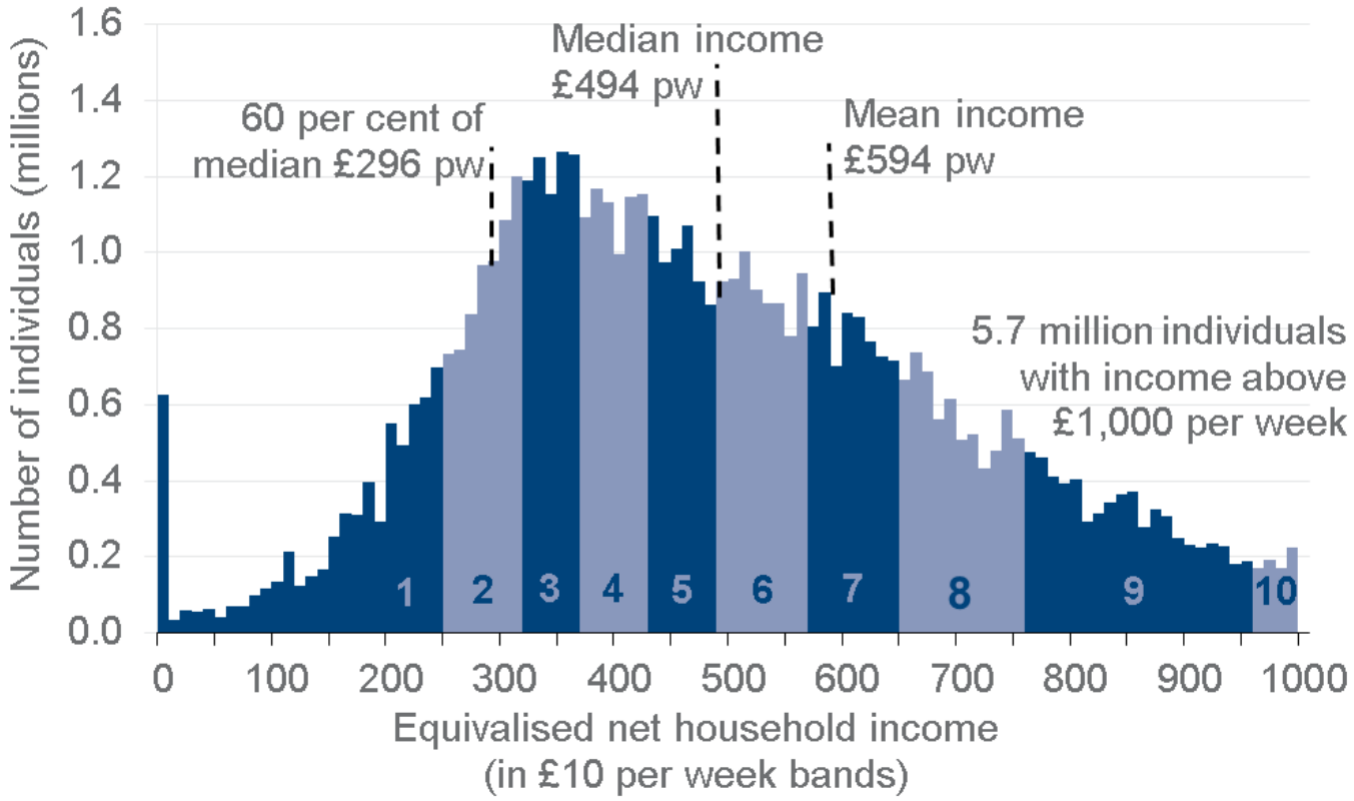
Negative skew: Mean < Median < Mode

Department for Work & Pensions

Households Below Average Income: An analysis of the UK income distribution: 1994/95-2016/17

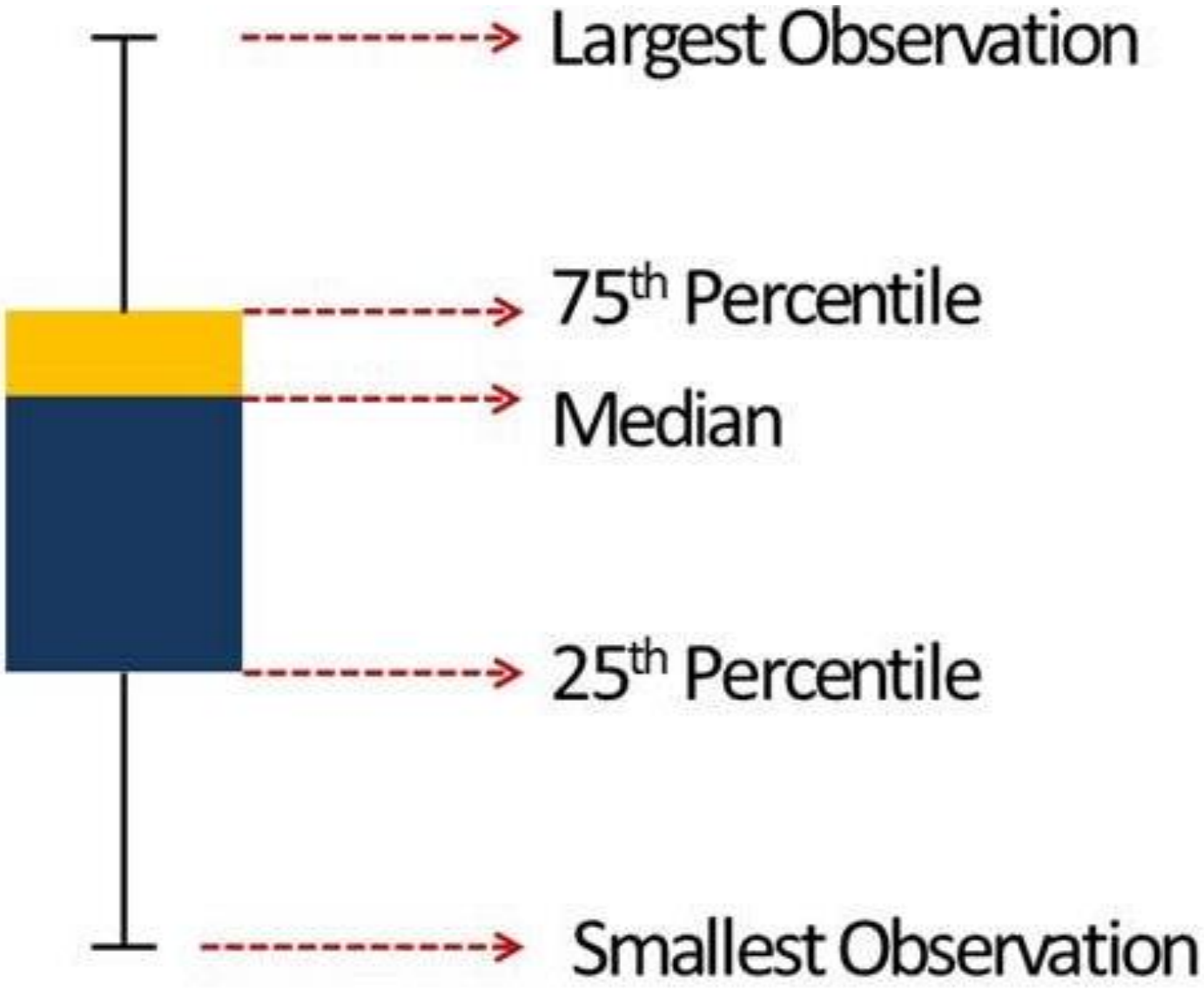
Annual | Financial year 2016/17 | Published: 22 March 2018 | United Kingdom | National Statistics

Income distribution (BHC) for the total population, 2016/17



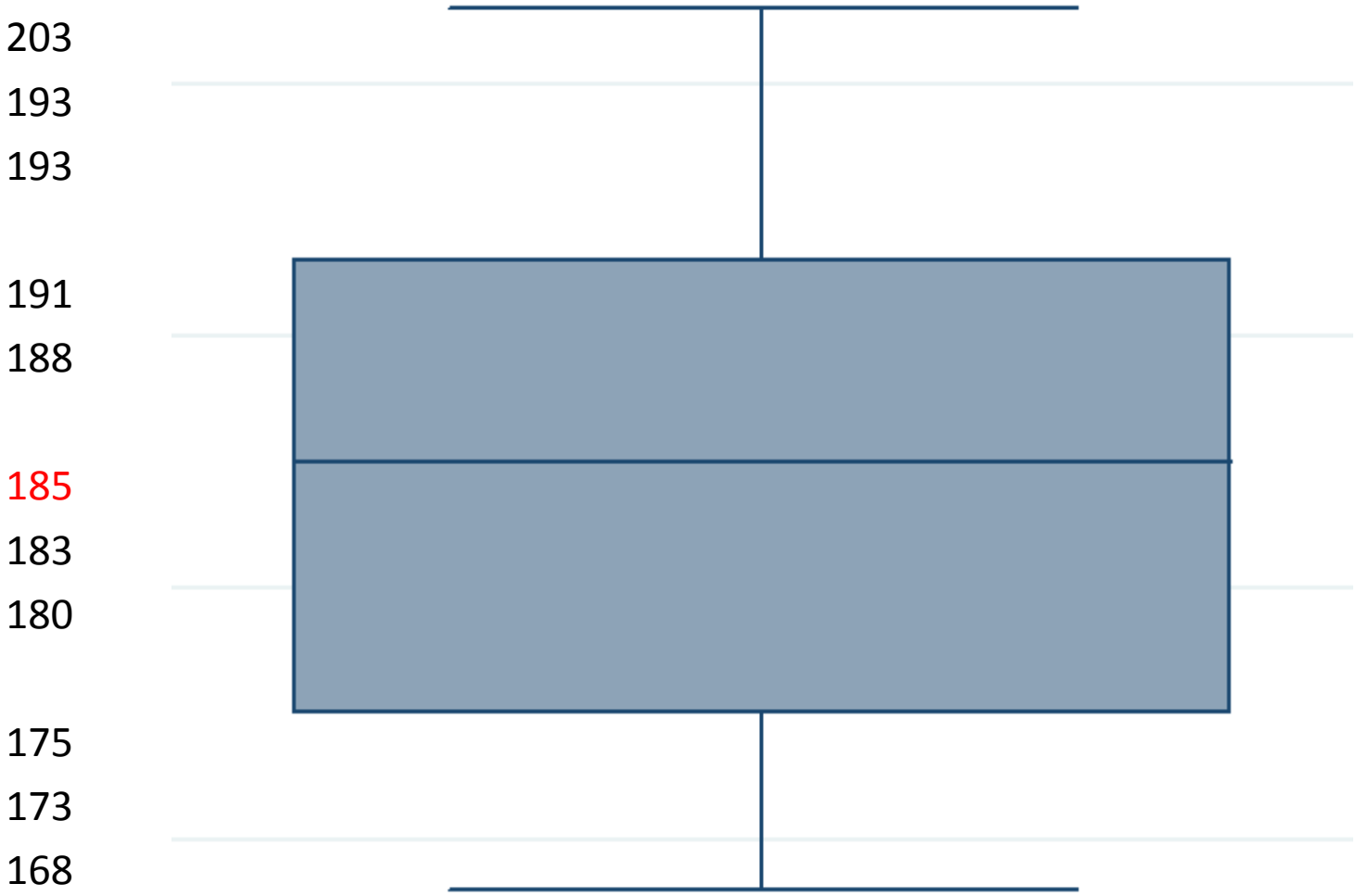
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/691917/households-below-average-income-1994-1995-2016-2017.pdf

A BOX PLOT


Largest Observation
75th Percentile
Median
25th Percentile
Smallest Observation

Phoenix Mercury WNBA Team Mean Height 184.73 cm (just under 6’ 1’’)

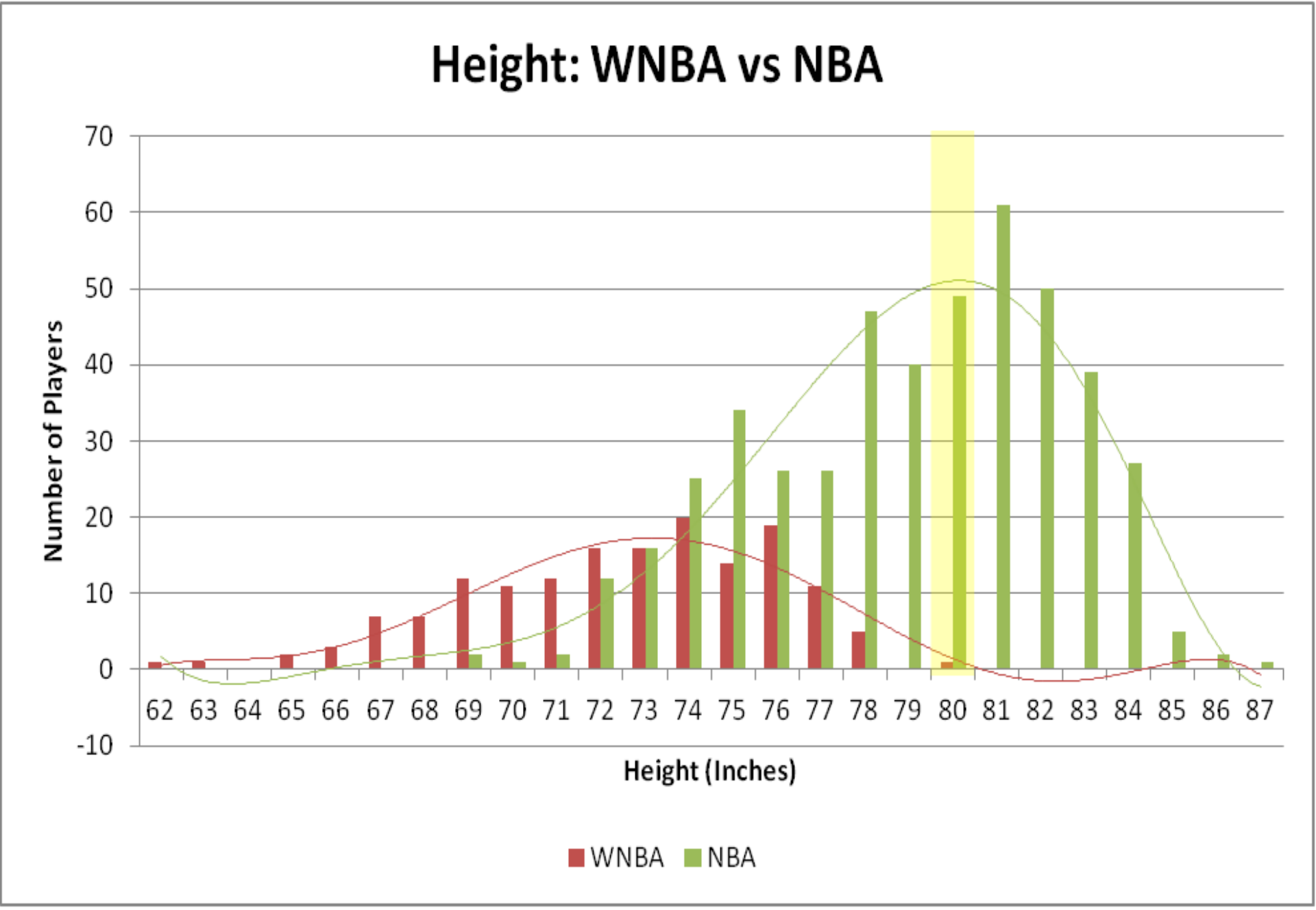
A Box Plot



Brittney Griner 6’ 8”

Tallest woman in the WNBA

But how tall would she be in the NBA?



http://wagesofwins.com/2013/04/10/could-brittney-griner-make-it-in-the-nba/

Series 6 Episode 8 (The 43 Peculiarity)

An Olympic Example

Fanshawe, T. (2012) Seven into Two : Principal components analysis and the Olympic Heptathlon, *Significance*, April p.40-42.

Some Multi-Event Competitions

Decathlon	***Heptathlon***	***Modern Pentathlon***
100m	100m hurdles	Shooting
Long jump	High jump	Fencing
Shot put	Shot put	Swimming
High jump	200m	Riding
400m	Long jump	Running
110m hurdles	Javelin	
Discus	800m	
Pole vault		
Javelin		
1500m		

In Greek mythology the hero Jason of Argonaut is credited with inventing the pentathlon. Modern pentathlon was brainchild of Pierre de Coubertin.

Heptathlon

- 2008 Olympic winner was Ukrainian Nataliya Dobrynska won with 6733pts

- Jackie Joyner-Kersee set world record with 7291pts in 1988

Left: Heptathlete
Jackie Joyner-Kersee

Above: Her colourful sister-in-law Florence Griffith-Joyner

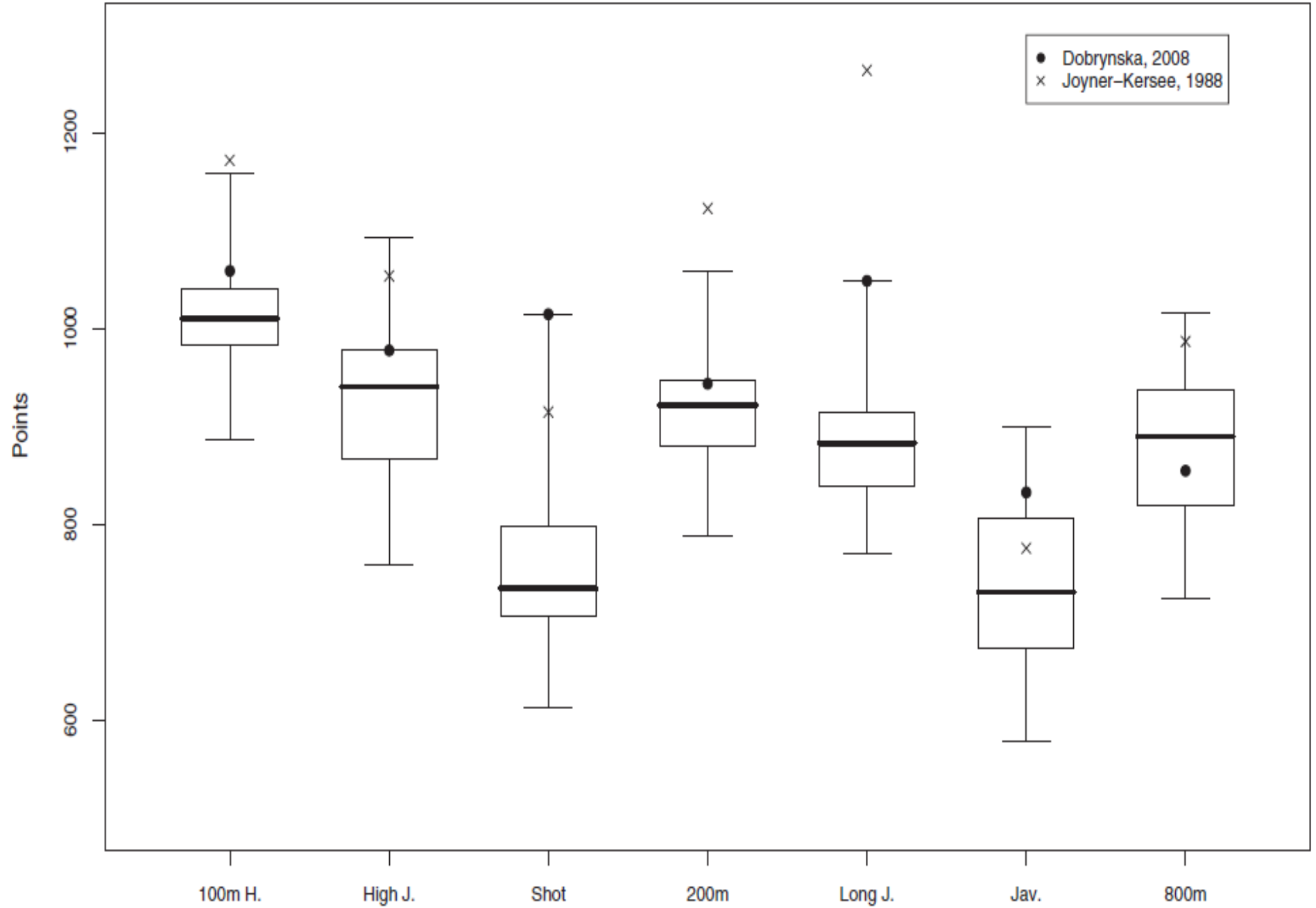


Figure 2. Boxplot of scores for each event in the 2008 heptathlon, with scores from Dobrynska (2008) and Joyner-Kersee (1988) marked. Note that the scoring system for the javelin changed between 1988 and 2008

- 1988 Ben Johnson was stripped of his WR and gold medal after testing positive for drugs

- International Olympic Committee formed the World Anti-Doping Agency

- Anti-doping testing increases

Part 4 Measures of Dispersion

Standard Deviation & Variance

- Arithmetic mean:

$$\bar{x} = \frac{\sum x_i}{n}$$

- Standard Deviation:

$$s = \sqrt{\frac{\sum \left(x_i - \bar{x} \right)^2}{n-1}}$$

- Variance: s^2

Standard Deviation & Variance

- Arithmetic mean:

$$\bar{x} = \frac{\sum x_i}{n}$$

Standard Deviation & Variance

- Standard Deviation:

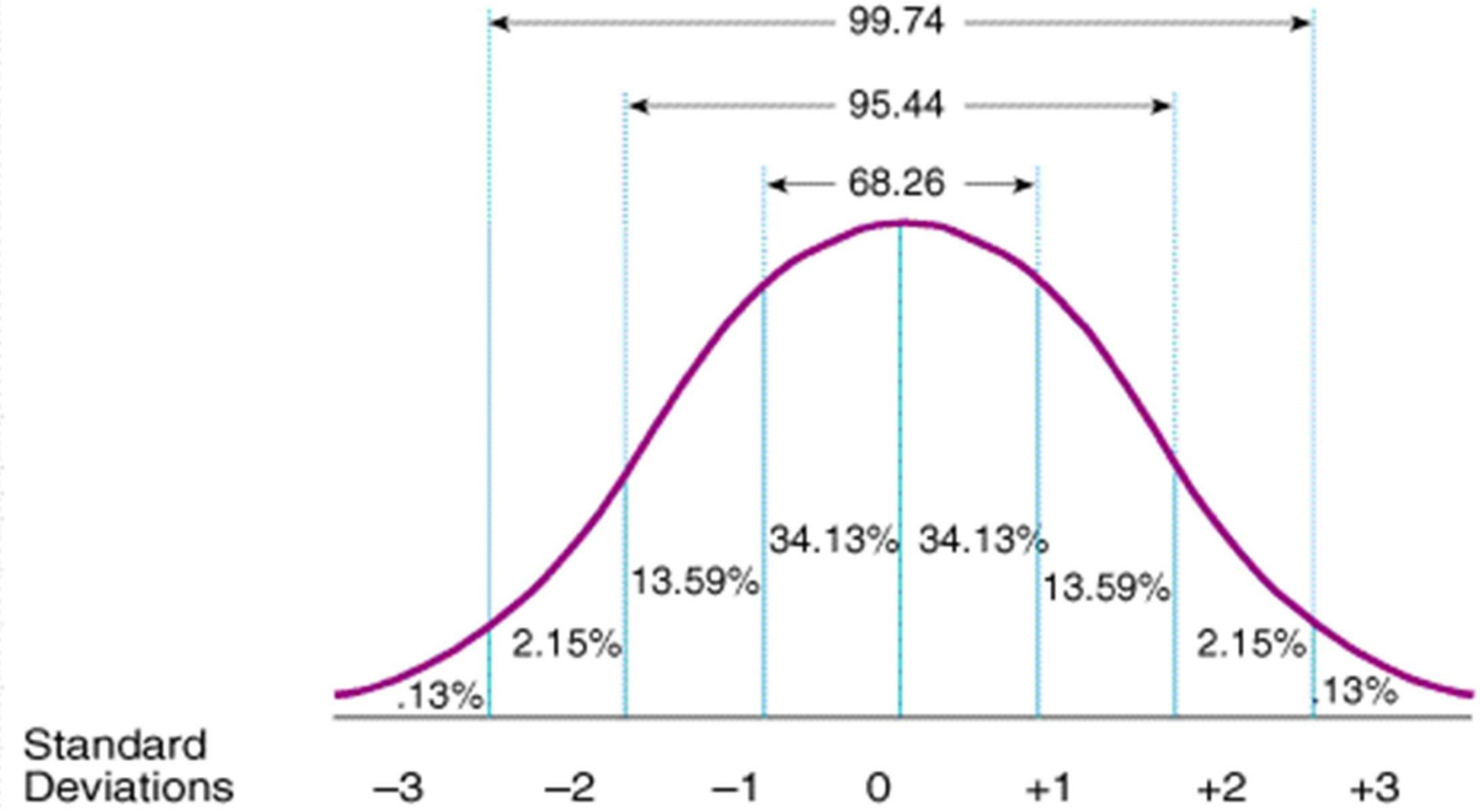
$$s = \sqrt{\frac{\sum \left(x_i - \bar{x} \right)^2}{n-1}}$$

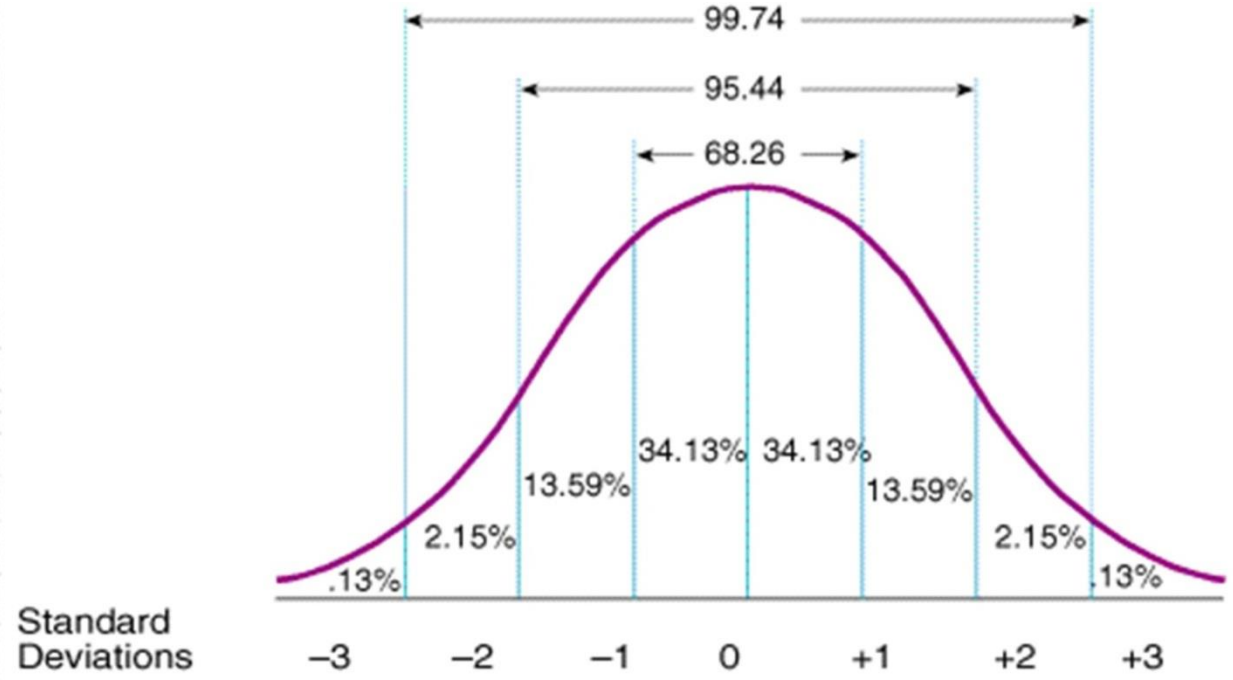
Standard Deviation & Variance

- Standard Deviation:

$$s = \sqrt{\frac{\sum \left(x_i - \bar{x} \right)^2}{n-1}}$$

- Variance: s^2


99.74
95.44
68.26
34.13%
34.13%
13.59%
13.59%
2.15%
2.15%
.13%
.13%
Standard Deviations
–3
–2
–1
0
+1
+2
+3



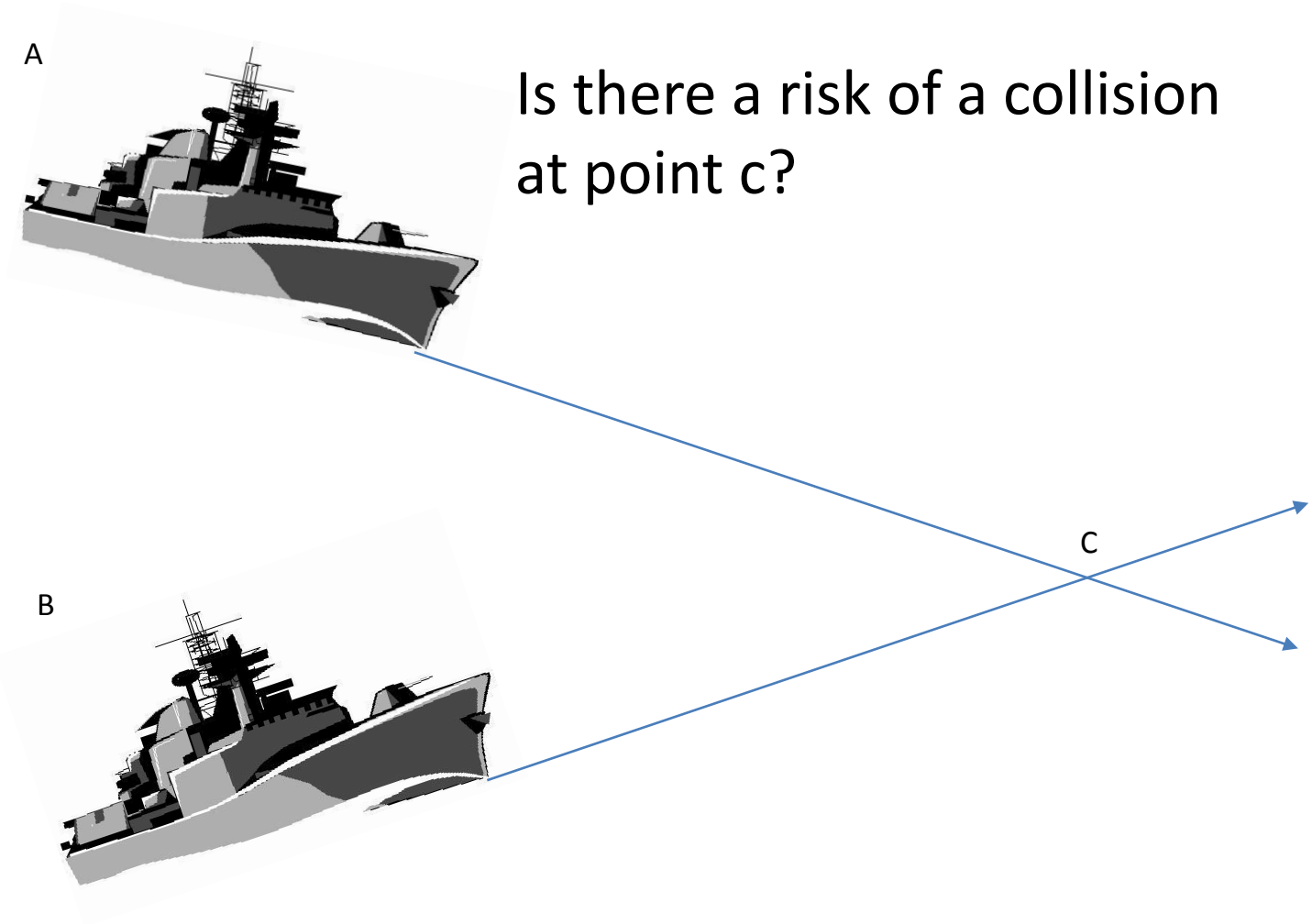
The number of days that Mr Smith in HR was ill last year was more than three standard deviations above the mean -

Is this level of absence unusual?

Standard Error of the Mean

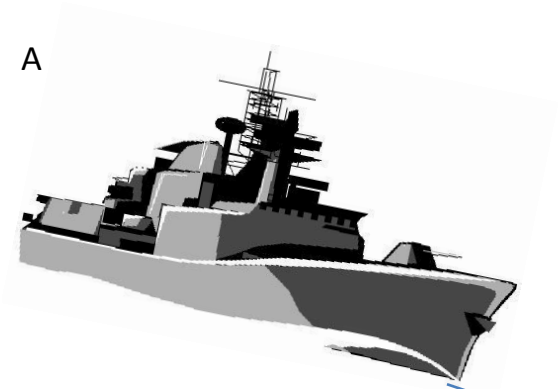
- Standard error of the mean = $\frac{SD}{\sqrt{n}}$

Confidence Intervals



Is there a risk of a collision at point c?

A

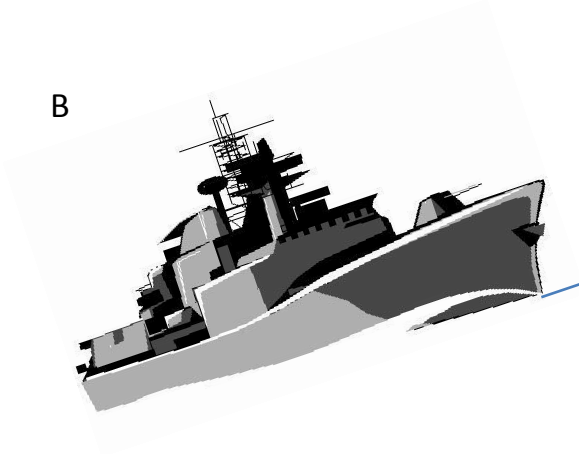


Ship A plans to be at point C at 10:00 am

95% of the time she will arrive between 9:55 am and 10:05 am

C

B

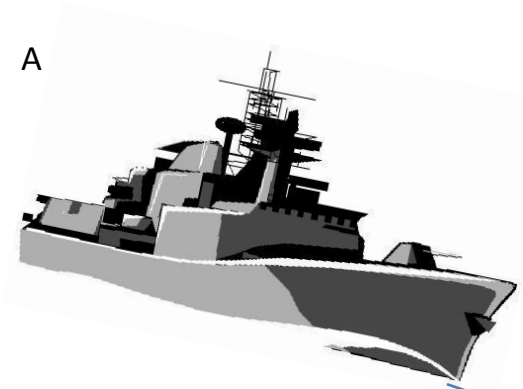


Ship B plans to be at point C at 10:15 am

95% of the time she will arrive between 10:10 am and 10:20 am

On Another Day…

A

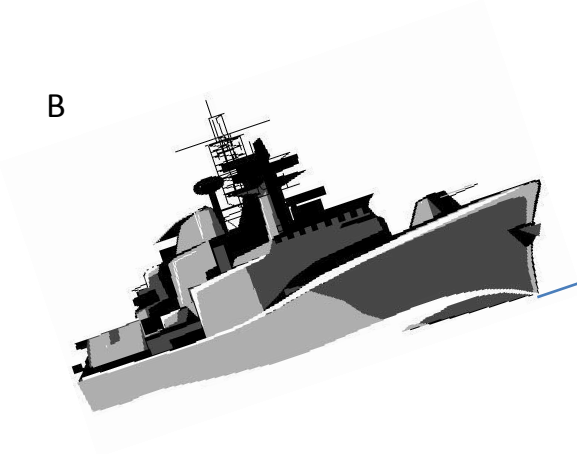


Ship A plans to be at point C at 10:00 am

95% of the time she will arrive between 9:50 am and 10:10 am

C

B



Ship B plans to be at point C at 10:15 am

95% of the time she will arrive between 10:05 am and 10:25am

TAKE HOME MESSAGE

When confidence intervals overlap then the measures are not significantly different

When there is ‘clear blue water’ there is a significant difference

95% Confidence intervals around mean

$$CI = \bar{x} \pm (1.96 se_m)$$

95% Confidence intervals around β

$$CI = \beta \pm (1.96 se_{\beta})$$

Professor Mac is always late…

Her wife accepts this but some times she is later than usual

Her wife keeps a diary and does some stats…

But how late is unreasonably late?

On the last ten occasions that they planned to go out she has been late nine times

- Minimum 0 minutes; Max 59 minutes
- Mean 16.4 minutes
- s.e. mean 5.4 minutes

Her partner constructs a confidence interval around the mean

Upper c.i. = 16.4 + (1.96* 5.4) = 26.98

Mean 16.40

Lower c.i. = 16.4 - (1.96* 5.4) = 5.82

Another way to think about this is..

Standard error

–how tightly distributed the values are grouped around the mean

Confidence intervals

– a measure of precision of an estimate (e.g. a mean)

Part 5 Getting Started


Excel

Spreadsheet can not easily be scripted –

So there will be no clear audit trail of activity

If you use a ‘gui’ then one day you will inevitably end up in a sticky mess

We advise against undertaking data analyses using software packages in interactive modes, for example through a graphical user interface (gui) such as drop-down menus

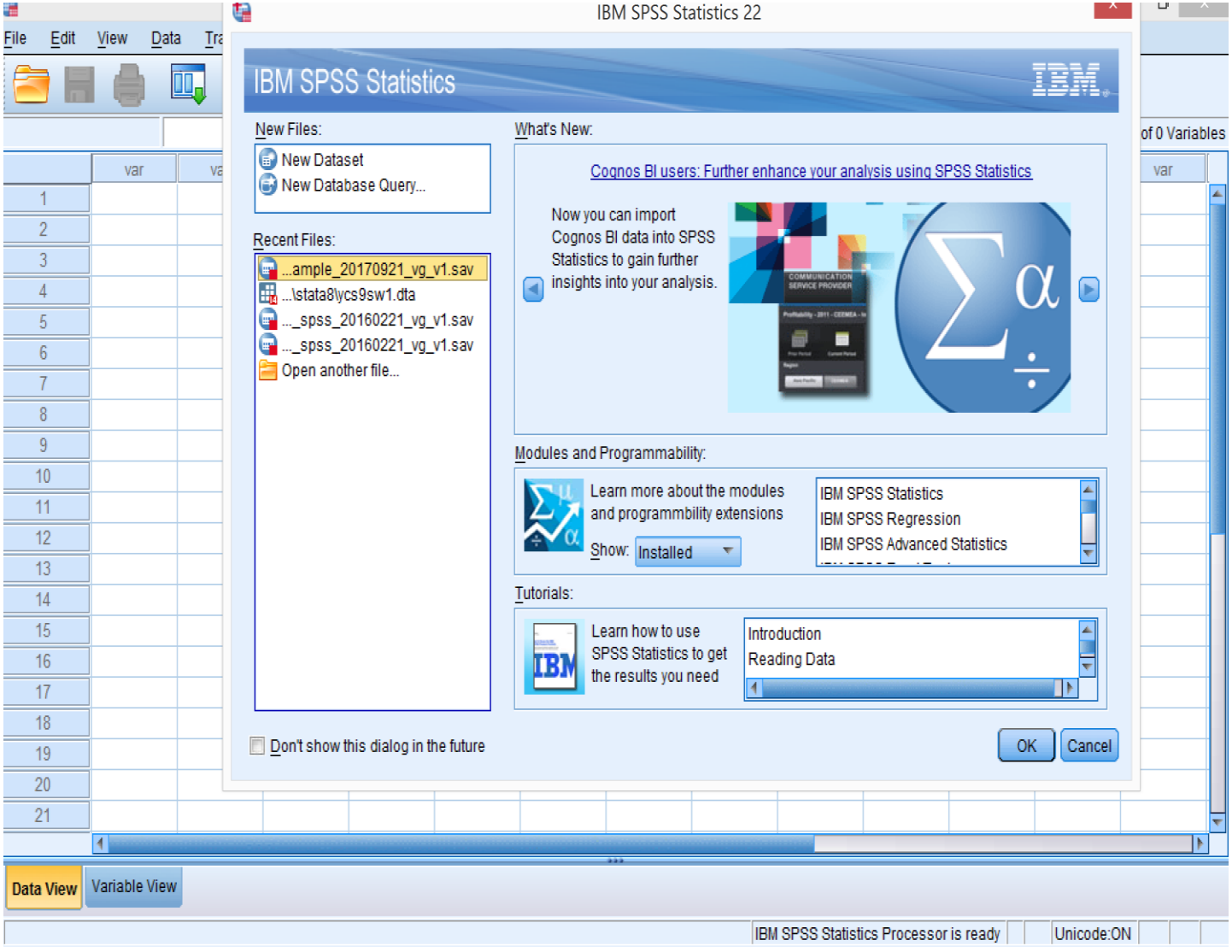
Which tool?

Which software (or data analysis package) should I learn?

"This is like asking which five countries should I know about, given that I don't have an interest in either geology or politics – the answers is it depends"

Larry Wall (the creator of the programming language Pearl)

SPSS


IBM SPSS Statistics 22
IBM SPSS Statistics
New Files:
New Dataset
New Database Query...
Recent Files:
...ample_20170921_vg_v1.sav
...\stata8\ycs9sw1.dta
..._spss_20160221_vg_v1.sav
..._spss_20160221_vg_v1.sav
Open another file...
What's New:
Cognos BI users: Further enhance your analysis using SPSS Statistics
Now you can import Cognos BI data into SPSS Statistics to gain further insights into your analysis.
Modules and Programmability:
Learn more about the modules and programmbility extensions
Show: Installed
IBM SPSS Statistics
IBM SPSS Regression
IBM SPSS Advanced Statistics
Tutorials:
Learn how to use SPSS Statistics to get the results you need
Introduction
Reading Data
Don't show this dialog in the future
OK
Cancel
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

SPSS

- Statistical Package for the Social Sciences
- Was widely taught in university departments (e.g. psychology, sociology, health etc.)
- Commercially available
- www.ibm.com/DataStatistics/SPSS

SAS

- Now only widely used in government departments
- Re-launched as a data science package
- Commercially available
- Fiddly to get started with
- www.sas.com

R

- Growing in popularity (e.g. data science, statistics, science etc.)
- Popular with statisticians
- Free (open source)
- Difficult to learn
- Development and support is not commercial
- Help resources are under-developed
- https://cran.r-project.org

Other Approaches

- Python (www.python.org)
  - Open source
  - Fewer libraries than R
  - Tricky to learn and almost no help resources

- Julia (https://julialang.org/)
  - Open source
  - Fewer libraries than R
  - Tricky to learn and no help resources

Stata

- Popular in economics globally
- Popular with US social science (e.g. sociology)
- Commercial software
- Easy to learn
- Excellent development
- Outstanding help files and teaching resources
- www.stata.com

Stata

- Does all the routine stuff (SPSS, SAS & R)
- Fits many more models than standard software (esp. for selection, longitudinal)
- Specialist complex survey analysis functions (Svy) (useful for UKHLS, MCS etc)
- Human readable format
- Very strong documentation
- Worldwide user community (lists etc.)
- New features almost daily

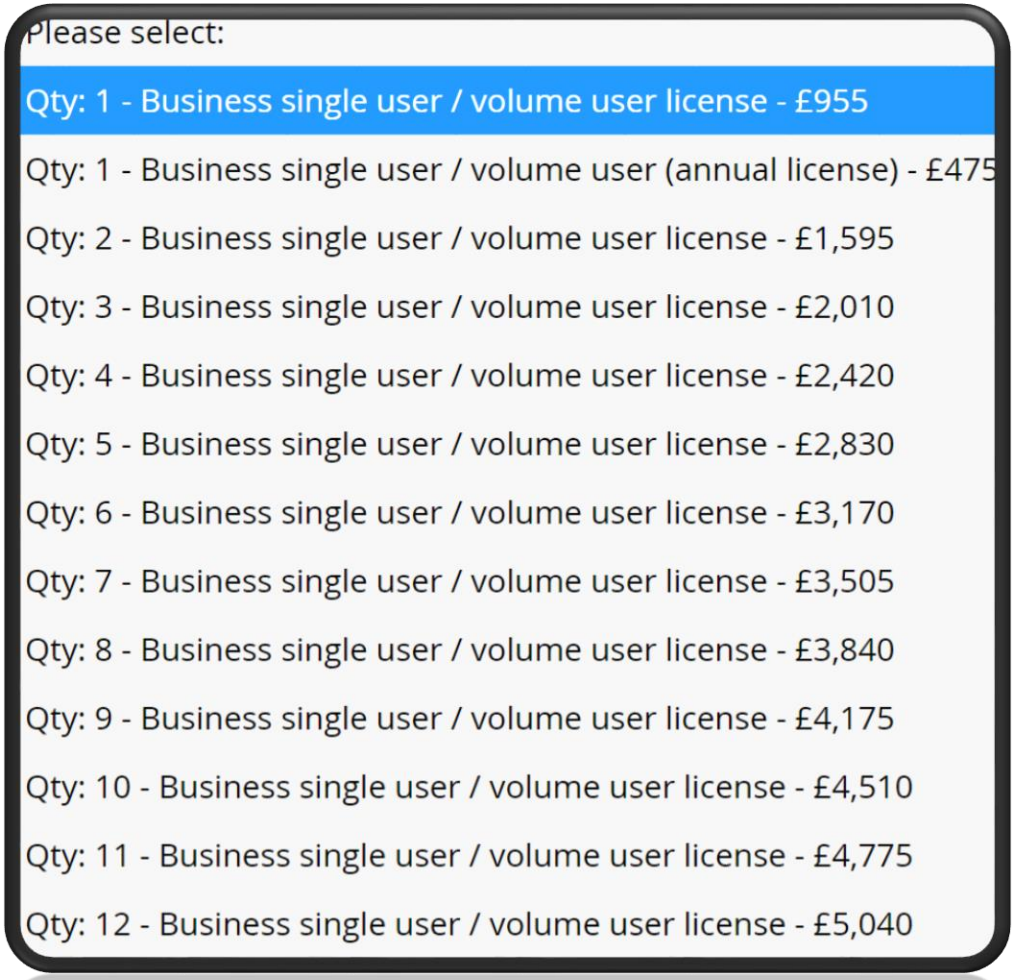
'…[Stata] has gotten better and better over time, so that by now it can happily be used as a general-purpose package. Stata is powerful and fast which makes it viable to carry out analysis on a PC…

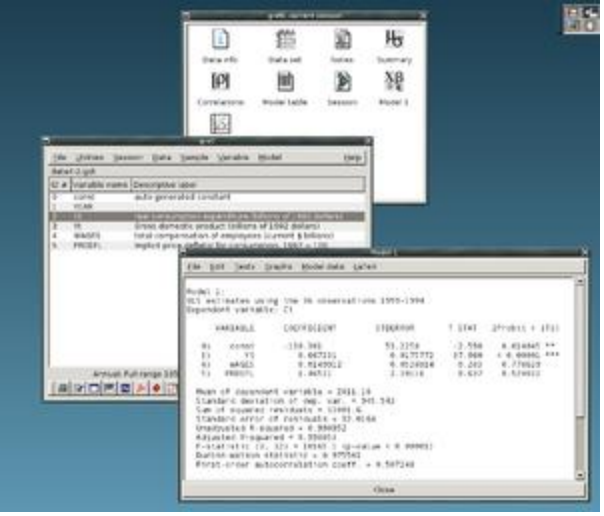
Overall, Stata is a very good choice for our kind of work'

(Treiman 2009 p.66)


Data Analysis Using Stata
Third Edition
ULRICH KOHLER
FRAUKE KREUTER
Stata Press
The Workflow of Data Analysis Using Stata
J. SCOTT LONG
Mehmet Mehmetoglu
Tor Georg Jakobsen
APPLIED STATISTICS USING STATA
A Guide for the Social Sciences
companion website
David Pevalin
Karen Robson
the stata survival manual

http://www.timberlake.co.uk/software/stata/business/single/prices.html



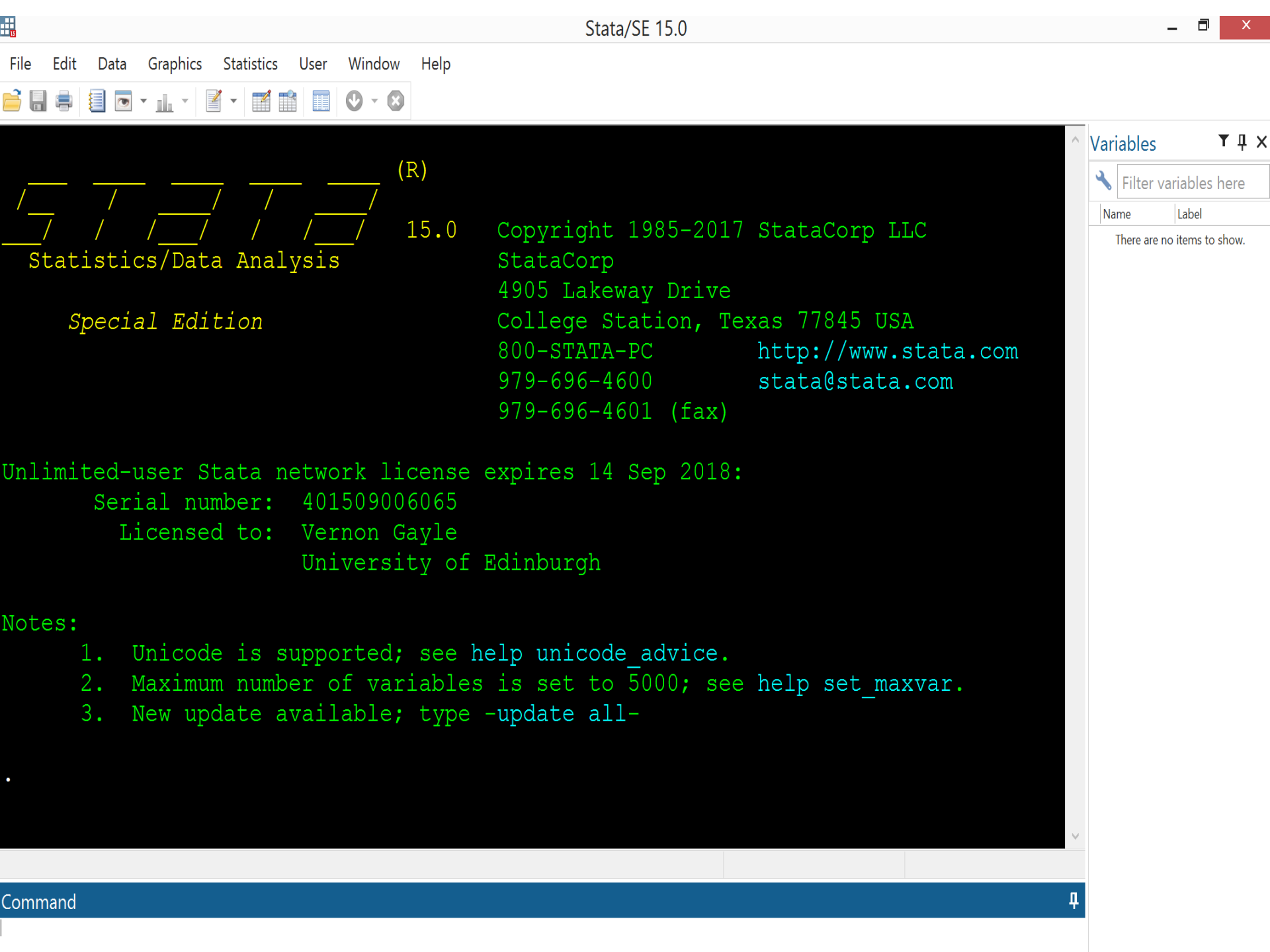


gretl

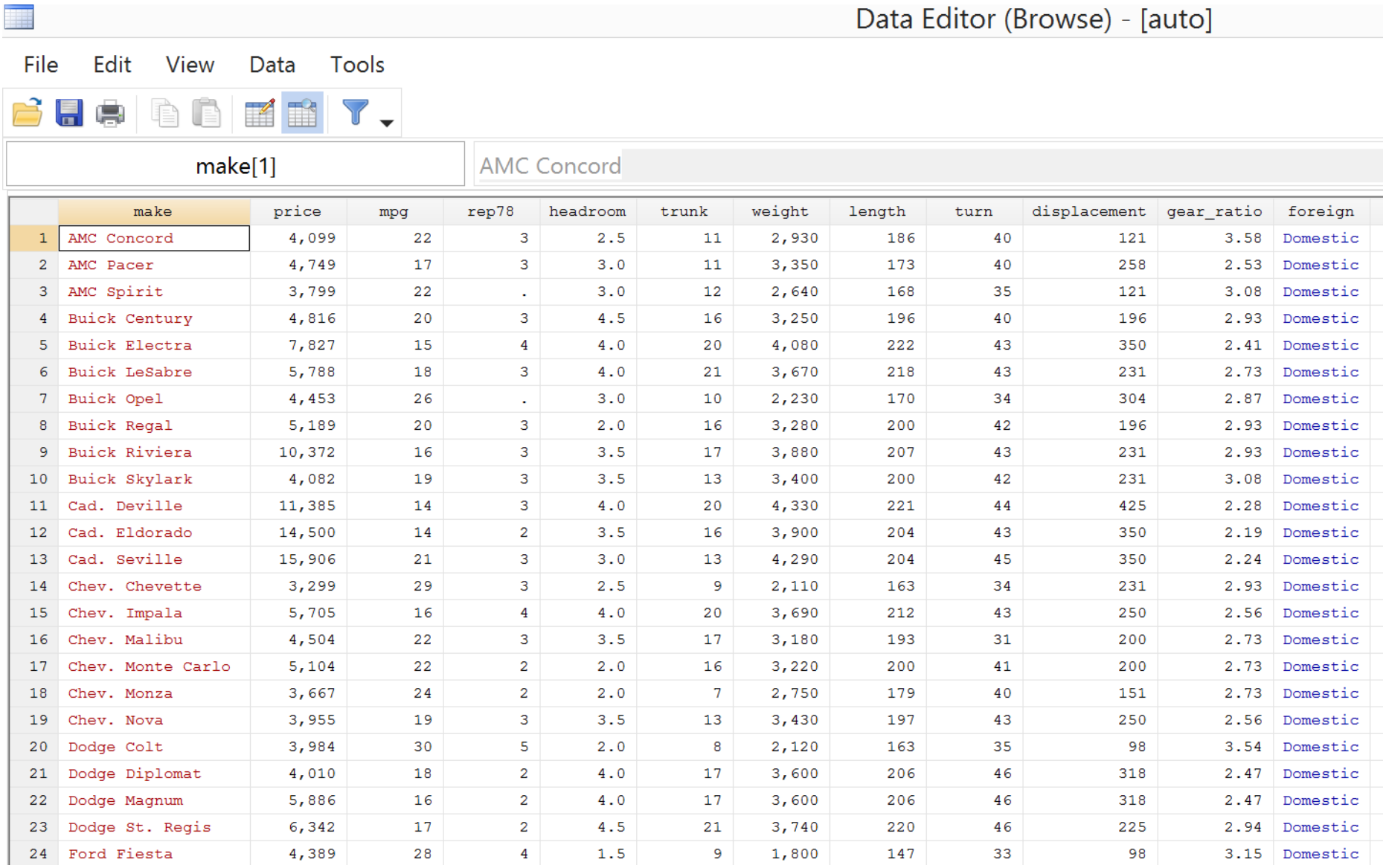
- Mainly for econometrics
- Open Source

  acronym for **g**nu **r**egression, **e**conometrics and **t**ime-series **l**ibrary
- https://en.wikipedia.org/wiki/Gretl
- https://sourceforge.net/projects/gretl/

Using Stata

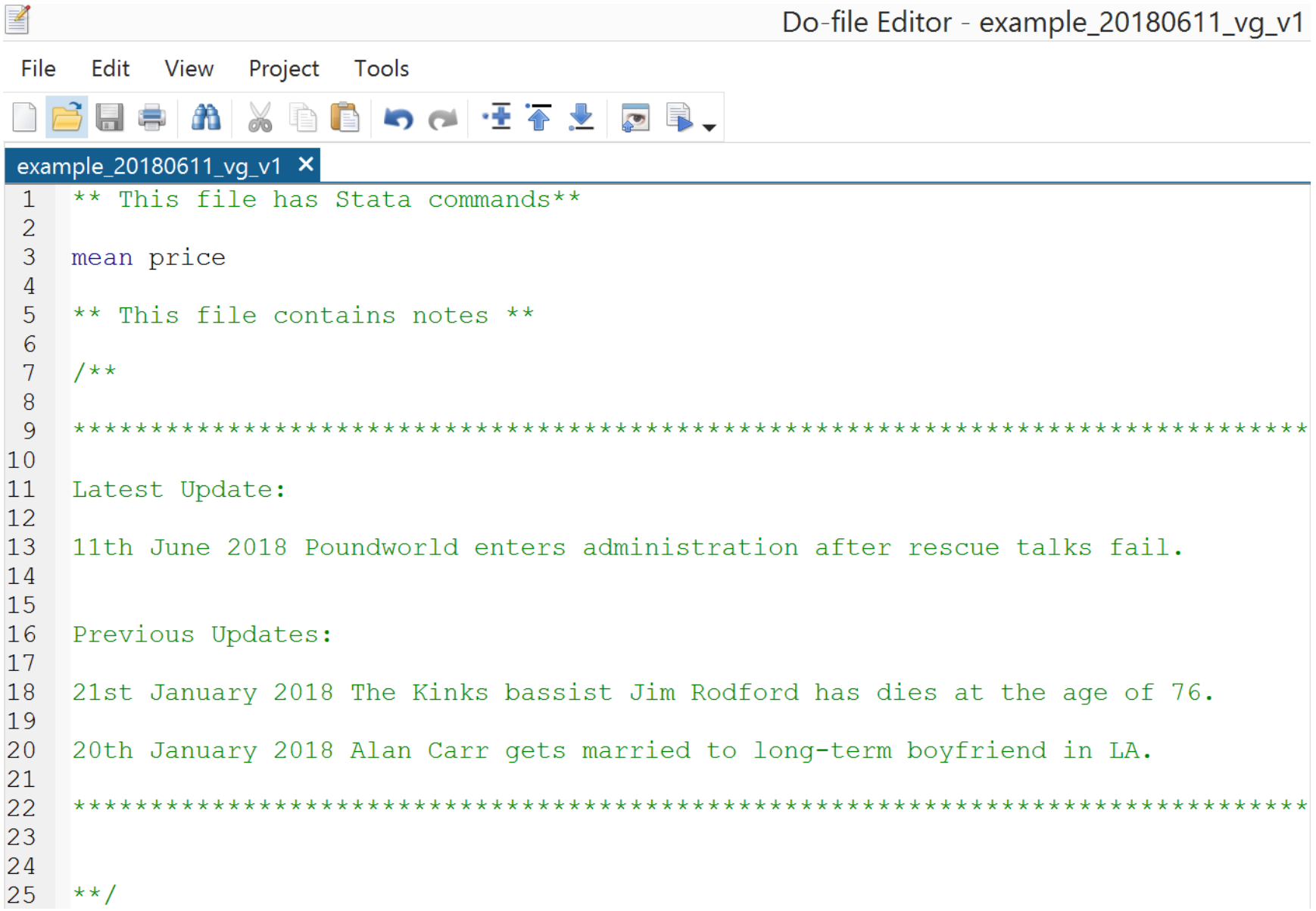

Stata/SE 15.0
File Edit Data Graphics Statistics User Window Help
(R)
Statistics/Data Analysis
15.0
Special Edition
Copyright 1985-2017 StataCorp LLC
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC http://www.stata.com
979-696-4600 stata@stata.com
979-696-4601 (fax)
Unlimited-user Stata network license expires 14 Sep 2018:
Serial number: 401509006065
Licensed to: Vernon Gayle
University of Edinburgh
Notes:
1. Unicode is supported; see help unicode_advice.
2. Maximum number of variables is set to 5000; see help set_maxvar.
3. New update available; type -update all-
.
Variables
Filter variables here
Name
Label
There are no items to show.
Command

Data Editor (familiar spread sheet)

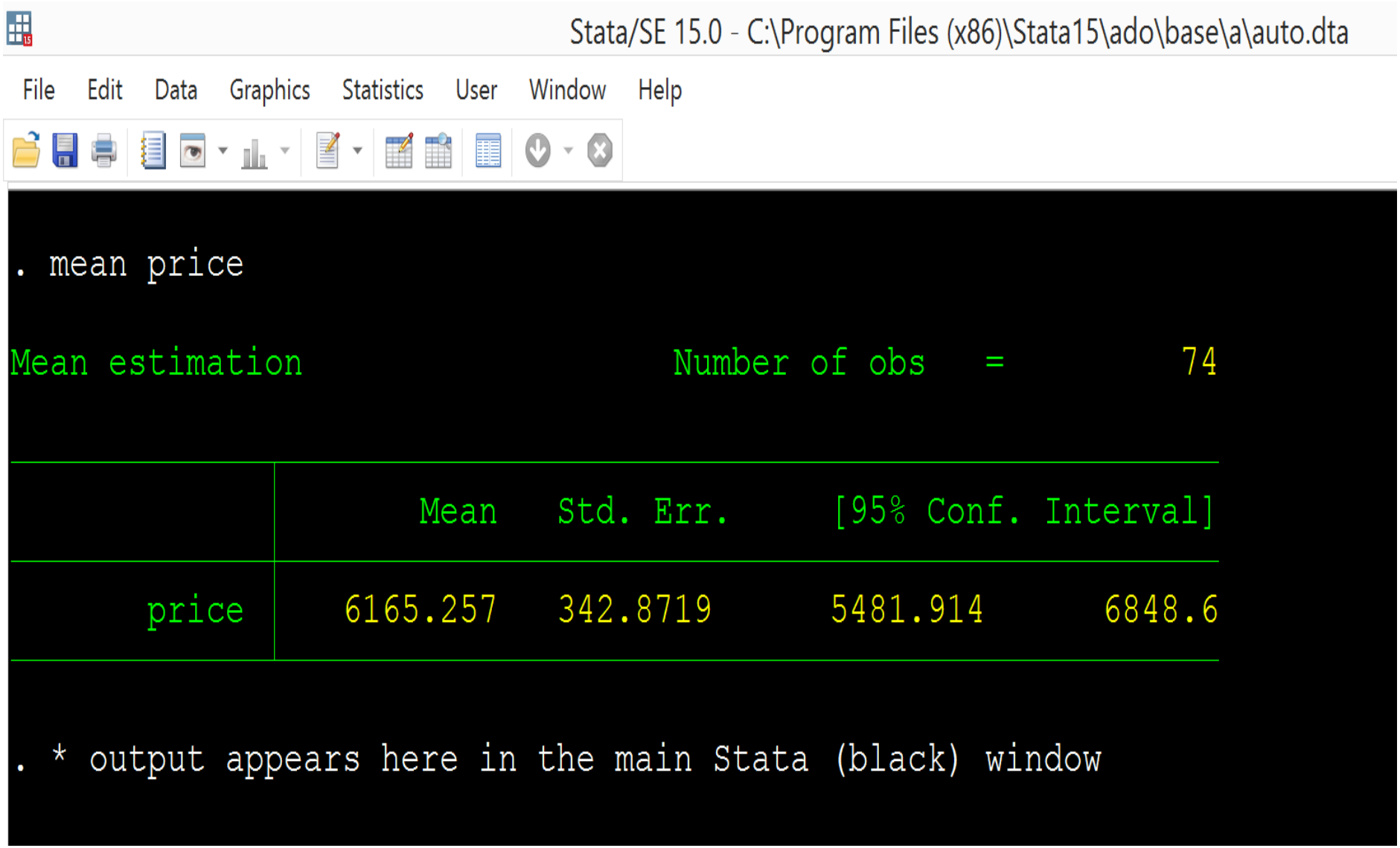


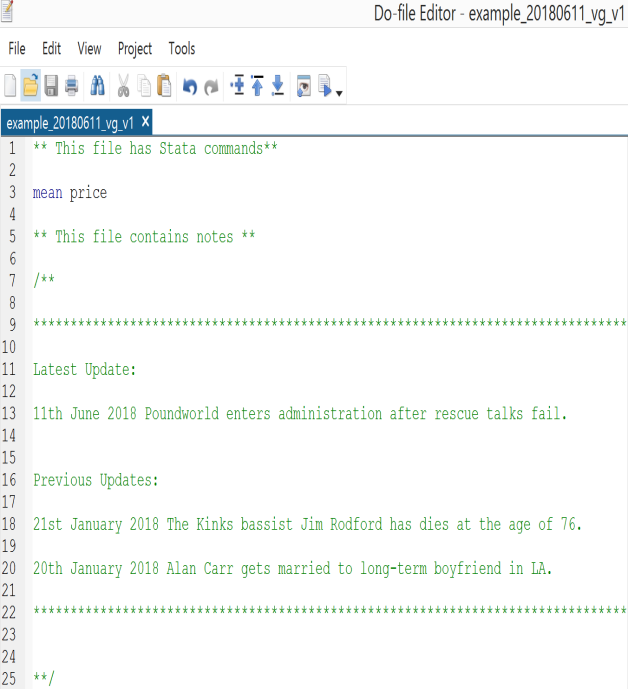
	make	price	mpg	rep78	headroom	trunk	weight	length	turn	displacement	gear_ratio	foreign
1	AMC Concord	4,099	22	3	2.5	11	2,930	186	40	121	3.58	Domestic
2	AMC Pacer	4,749	17	3	3.0	11	3,350	173	40	258	2.53	Domestic
3	AMC Spirit	3,799	22	.	3.0	12	2,640	168	35	121	3.08	Domestic
4	Buick Century	4,816	20	3	4.5	16	3,250	196	40	196	2.93	Domestic
5	Buick Electra	7,827	15	4	4.0	20	4,080	222	43	350	2.41	Domestic
6	Buick LeSabre	5,788	18	3	4.0	21	3,670	218	43	231	2.73	Domestic
7	Buick Opel	4,453	26	.	3.0	10	2,230	170	34	304	2.87	Domestic
8	Buick Regal	5,189	20	3	2.0	16	3,280	200	42	196	2.93	Domestic
9	Buick Riviera	10,372	16	3	3.5	17	3,880	207	43	231	2.93	Domestic
10	Buick Skylark	4,082	19	3	3.5	13	3,400	200	42	231	3.08	Domestic
11	Cad. Deville	11,385	14	3	4.0	20	4,330	221	44	425	2.28	Domestic
12	Cad. Eldorado	14,500	14	2	3.5	16	3,900	204	43	350	2.19	Domestic
13	Cad. Seville	15,906	21	3	3.0	13	4,290	204	45	350	2.24	Domestic
14	Chev. Chevette	3,299	29	3	2.5	9	2,110	163	34	231	2.93	Domestic
15	Chev. Impala	5,705	16	4	4.0	20	3,690	212	43	250	2.56	Domestic
16	Chev. Malibu	4,504	22	3	3.5	17	3,180	193	31	200	2.73	Domestic
17	Chev. Monte Carlo	5,104	22	2	2.0	16	3,220	200	41	200	2.73	Domestic
18	Chev. Monza	3,667	24	2	2.0	7	2,750	179	40	151	2.73	Domestic
19	Chev. Nova	3,955	19	3	3.5	13	3,430	197	43	250	2.56	Domestic
20	Dodge Colt	3,984	30	5	2.0	8	2,120	163	35	98	3.54	Domestic
21	Dodge Diplomat	4,010	18	2	4.0	17	3,600	206	46	318	2.47	Domestic
22	Dodge Magnum	5,886	16	2	4.0	17	3,600	206	46	318	2.47	Domestic
23	Dodge St. Regis	6,342	17	2	4.5	21	3,740	220	46	225	2.94	Domestic
24	Ford Fiesta	4,389	28	4	1.5	9	1,800	147	33	98	3.15	Domestic

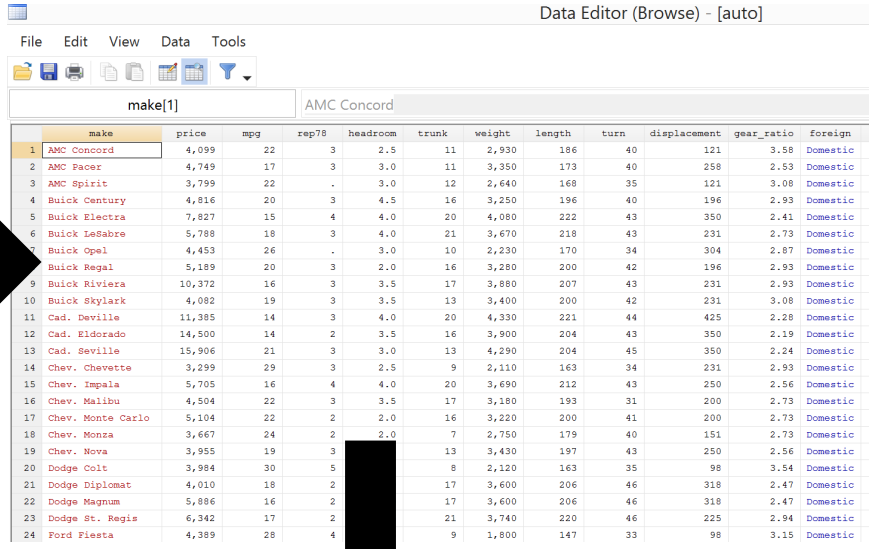
Syntax File .do



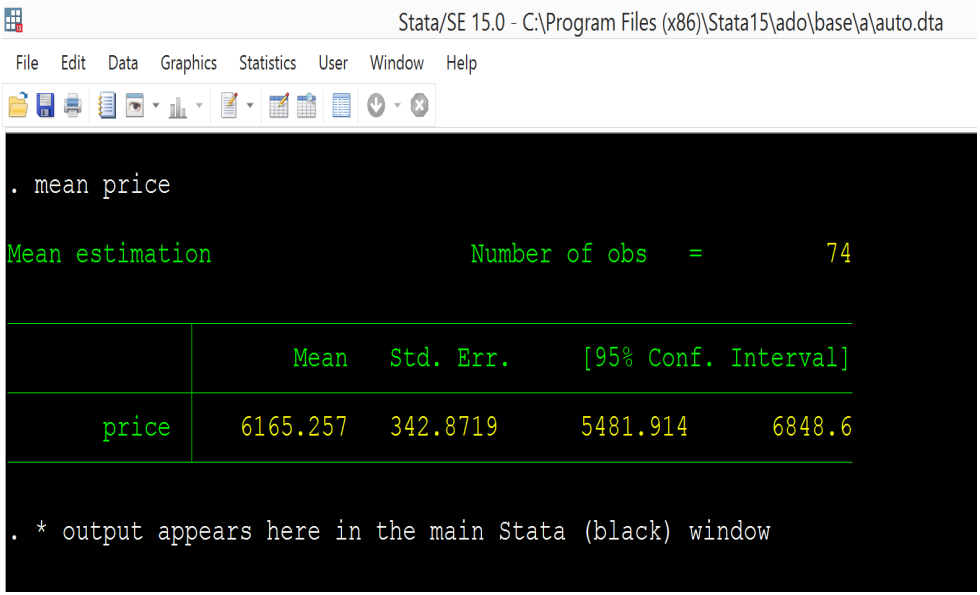
Stata Main Window (output)





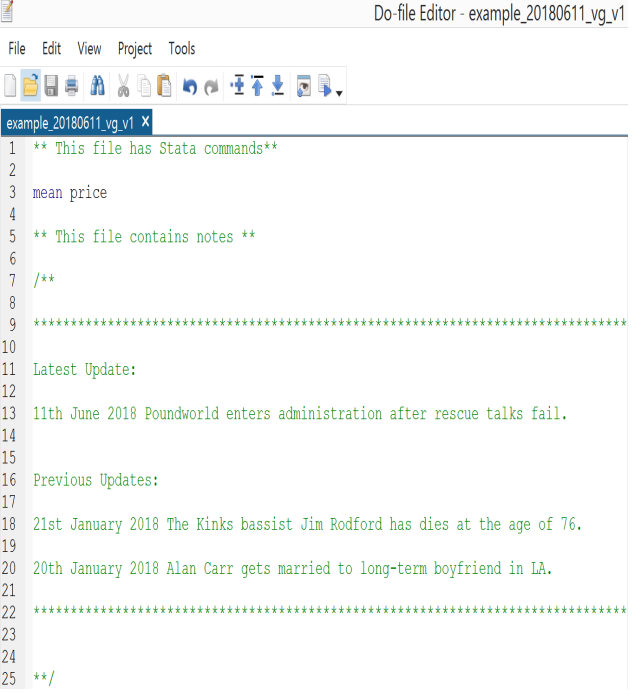


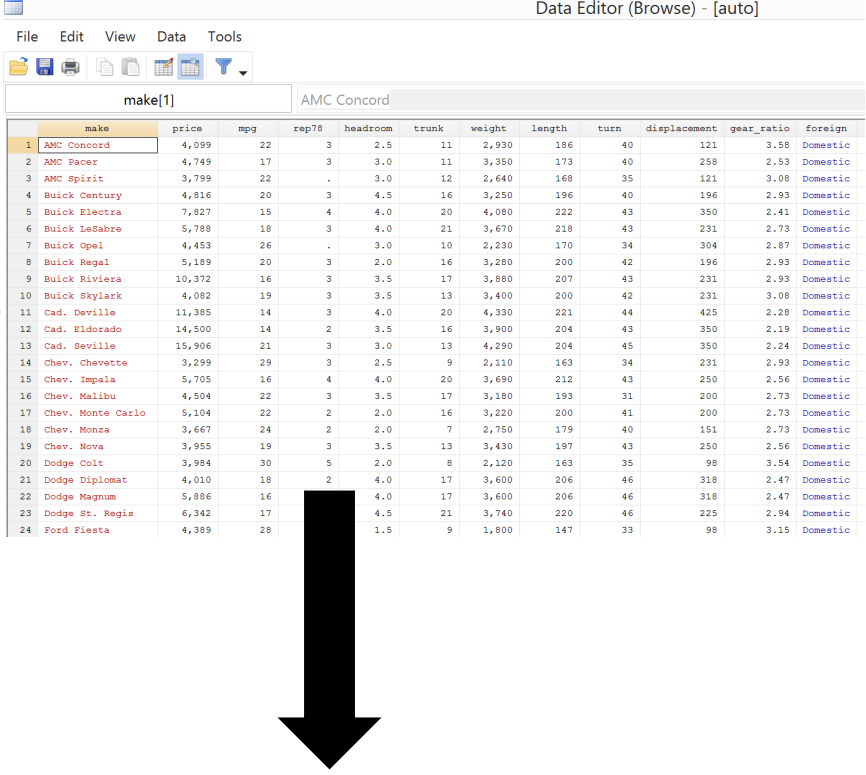
	make	price	mpg	rep78	headroom	trunk	weight	length	turn	displacement	gear_ratio	foreign
1	AMC Concord	4,099	22	3	2.5	11	2,930	186	40	121	3.58	Domestic
2	AMC Pacer	4,749	17	3	3.0	11	3,350	173	40	258	2.53	Domestic
3	AMC Spirit	3,799	22	.	3.0	12	2,640	168	35	121	3.08	Domestic
4	Buick Century	4,816	20	3	4.5	16	3,250	196	40	196	2.93	Domestic
5	Buick Electra	7,827	15	4	4.0	20	4,080	222	43	350	2.41	Domestic
6	Buick LeSabre	5,788	18	3	4.0	21	3,670	218	43	231	2.73	Domestic
7	Buick Opel	4,453	26	.	3.0	10	2,230	170	34	304	2.87	Domestic
8	Buick Regal	5,189	20	3	2.0	16	3,280	200	42	196	2.93	Domestic
9	Buick Riviera	10,372	16	3	3.5	17	3,880	207	43	231	2.93	Domestic
10	Buick Skylark	4,082	19	3	3.5	13	3,400	200	42	231	3.08	Domestic
11	Cad. Deville	11,385	14	3	4.0	20	4,330	221	44	425	2.28	Domestic
12	Cad. Eldorado	14,500	14	2	3.5	16	3,900	204	43	350	2.19	Domestic
13	Cad. Seville	15,906	21	3	3.0	13	4,290	204	45	350	2.24	Domestic
14	Chev. Chevette	3,299	29	3	2.5	9	2,110	163	34	231	2.93	Domestic
15	Chev. Impala	5,705	16	4	4.0	20	3,690	212	43	250	2.56	Domestic
16	Chev. Malibu	4,504	22	3	3.5	17	3,180	193	31	200	2.73	Domestic
17	Chev. Monte Carlo	5,104	22	2	2.0	16	3,220	200	41	200	2.73	Domestic
18	Chev. Monza	3,667	24	2	2.0	7	2,750	179	40	151	2.73	Domestic
19	Chev. Nova	3,955	19	3		13	3,430	197	43	250	2.56	Domestic
20	Dodge Colt	3,984	30	5		8	2,120	163	35	98	3.54	Domestic
21	Dodge Diplomat	4,010	18	2		17	3,600	206	46	318	2.47	Domestic
22	Dodge Magnum	5,886	16	2		17	3,600	206	46	318	2.47	Domestic
23	Dodge St. Regis	6,342	17	2		21	3,740	220	46	225	2.94	Domestic
24	Ford Fiesta	4,389	28	4		9	1,800	147	33	98	3.15	Domestic

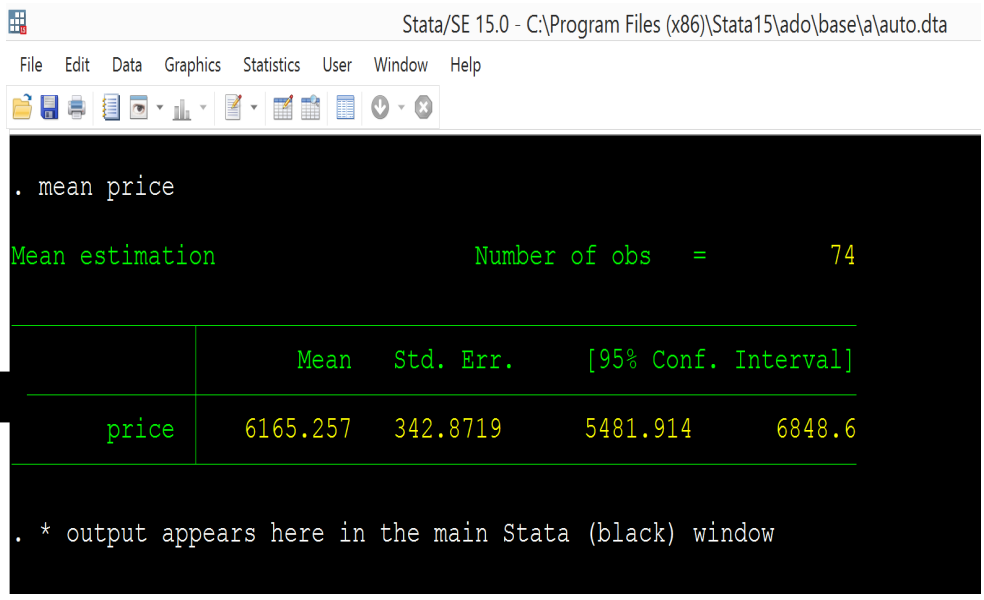


Log File (plain text)

```
. mean price

Mean estimation                   Number of obs    =         74

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
       price |   6165.257   342.8719      5481.914       6848.6
--------------------------------------------------------------

.
. ** Output appears is copied to the log file
.
```

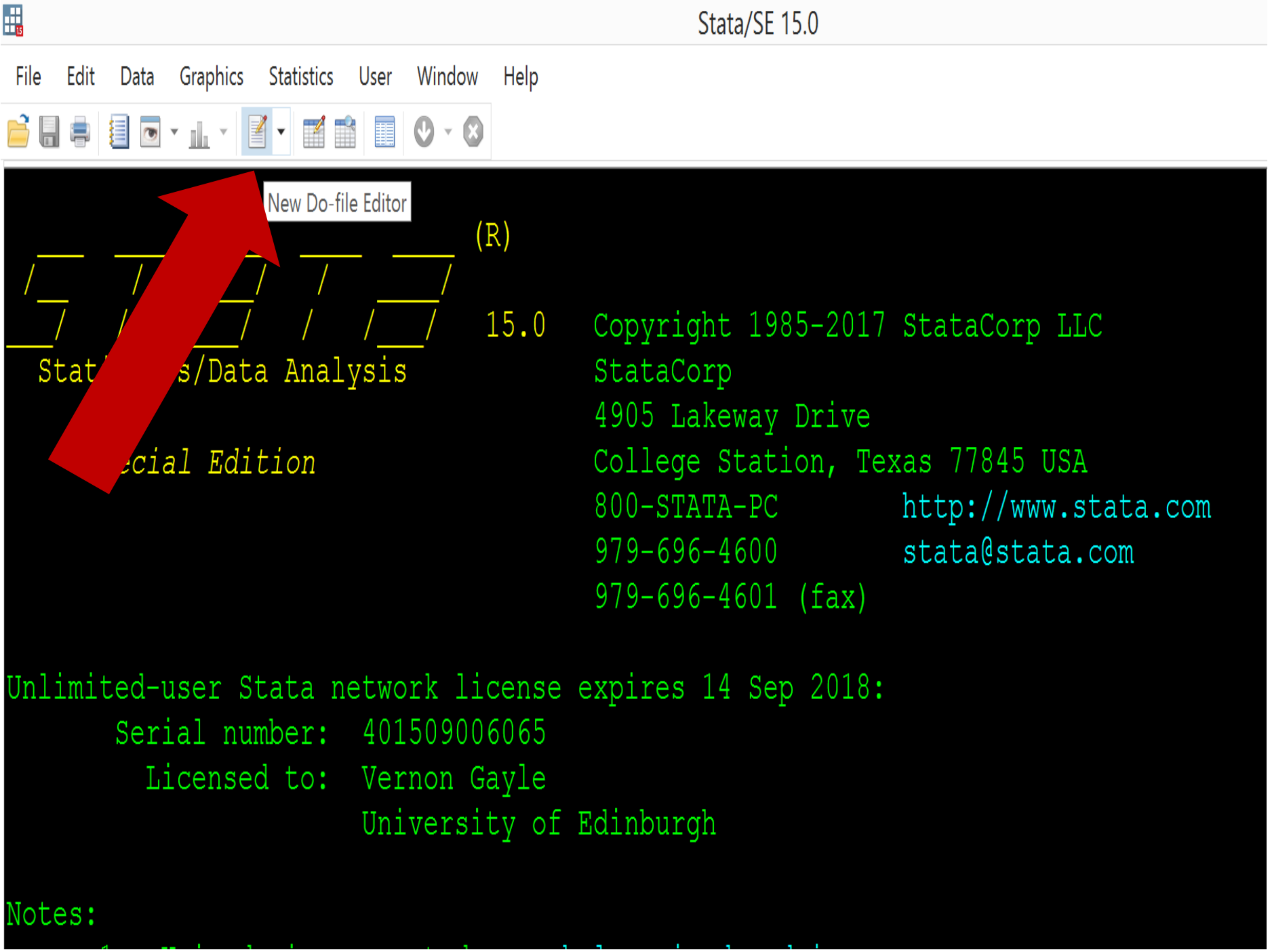


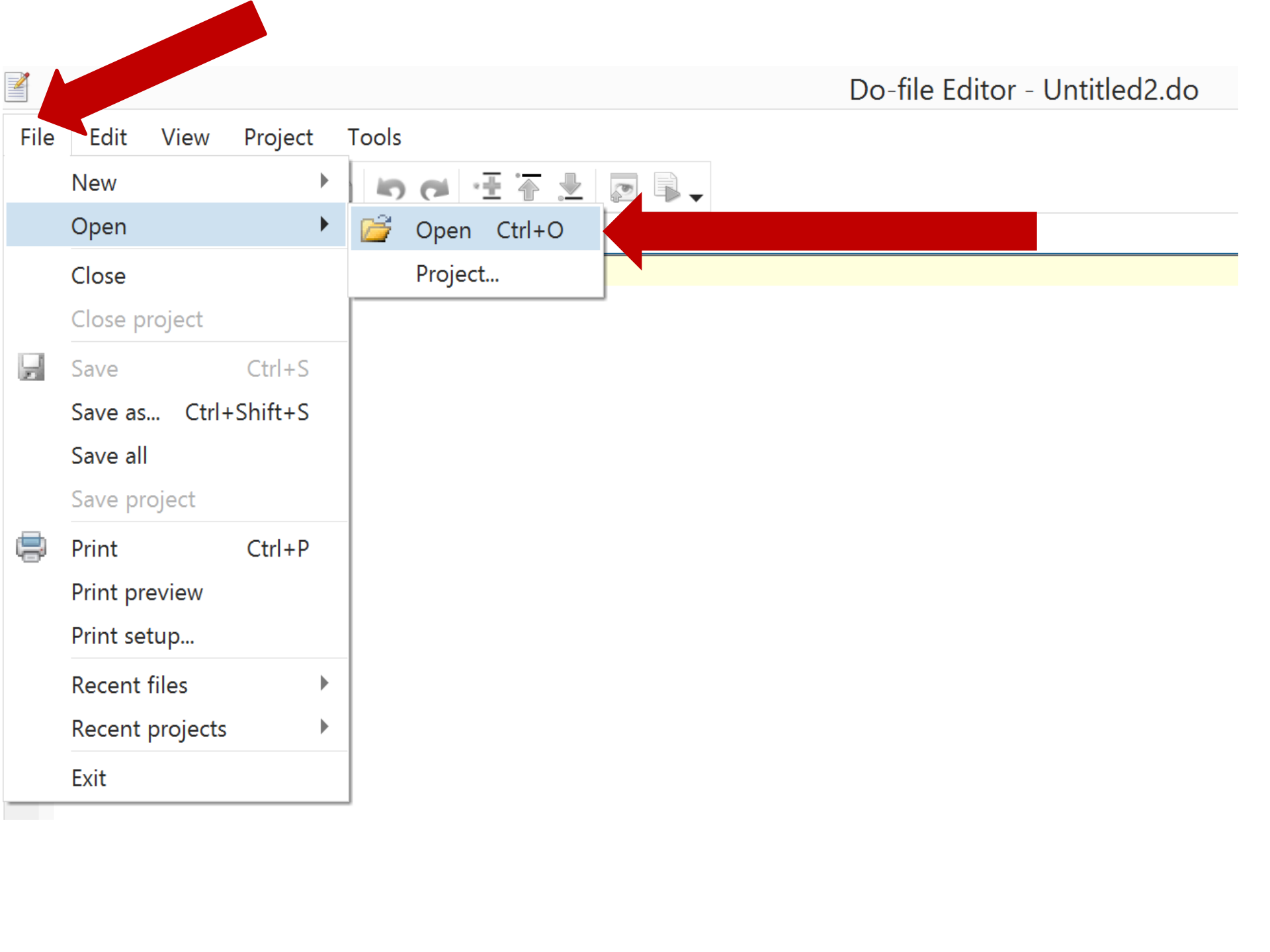





```
. mean price

Mean estimation                   Number of obs    =        74

--------------------------------------------------------------
             |       Mean   Std. Err.      [95% Conf. Interval]
-------------+------------------------------------------------
       price |   6165.257   342.8719       5481.914       6848.
--------------------------------------------------------------

.
. ** Output appears is copied to the log file
.
```

Opening a .do file…


Stata/SE 15.0
File Edit Data Graphics Statistics User Window Help
New Do-file Editor
(R)
15.0
Copyright 1985-2017 StataCorp LLC
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)
Unlimited-user Stata network license expires 14 Sep 2018:
Serial number: 401509006065
Licensed to: Vernon Gayle
University of Edinburgh
Notes:


Do-file Editor - Untitled2.do
File
Edit
View
Project
Tools
New
Open
Close
Close project
Save Ctrl+S
Save as... Ctrl+Shift+S
Save all
Save project
Print Ctrl+P
Print preview
Print setup...
Recent files
Recent projects
Exit
Open Ctrl+O
Project...

Stata/SE 15.0

File Edit Data Graphics Statistics User Window Help

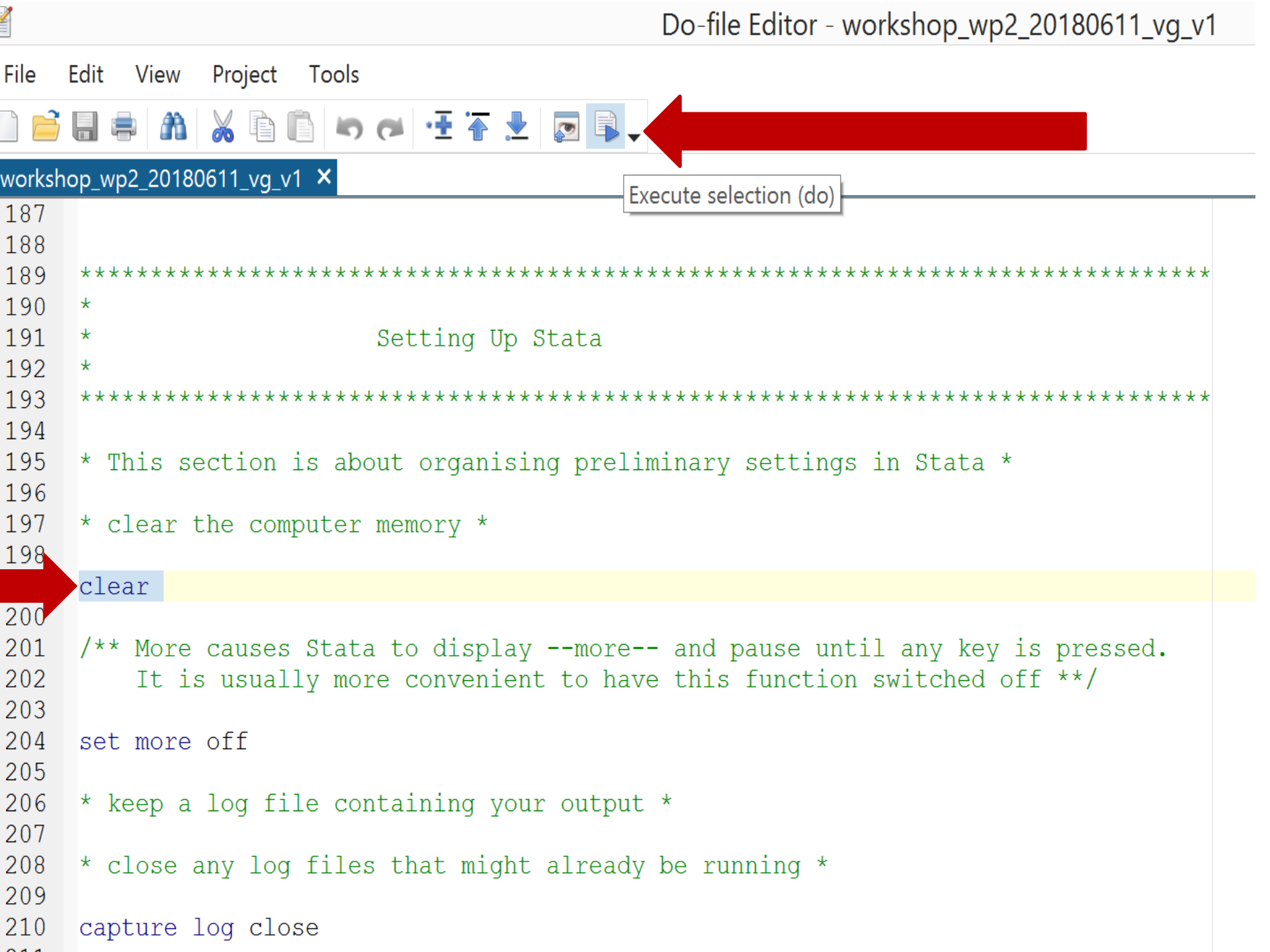
```
                                      (R)
  ___  ____  ____  ____  ____
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   15.0   Copyri
  Statistics/Data Analysis             StataC
                                       4905 I
     Special Edition                   Colleg
                                       800-ST
                                       979-69
                                       979-69

Unlimited-user Stata network license expires
       Serial number:  401509006065
         Licensed to:  Vernon Gayle
                       University of Edinbur

Notes:
      1.  Unicode is supported; see help uni
      2.  Maximum number of variables is set
```

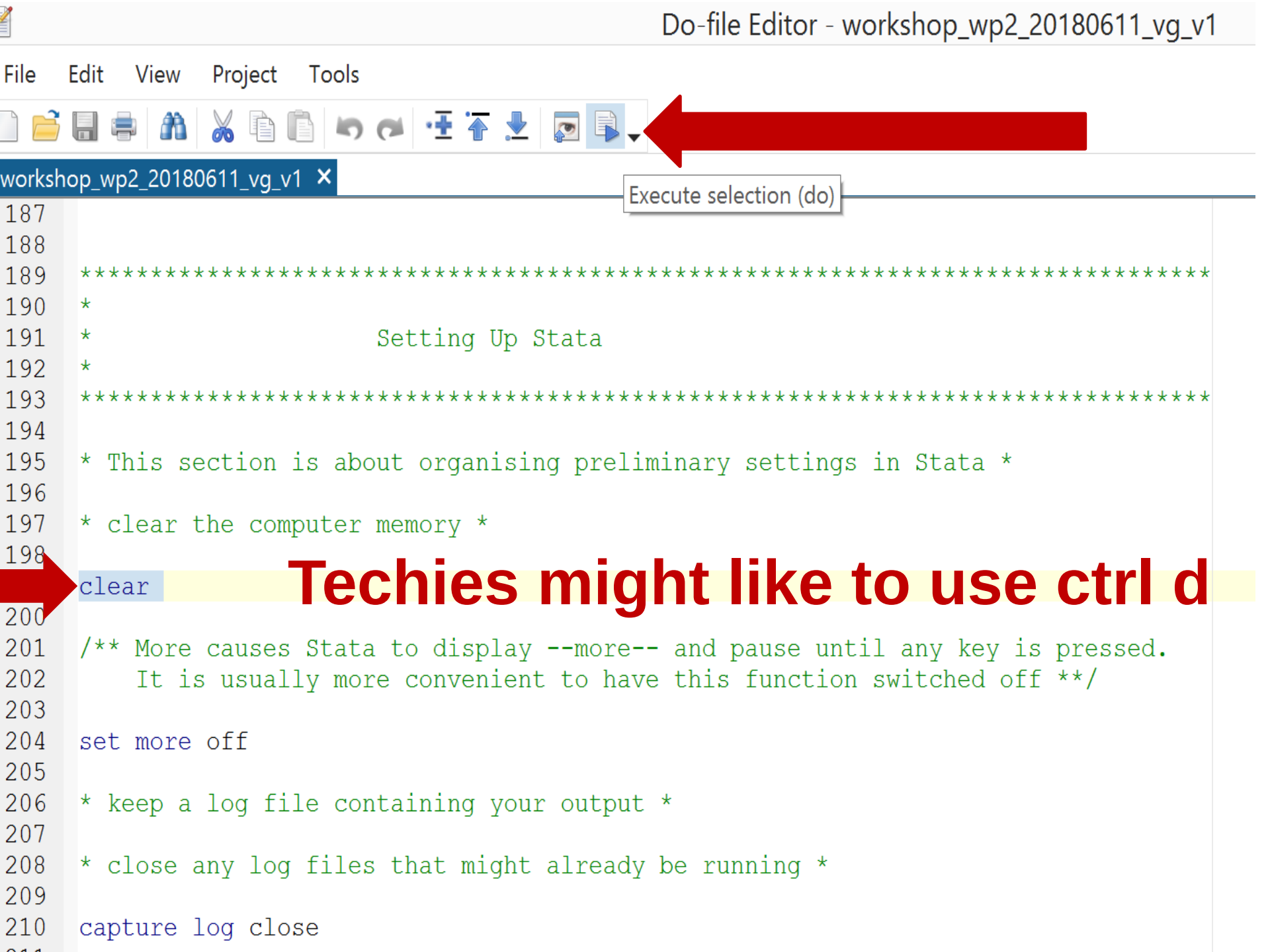
Do-file Editor - workshop_wp2_20180611_vg_v1

File Edit View Project Tools

workshop_wp2_20180611_vg_v1 X

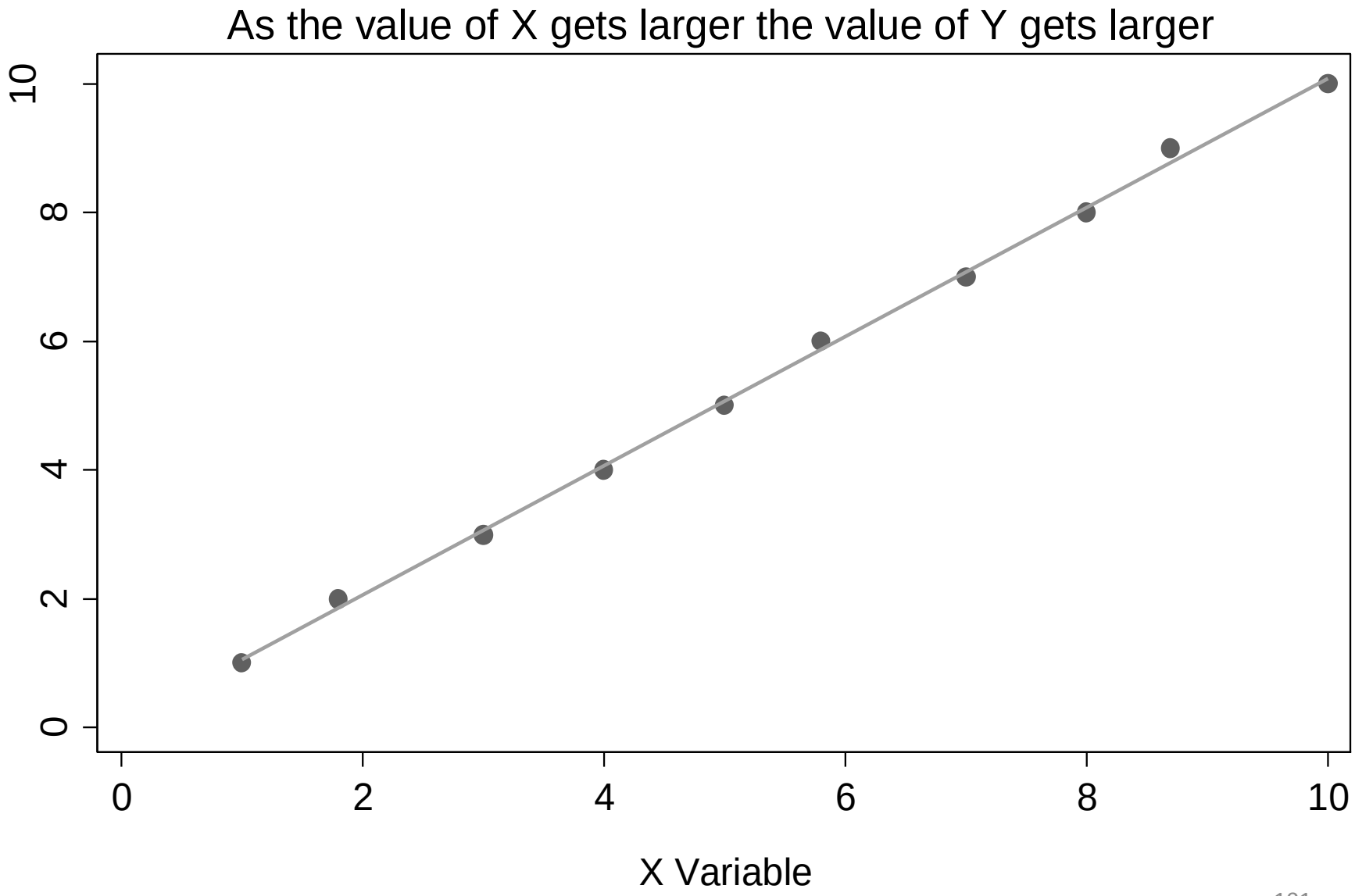
```
1  STOP
2
3  /**
4
5  AQMEN Applied Quantitative Methods Network /
6
7
8  An Introduction to Statistical Concepts for Data Analysis
9
10 Stata Workshop (July 2018)
11
12 A two day hands-on workshop led by Professor Vernon Gayle,
13 University of Edinburgh.
14
15         ********************************************
16         * IT IS IMPORTANT THAT YOU READ THIS HANDOUT *
17         * AND FOLLOW THE STATA.DO FILE LINE BY LINE! *
18         ********************************************
19
20 Topics:
21
22 The course introduces participants to fundamental concepts in
23 rudimentary data analysis techniques and how to interpret resu
24 These skills are critical for the successful analysis of data.
25
```


Do-file Editor - workshop_wp2_20180611_vg_v1
File Edit View Project Tools
workshop_wp2_20180611_vg_v1
Execute selection (do)
187
188
189 **
190 *
191 * Setting Up Stata
192 *
193 **
194
195 * This section is about organising preliminary settings in Stata *
196
197 * clear the computer memory *
198
clear
200
201 /** More causes Stata to display --more-- and pause until any key is pressed.
202 It is usually more convenient to have this function switched off **/
203
204 set more off
205
206 * keep a log file containing your output *
207
208 * close any log files that might already be running *
209
210 capture log close

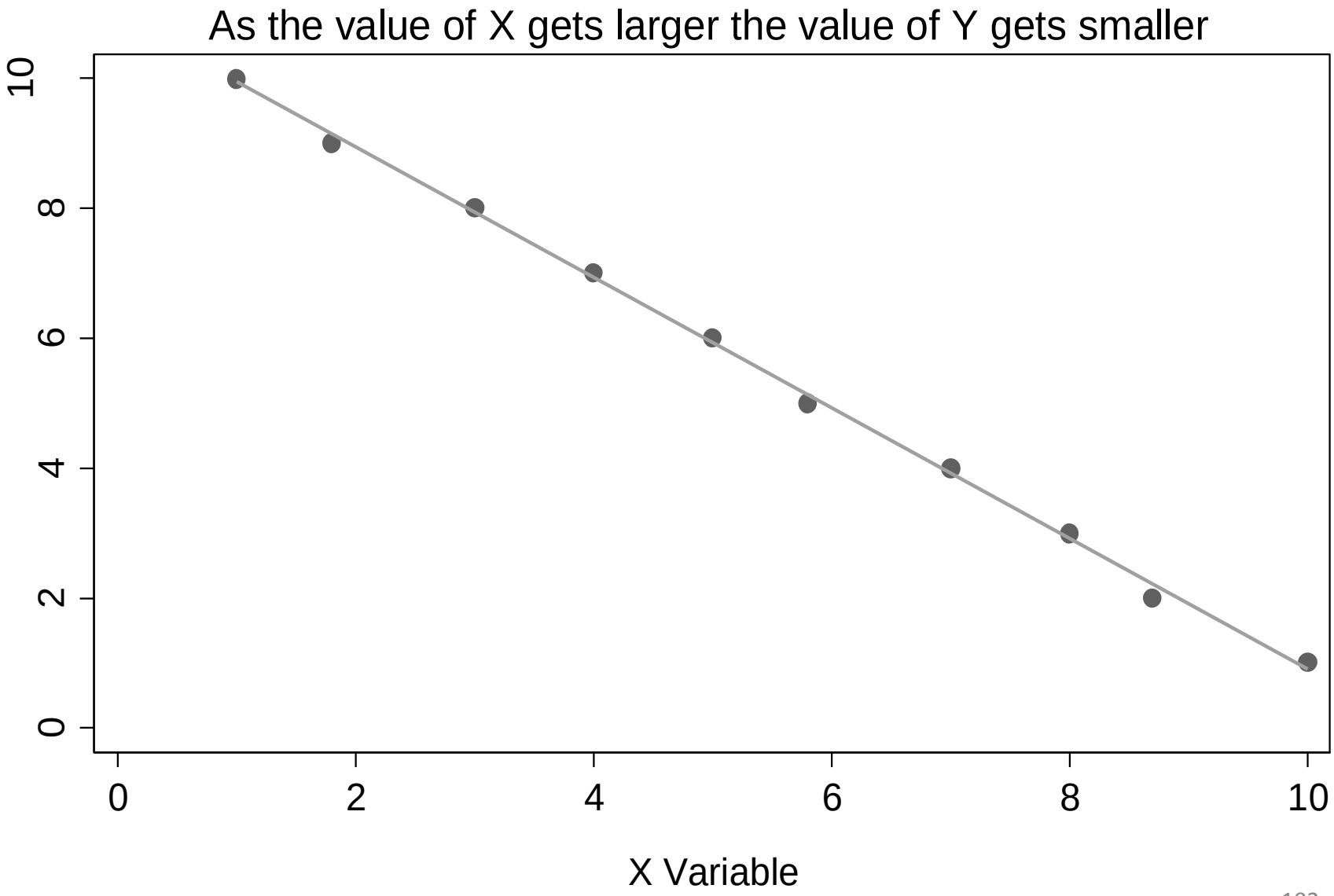

Do-file Editor - workshop_wp2_20180611_vg_v1
File Edit View Project Tools
workshop_wp2_20180611_vg_v1
Execute selection (do)
187
188
189 **
190 *
191 * Setting Up Stata
192 *
193 **
194
195 * This section is about organising preliminary settings in Stata *
196
197 * clear the computer memory *
198
Techies might like to use ctrl d
clear
200
201 /** More causes Stata to display --more-- and pause until any key is pressed.
202 It is usually more convenient to have this function switched off **/
203
204 set more off
205
206 * keep a log file containing your output *
207
208 * close any log files that might already be running *
209
210 capture log close

Part 6 Correlations

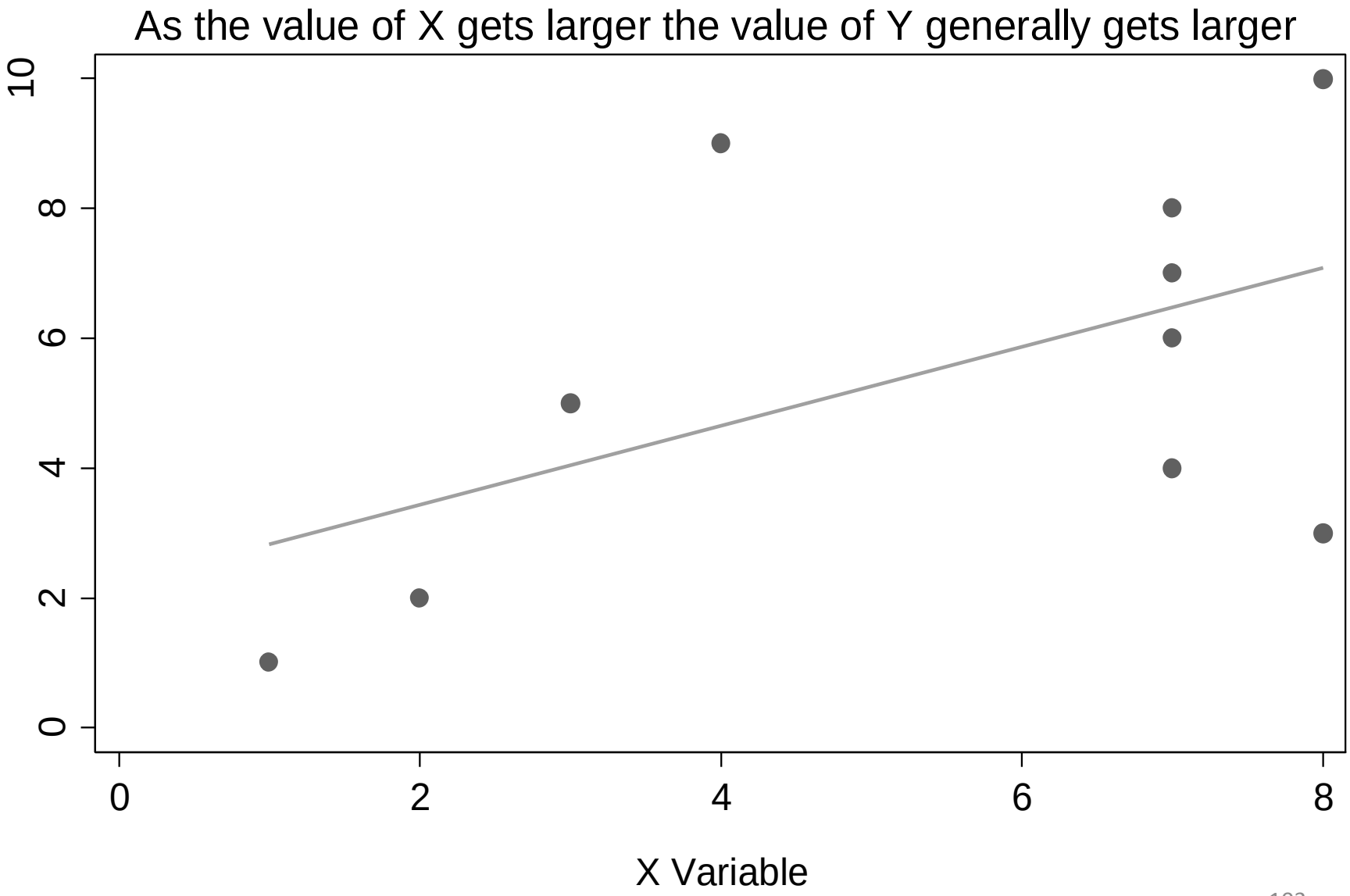
Positive Relationship



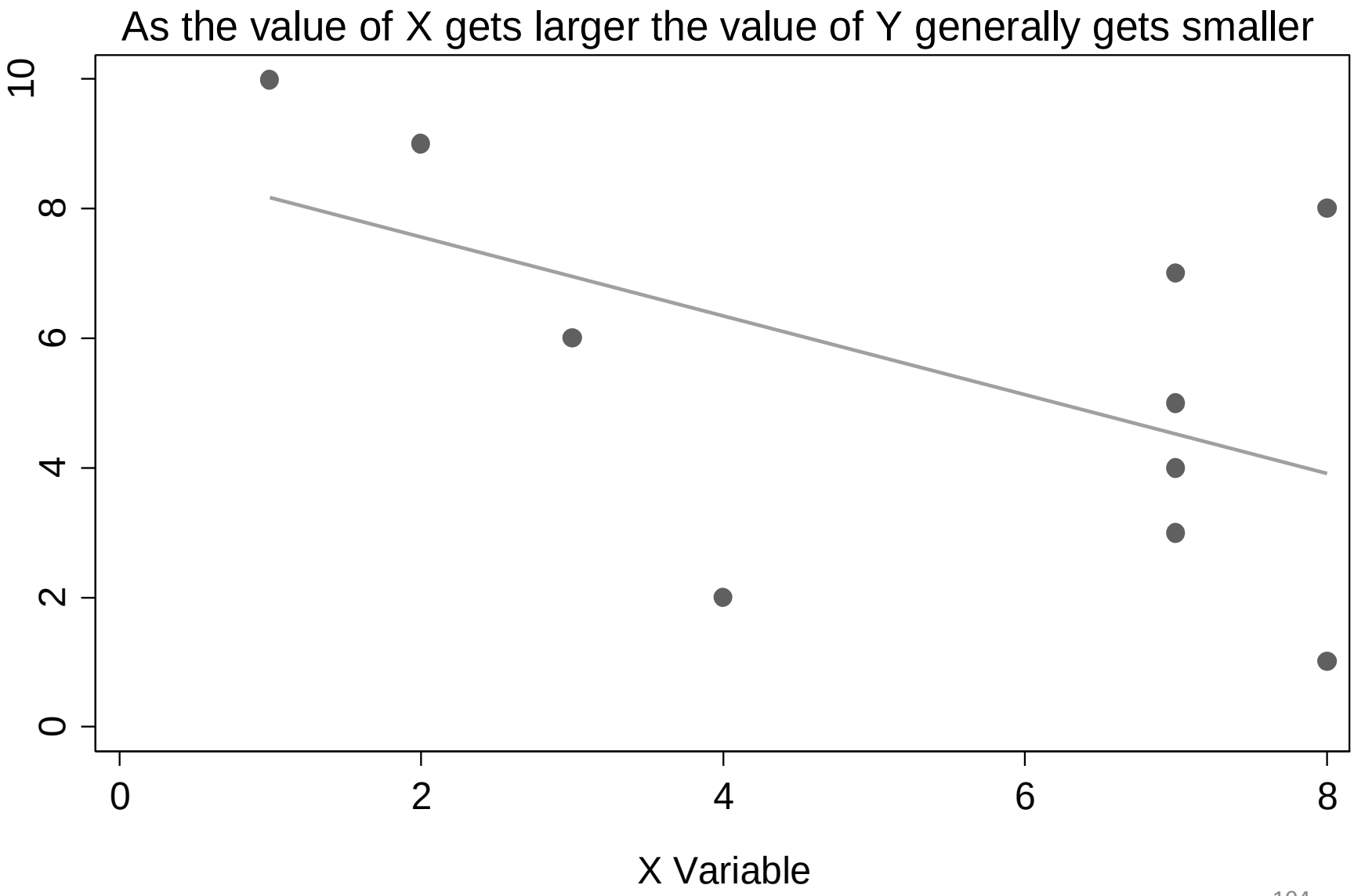
Negative Relationship



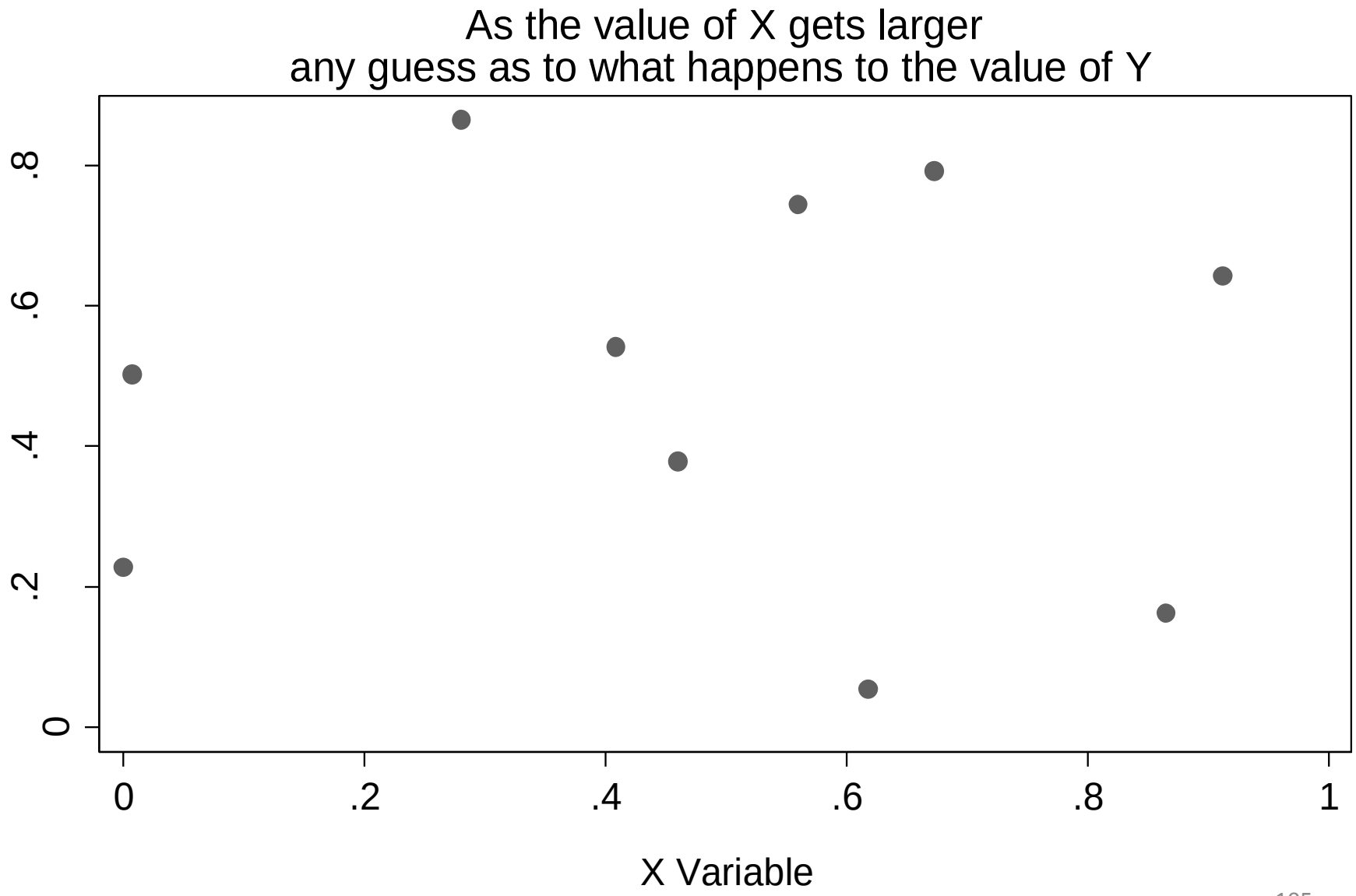
Weaker Positive Relationship



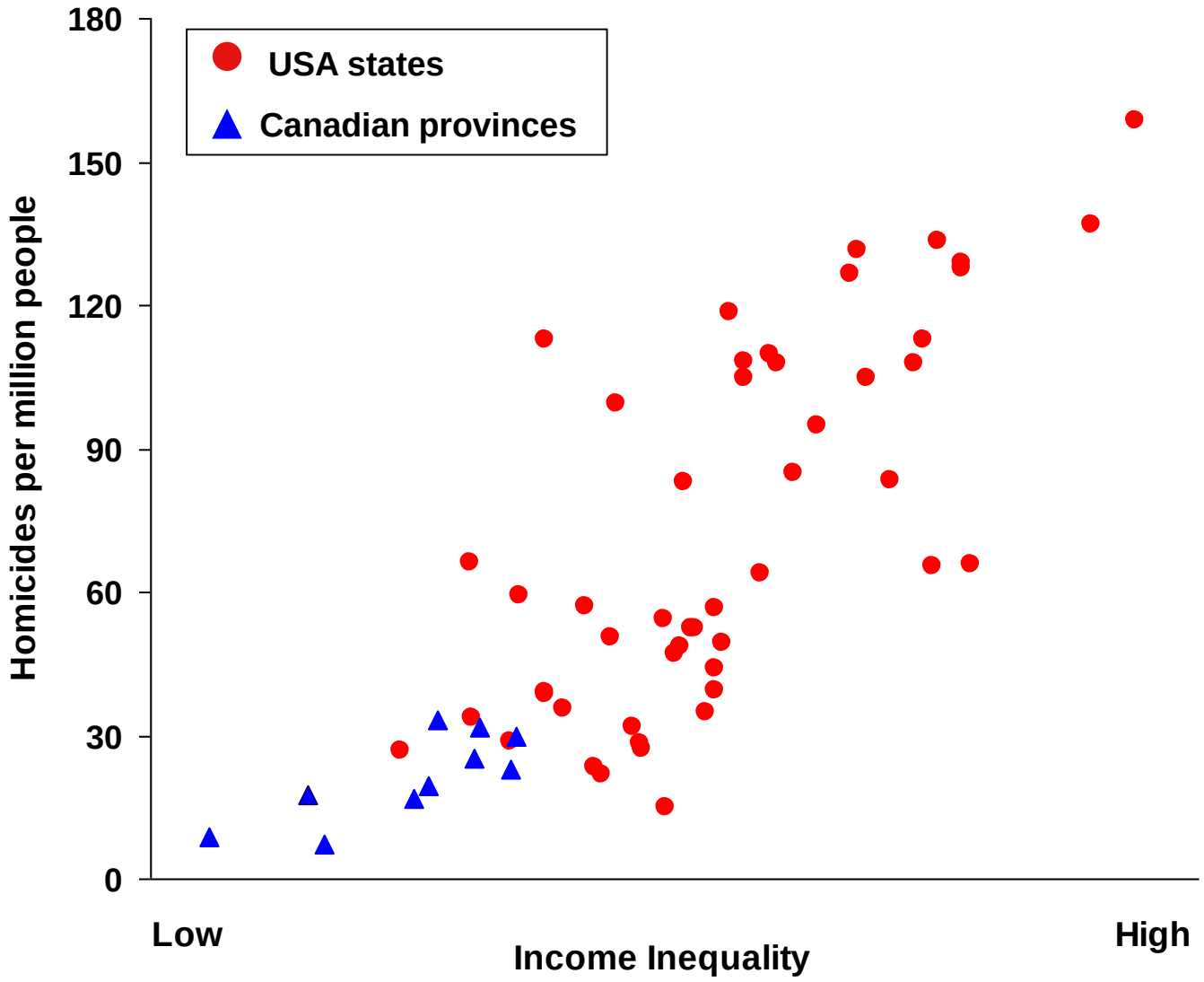
Weaker Negative Relationship



No (linear) Relationship



Homicide rates are higher in more unequal US states and Canadian provinces



Daly M, Wilson M, Vasdev S. Income inequality and homicide rates in Canada and the United States. *Can J Crim* 2001; 43: 219-36.

Vocabulary

Positive (+)

Negative (-)

Pearson's r

+1 **Strong**

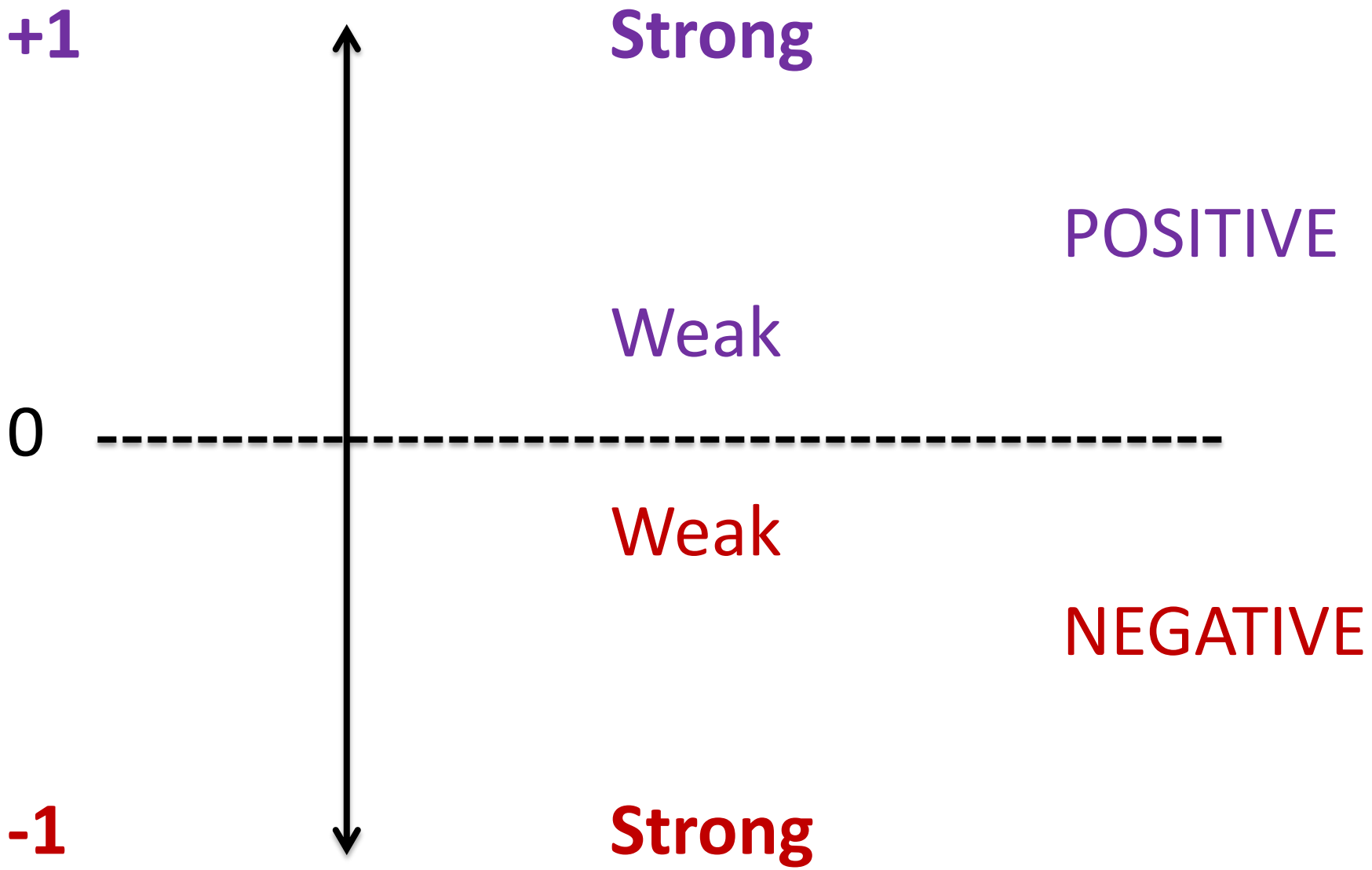
Weak

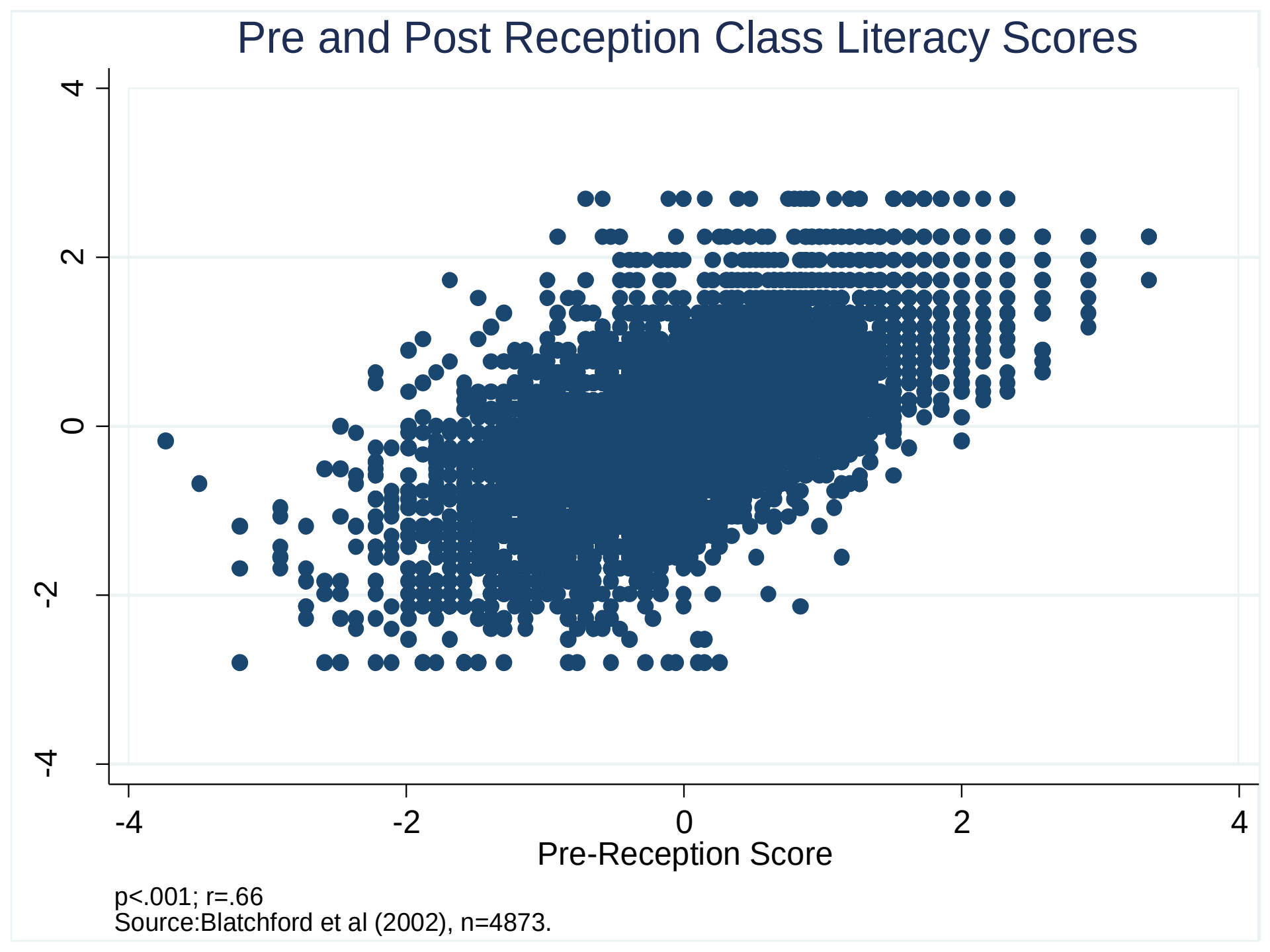
0

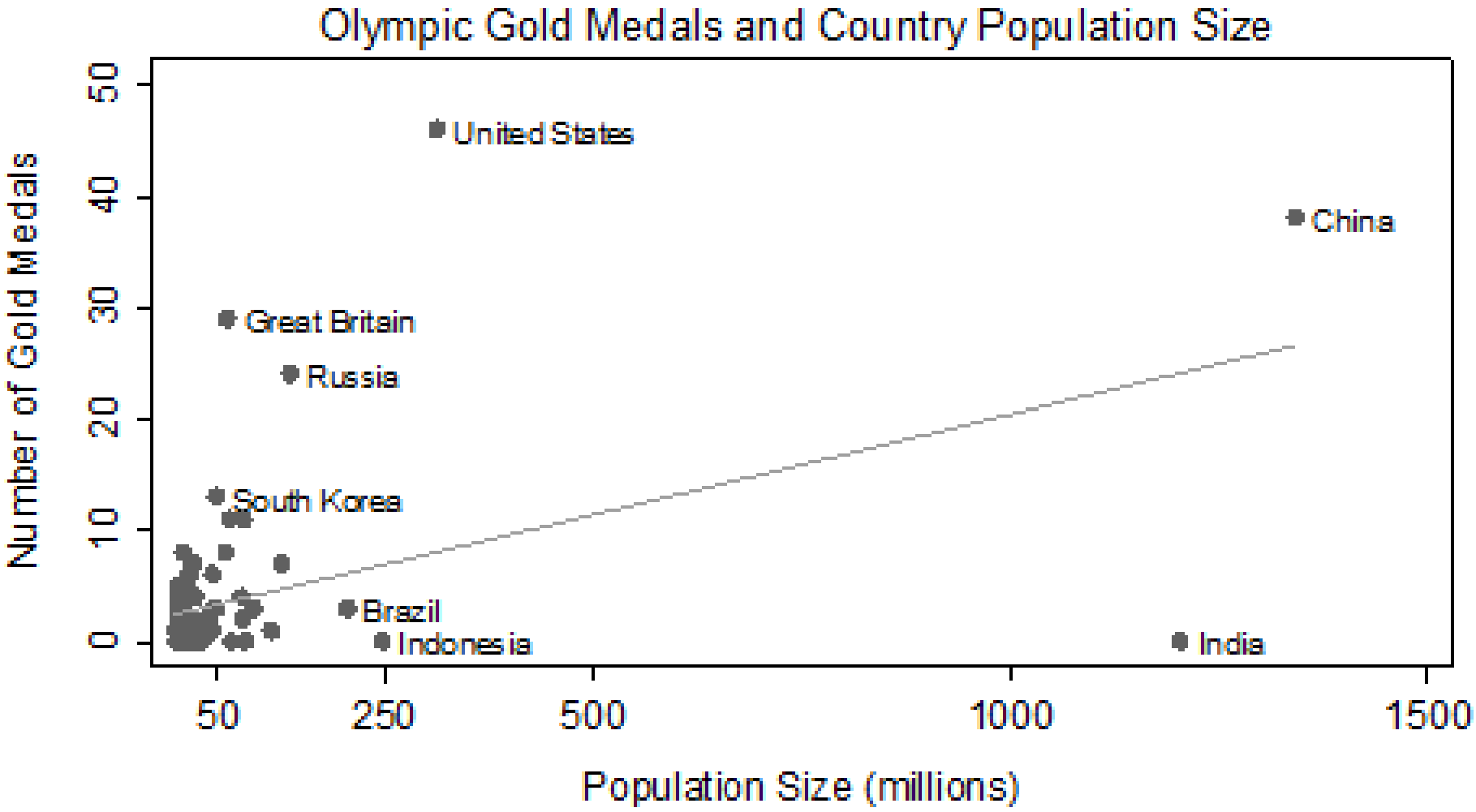
Weak

-1 **Strong**

Pearson's r



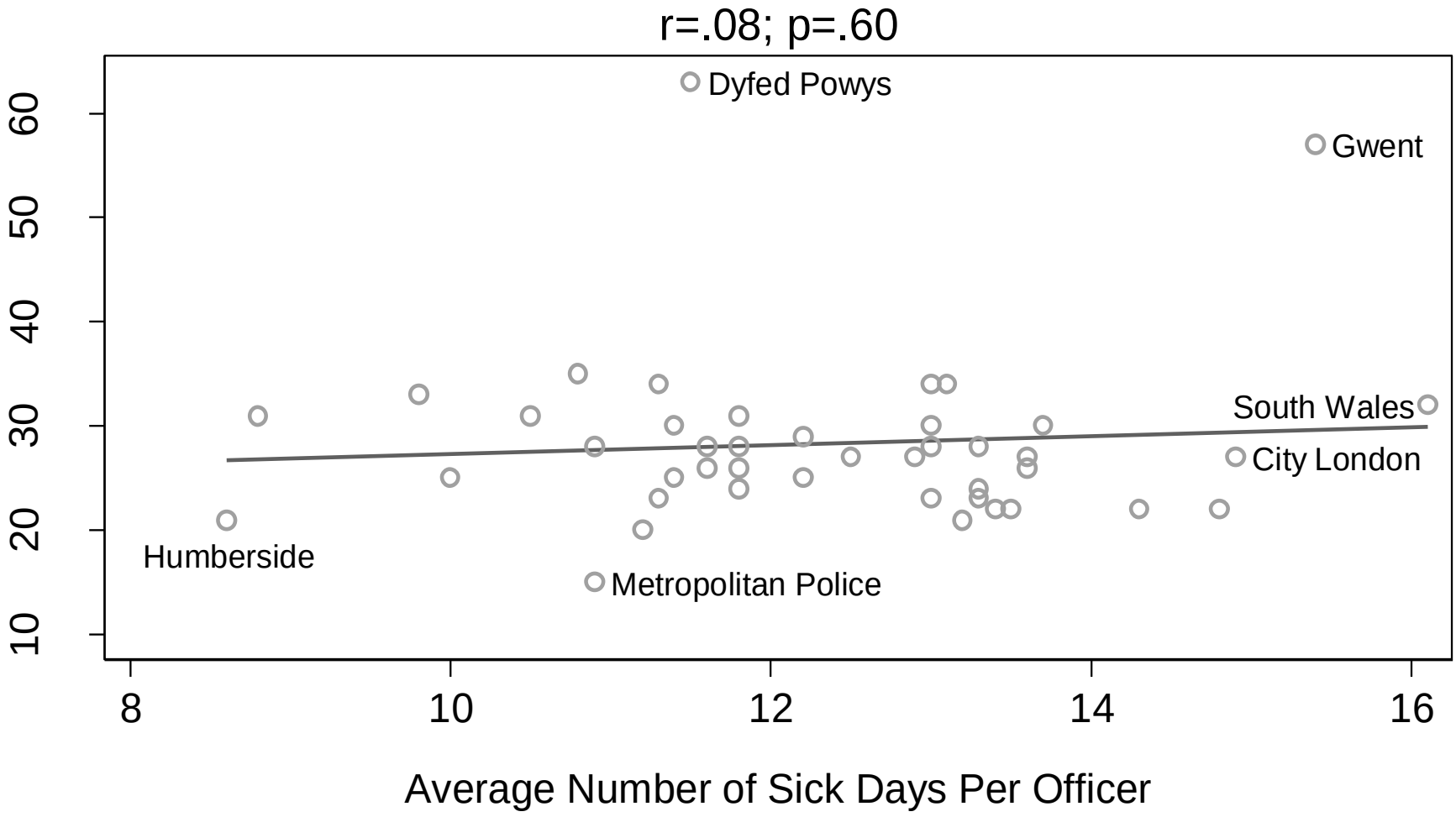

Pre and Post Reception Class Literacy Scores
4
2
0
-2
-4
-4
-2
0
2
4
Pre-Reception Score
p<.001; r=.66
Source:Blatchford et al (2002), n=4873.



p<.01; r=.46; rsquared=.22;
Data source: Timothy Lethbridge's Blog

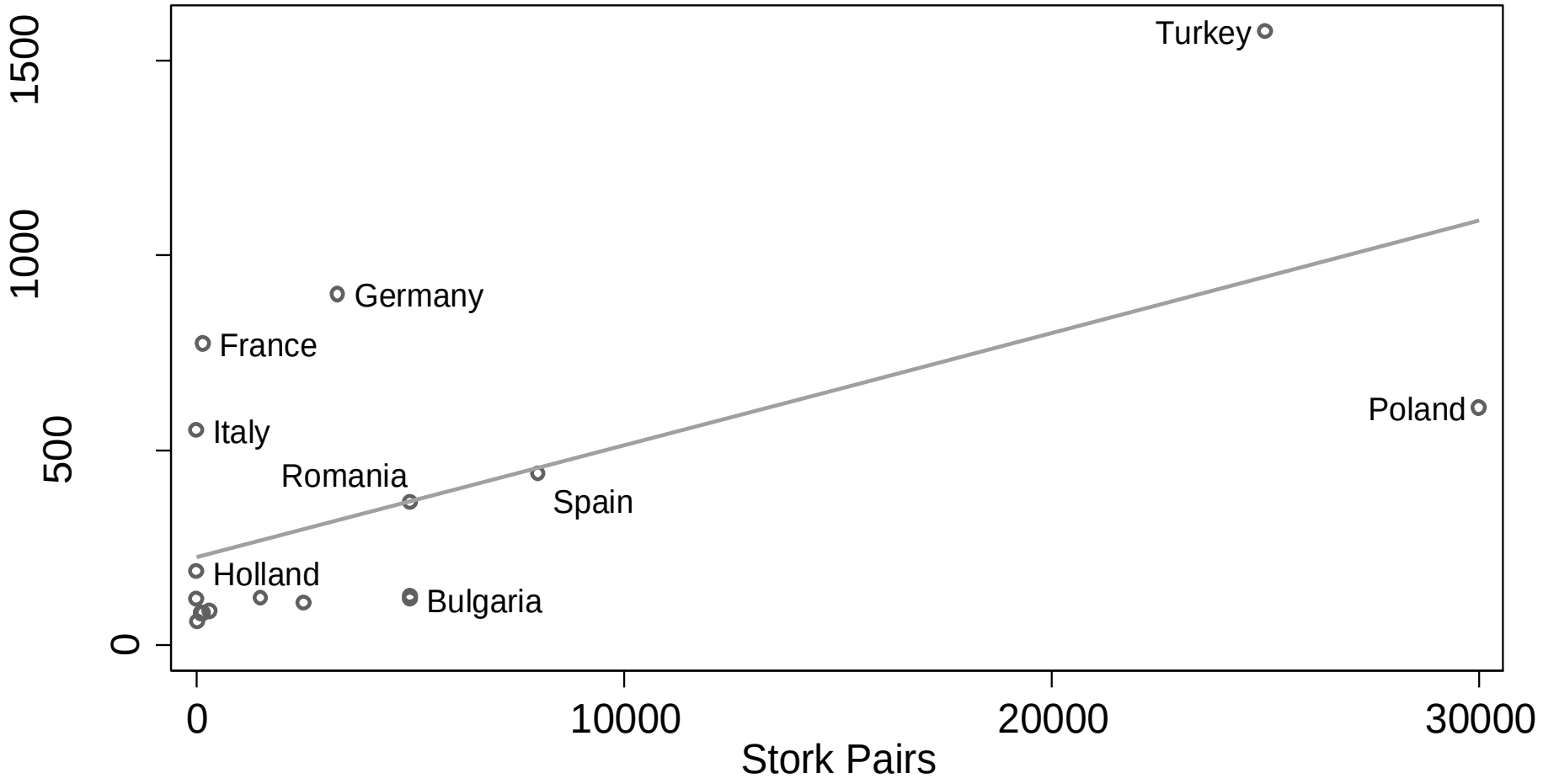
Crime Detection and Staff Sickness

Police Forces (England and Wales)



Observer police performance tables
http://observer.theguardian.com/secondterm/table/0,8173,609883,00.html

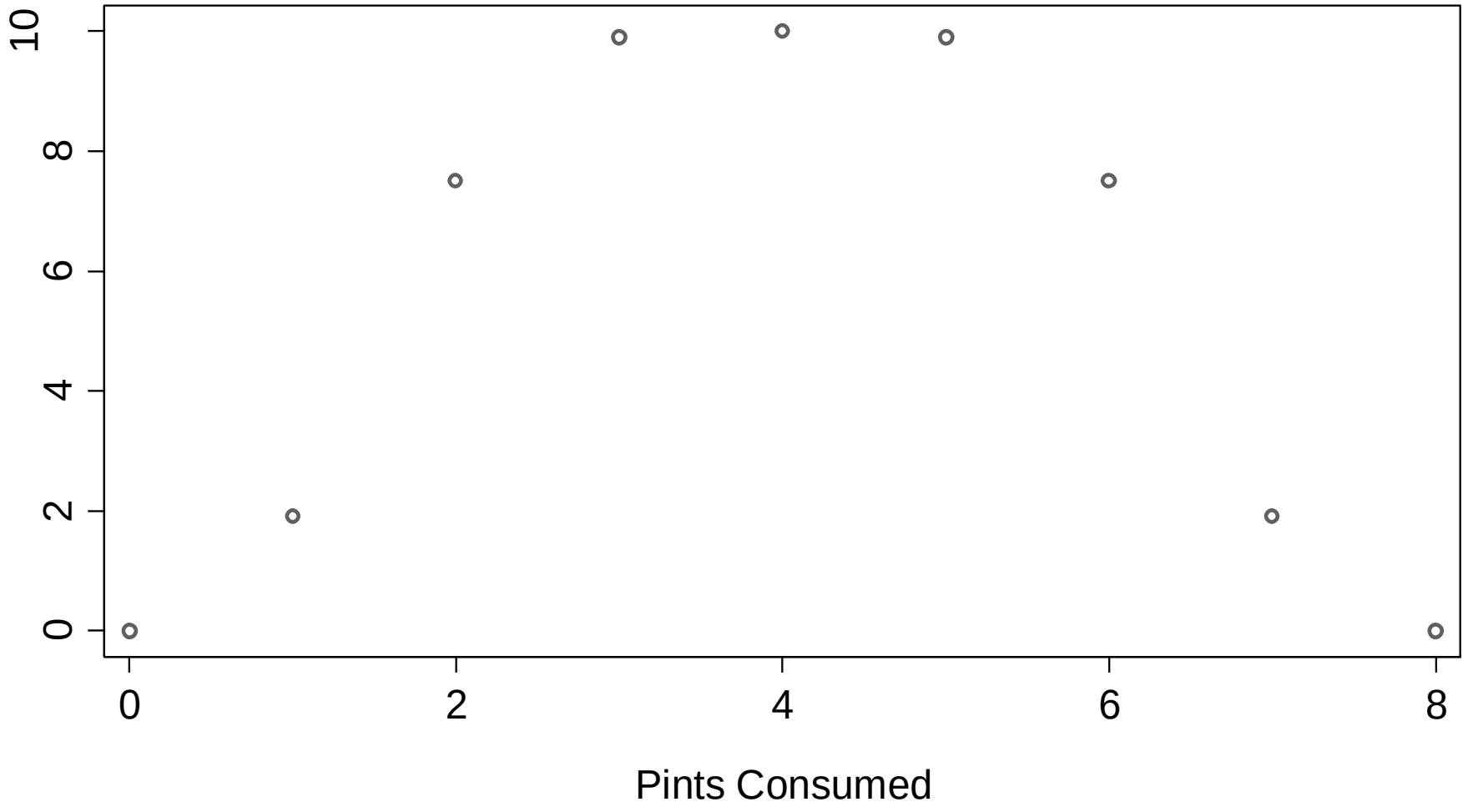
Do Storks Deliver Babies?



p=.008; r=.62

Mathews, R. 'Storks Deliver Babies (p=0.008)'
Teaching Statistics. Volume 22, Number 2, Summer 2000

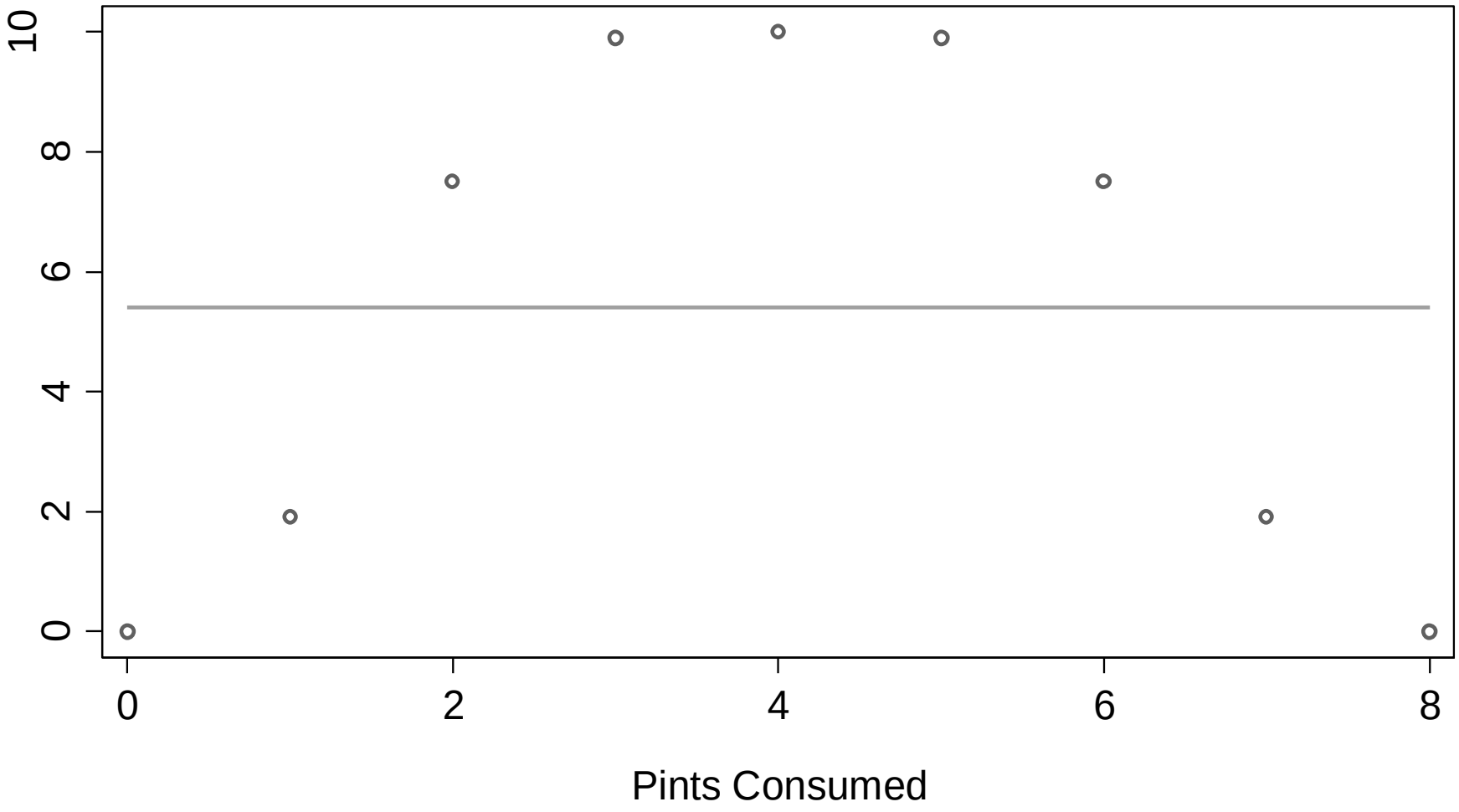
Beer Drinking and Sexual Performance in Young Males



p=1.00; r=.00

Journal of Unethical Studies

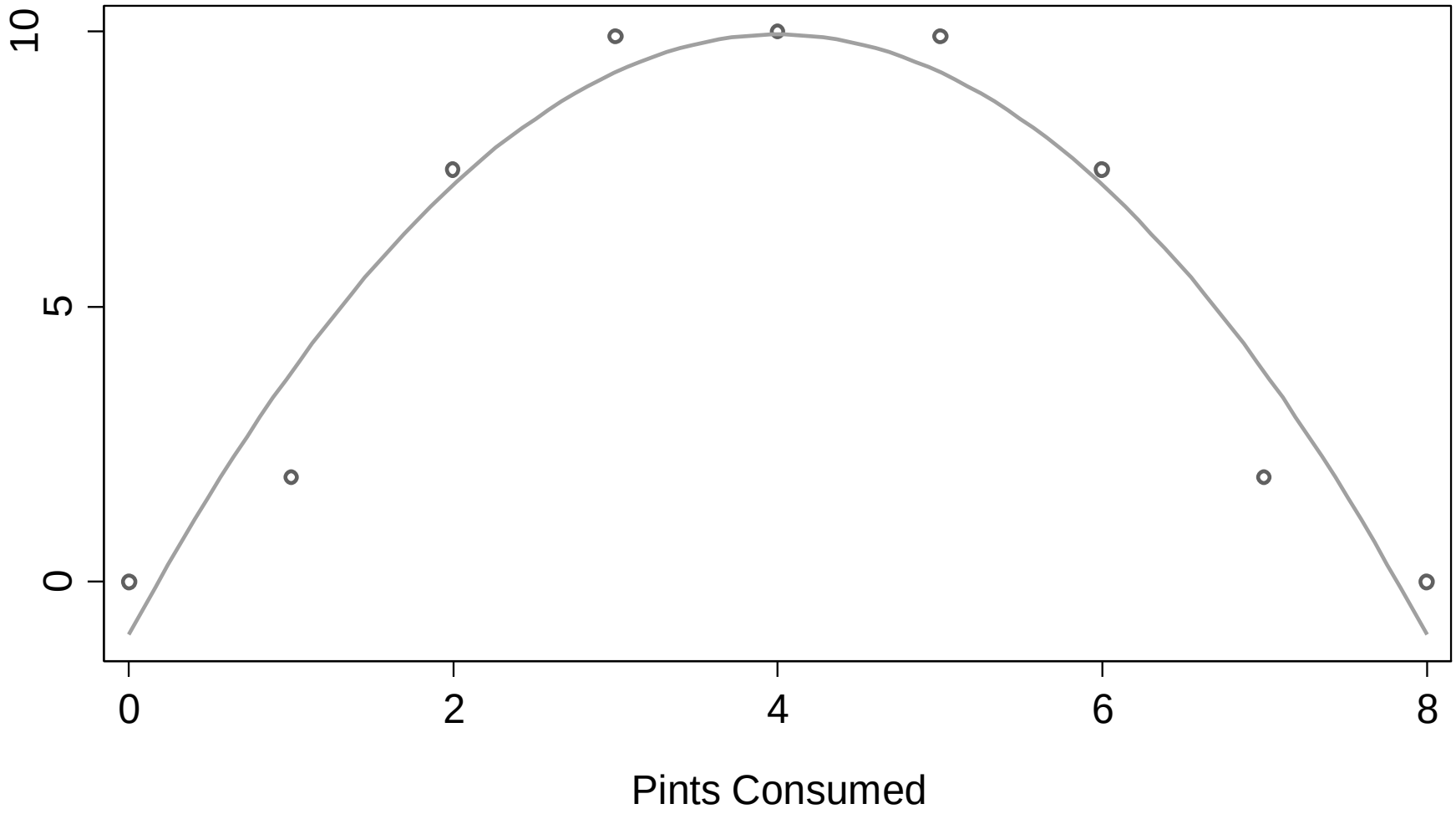
Beer Drinking and Sexual Performance in Young Males



p=1.00; r=.00

Journal of Unethical Studies

Beer Drinking and Sexual Performance in Young Males



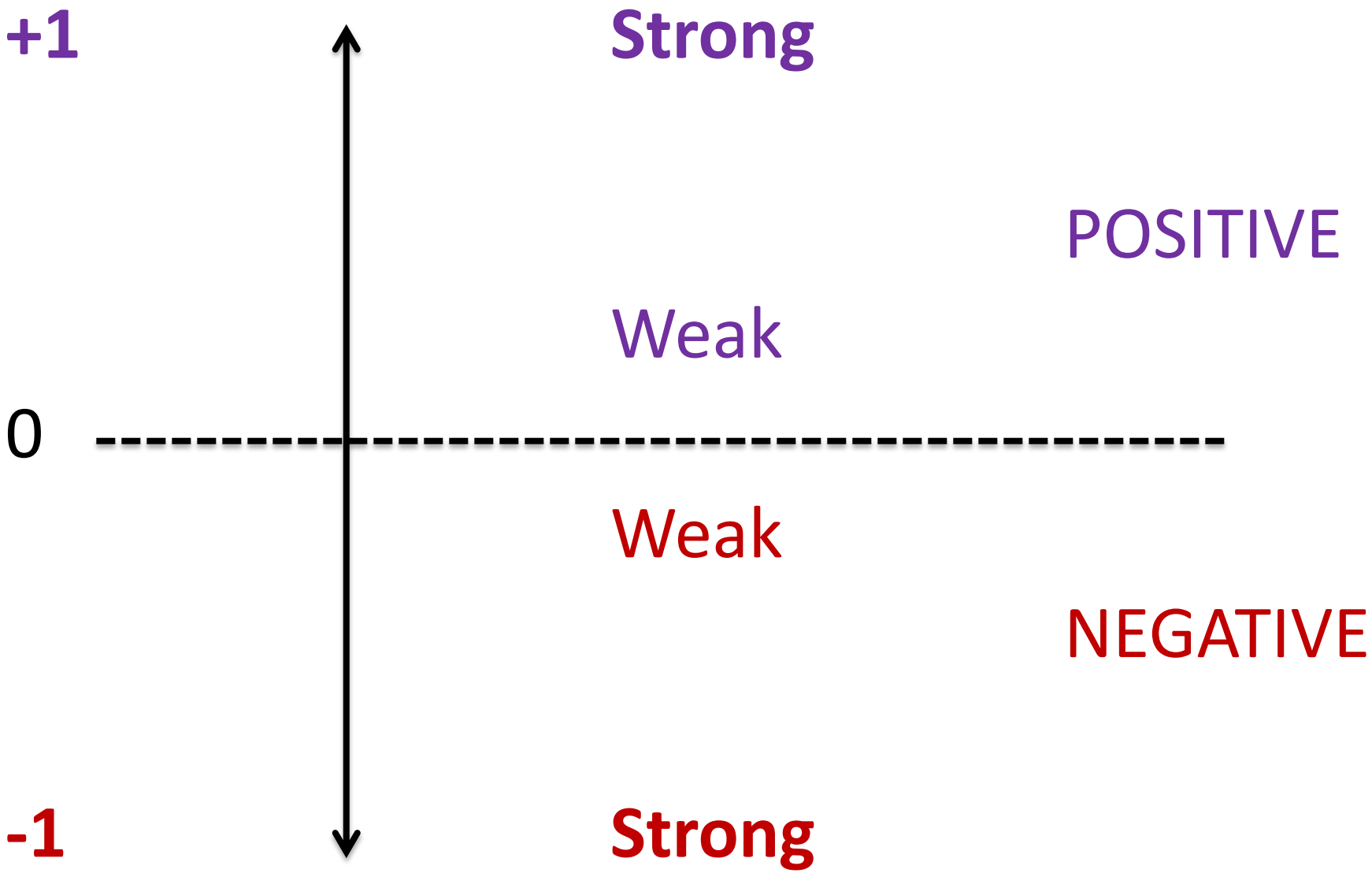
p=1.00; r=.00

Journal of Unethical Studies

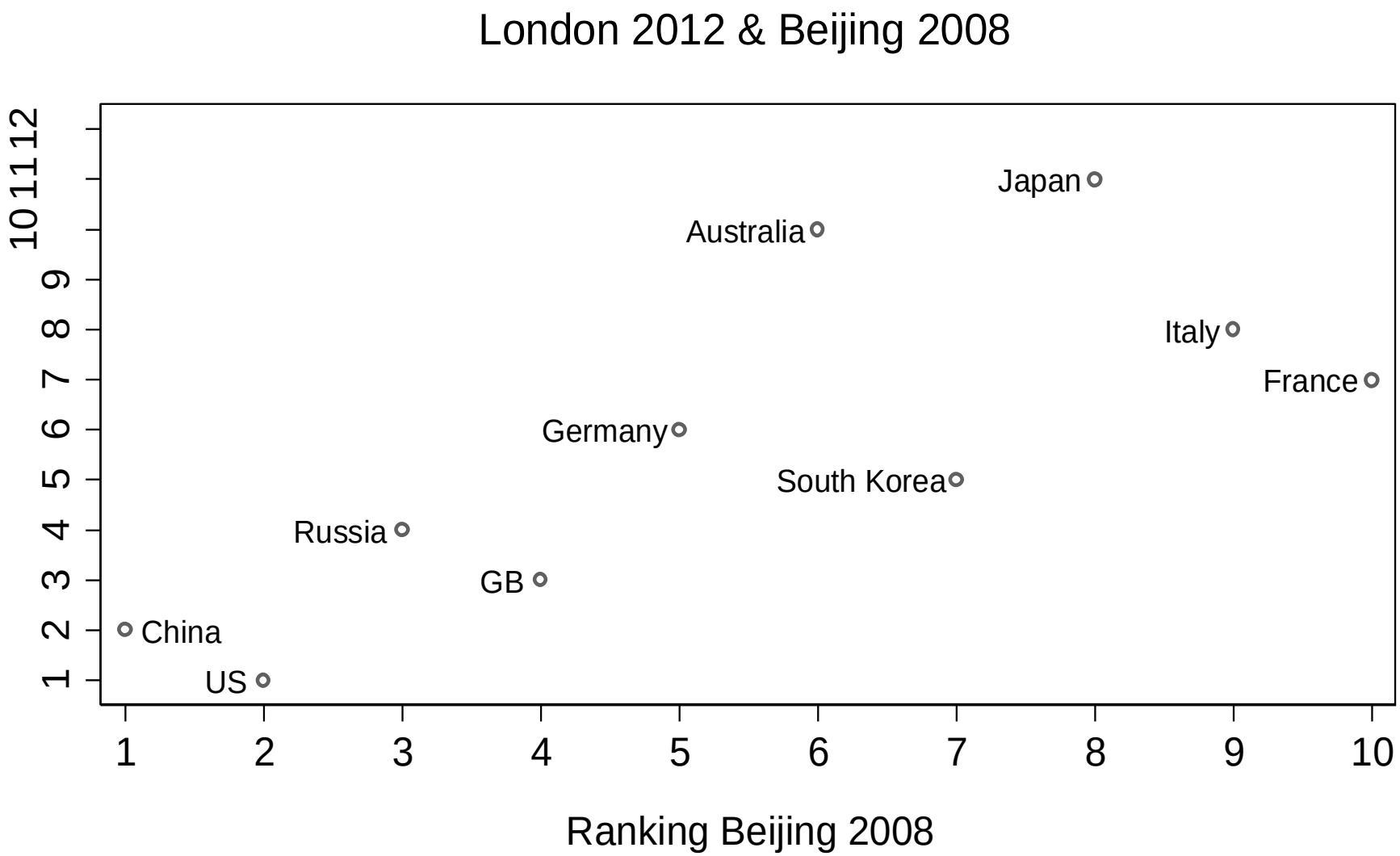
Terminology

- Positive or Negative
- Weak of Strong
- Linear relationship
- No linear relationship
- Bivariate regression line (line of best fit)
- Spurious correlation
- Non-linear relationship

Spearman's Rho ρ



Olympic Medals Table Rank



p<.01; rho=.81. Source: BBC Sport

Coefficient of Determination

- r^2 takes of values between 0 and 1
- Proportion of variation in Y explained by X

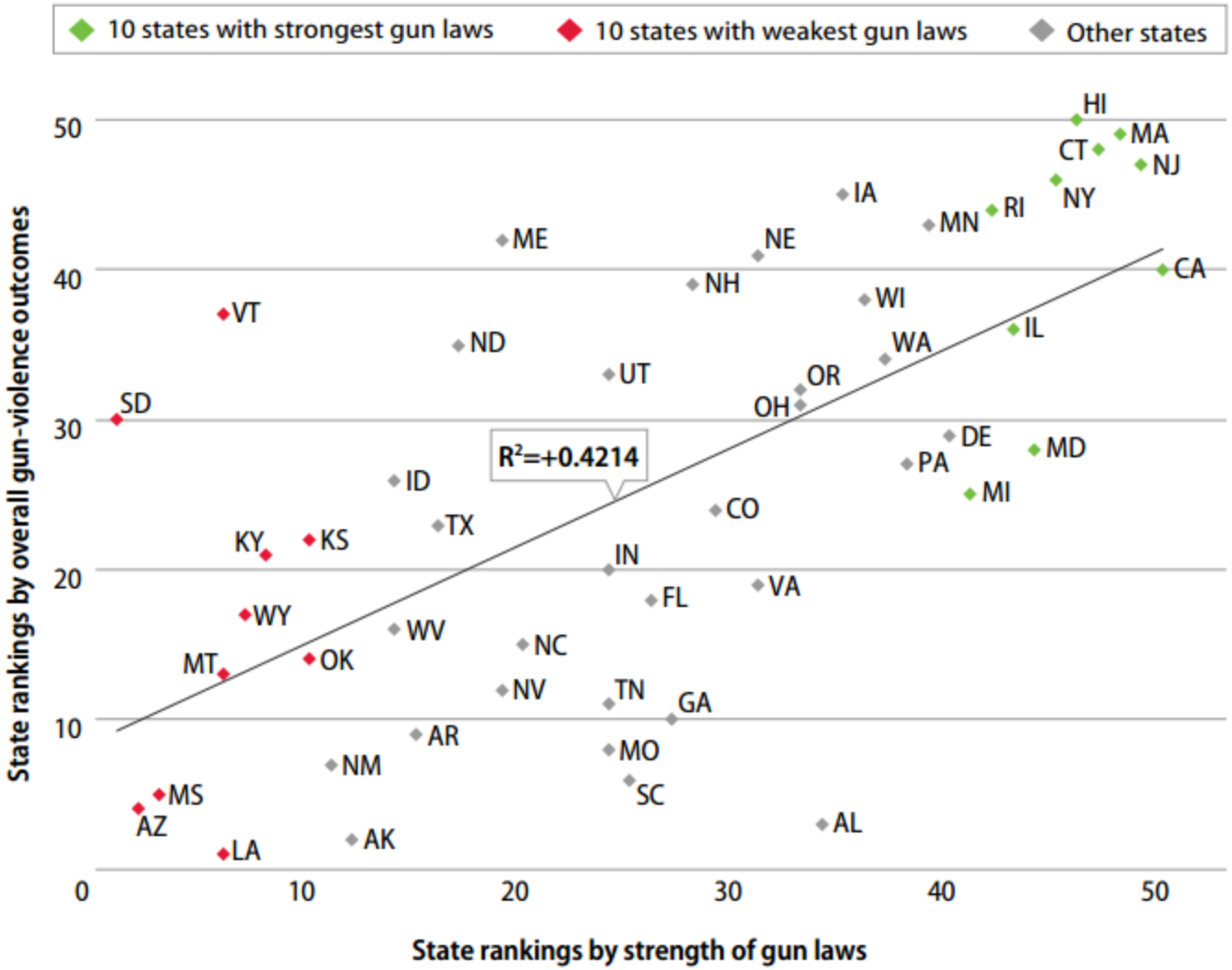
Therefore…

A strong positive correlation r =.9 then r^2 =.81

A strong negative correlation r = -.9 then r^2 =.81

FIGURE 3

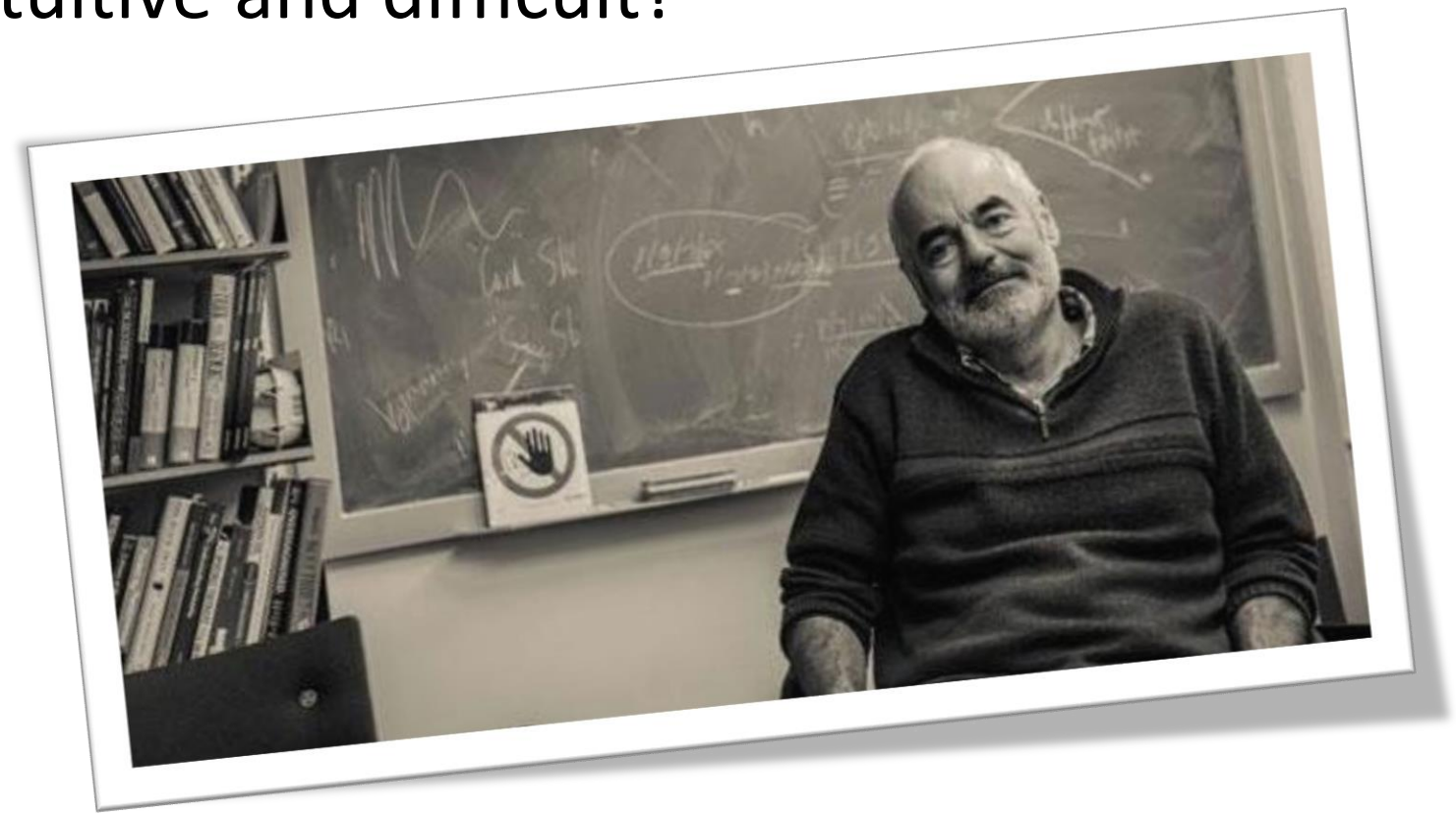
Correlation between state gun laws and gun-violence outcomes



Source: Center for American Progress analysis based on data from Centers for Disease Control and Prevention, Federal Bureau of Investigation, Mayors Against Illegal Guns, and Law Center to Prevent Gun Violence.

Part 7 Probability

Why do so many people find probability theory so unintuitive and difficult?

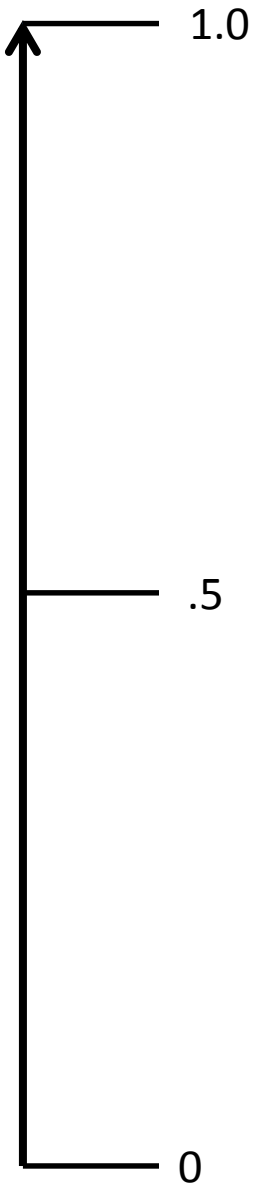


After years of careful study, I have finally found it's because probability is unintuitive and difficult.

http://www.wired.co.uk/magazine/archive/2011/09/ideas-bank/david-spiegelhalter-probability-is-likely-to-confuse-people

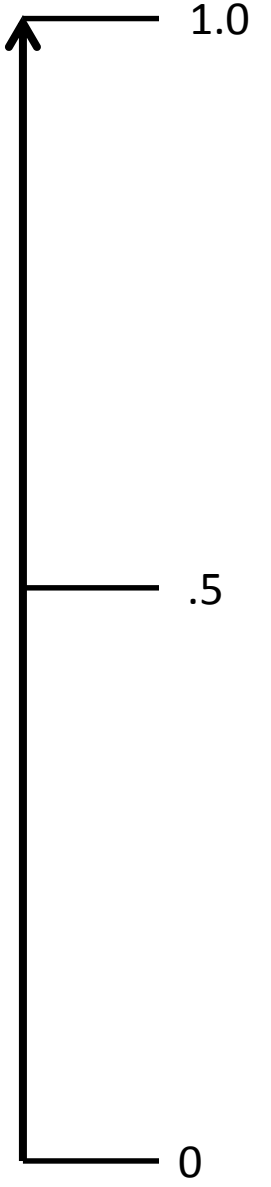
“It is clear talking to people that I am not the only one who finds probability difficult. I find probability problems very difficult indeed. I have to sit and think carefully. Usually because there are two or three different ways of doing it, and different ways of approaching the problem. Which is nice, but it is quite difficult.”

Lecture to South African Statisticians



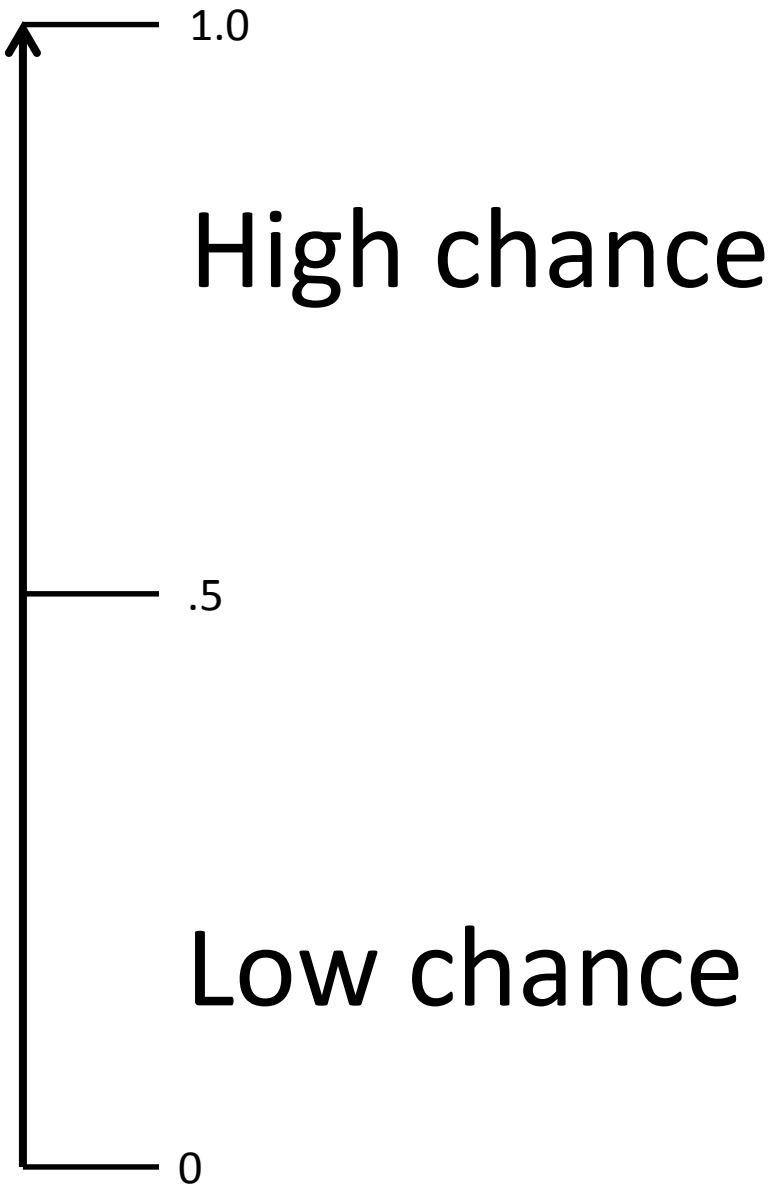
Probabilities take on values between 0 and 1

Denoted a “p”



Event will definitely occur

Event will definitely **not** occur


1.0
High chance
.5
Low chance
0

p=.50 even chance

Guess the Probability?

1. Drawing an ace from a single standard pack?

2. Rolling a three with a (fair) single die?

3. Probability of tossing two heads in a row with a 50p coin?

Guess the Probability?

1. Drawing an ace from a single standard pack

   p=.08 (4/52)

2. Rolling a three with a (fair) single die?

   p=.17 (1/6)

3. Probability of tossing two heads in a row with a 50p coin?

   p=.25 (HH; HT; TT; TH = ¼ or ½ * ½)

Probability Distribution of Outcomes Single Roll of a Pair of (fair) Dice

Score	Probability	(p)
2	1/36	.02778
3	2/36	.05556
4	3/36	.08333
5	4/36	.11111
6	5/36	.13889
7	6/36	.16667
8	5/36	.13889
9	4/36	.11111
10	3/36	.08333
11	2/36	.05556
12	1/36	.02778
	36/36	**1.00000**

Probability Distribution of Outcomes
Single Roll of a Pair of (fair) Dice

Score	Probability	(p)
2	1/36	.02778
3	2/36	.05556
4	3/36	.08333
5	4/36	.11111
6	5/36	.13889
7	6/36	.16667
8	5/36	.13889
9	4/36	.11111
10	3/36	.08333
11	2/36	.05556
12	1/36	.02778
	36/36	**1.00000**

Gerolamo Cardano (1501 – 1576) might have been the first person to work this out formally

Scotland v Brazil

If the game ends

Scotland 2 Brazil 2

What is the probability that it was 0 - 0 at half time?

Probability =

Outcome / Total Number of Outcomes

Scotland	Brazil
0	0
0	1
0	2
1	0
1	1
1	2
2	0
2	1
2	2

Probability = Outcome /Total Number of Outcomes

1 / 9 chance it was 0 - 0 at half time

(if all scores are equally likely)

P values

	(Ten)	Hundred				
p=.	9	9				
p=.	2	5				
p=.	1	0				

P values

	(Ten)	Hundred	Thousand			
p=.	0	5				
p=.	0	1				
p=.	0	0	1			

P values

	(Ten)	Hundred	Thousand	Ten Thousand	Hundred Thousand	Million
p=.	0	0	0	1		
p=.	0	0	0	0	1	
p=.	0	0	0	0	0	1

1 in a million - 10p tossed 20 times all coming out heads

7944-015283747-084079

Good luck for your draw on Wed 02 Oct 13

Your numbers

A 05 16 27 42 46 47

1 play x £1.00 for 1 draw = £ 1.00

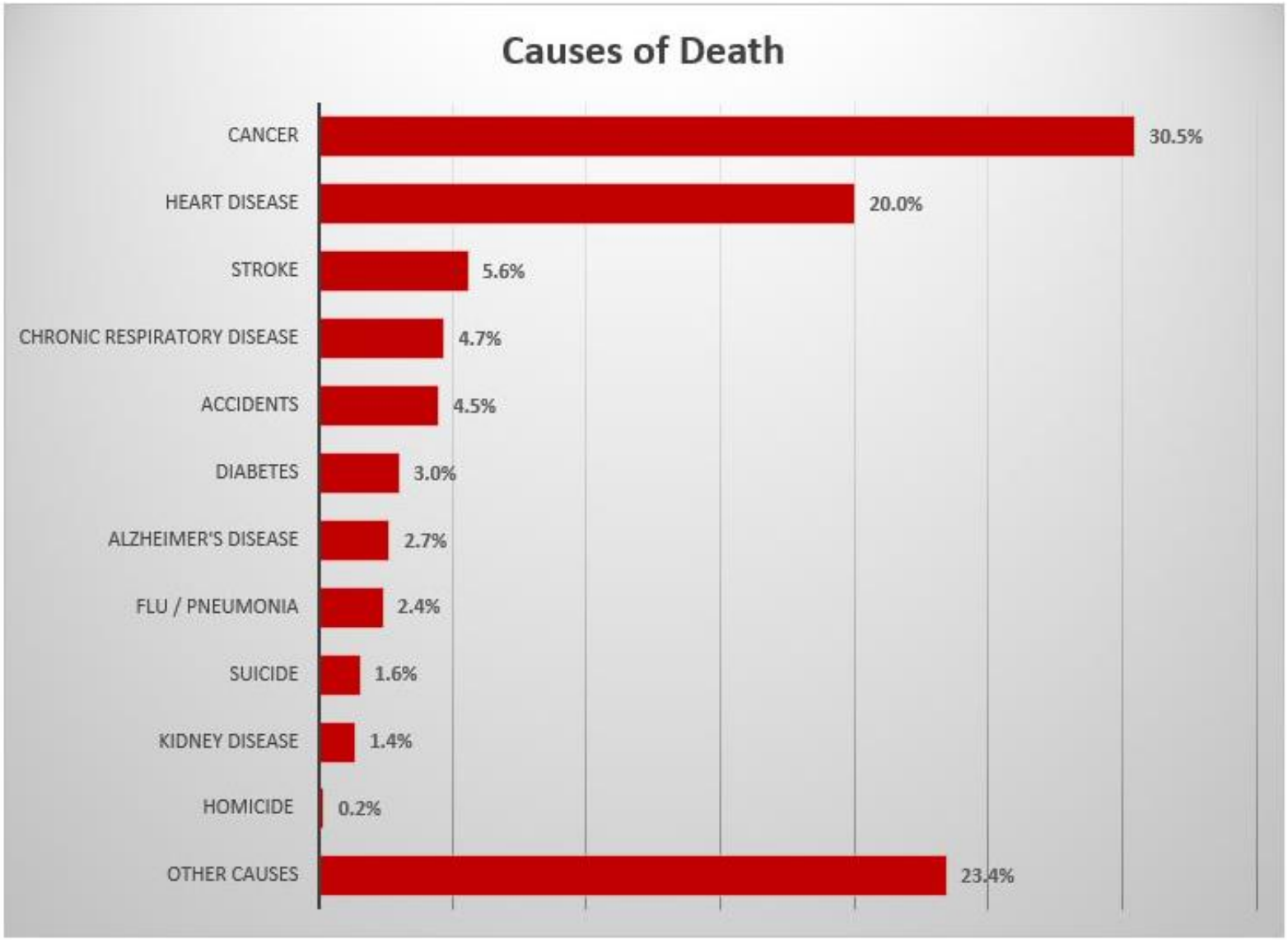
NEW LOTTO IS NEARLY HERE

TICKETS ON SALE FROM THURSDAY

7944-015283747-084079 014421 Term. 44412401

[. . . . .] Fill the box to void the ticket

How are you most likely to die? The chart below summarizes the probability of death by various causes for the average Canadian.



Part 8 Experiments

Rev Edward Stone (1702-1768)

Discovered the active ingredient of aspirin

He wrote to the Royal Society on 25 April 1763

was always given in powders, with any common vehicle, as water, tea, ſmall beer and ſuch like. This was done purely to aſcertain its effects; and that I might be aſſured the changes wrought in the patient could not be attributed to any other thing

I have no other motives for publiſhing this valuable ſpecific, than that it may have a fair and full trial in all its variety of circumſtances and ſituations, and that the world may reap the benefits accruing from it. For theſe purpoſes I have given this long and minute account of it, and which I would not have troubled your Lordſhip with, was I not fully perſuaded of the wonderful efficacy of this Cortex Salignus in agues and intermitting caſes, and did I not think, that this perſuaſion was ſufficiently ſupported by the manifold experience, which I have had of it.

I am, my Lord,

with the profoundeſt ſubmiſſion and reſpect,

Chipping-Norton,
Oxfordſhire,
April 25, 1763.

your Lordſhip's moſt obedient
humble Servant

Edward Stone.

Experiments

- Two Groups
- Experimental Group – receives a treatment
- Control Group – provides a comparison

Experiments

- Two Groups
- Experimental Group – drug group
- Control Group – placebo group

Experiments

- Experimental Group New reading scheme
- Control Group = Existing reading scheme

Experiments

There are many more complex experimental designs!!

International Journal of Sport Nutrition and Exercise Metabolism, 2005, 15, 386-400

Effects of Acute Versus Chronic L-Carnitine L-tartrate Supplementation on Metabolic Responses to Steady State Exercise in Males and Females

Weronika N. Abramowicz and Stuart D.R. Galloway

Twelve healthy active subjects (6 male, 6 female) performed 60 min of exercise (60% VO_{2max}) on 3 occasions after supplementing with L-Carnitine L-tartrate (LCLT) or placebo. Each subject received a chronic dose, an acute dose, and placebo in a randomized, double-blind crossover design. Dietary intake and exercise were replicated for 2 d prior to each trial. In males there was a significant difference in rate of carbohydrate (CHO) oxidation between placebo and chronic trials ($P = 0.02$) but not placebo and acute trials ($P = 0.70$), and total CHO oxidation was greater following chronic supplementation vs. placebo (mean ± standard deviation) of 93.8 (17.3) g/hr and 78.2 (23.3) g/h, respectively). In females, no difference in rate of, or total, CHO oxidation was observed between trials. No effects on fat oxidation or hematological responses were noted in either gender group. Under these experimental conditions, chronic LCLT supplementation increased CHO oxidation in males during exercise but this was not observed in females

Key Words: carbohydrate oxidation, fat oxidation, gender

BRITISH MEDICAL JOURNAL

LONDON SATURDAY OCTOBER 30 1948

STREPTOMYCIN TREATMENT OF PULMONARY TUBERCULOSIS

A MEDICAL RESEARCH COUNCIL INVESTIGATION

The following gives the short-term results of a controlled investigation into the effects of streptomycin on one type of pulmonary tuberculosis. The inquiry was planned and directed by the Streptomycin in Tuberculosis Trials Committee, composed of the following members: Dr. Geoffrey Marshall (chairman), Professor J. W. S. Blacklock, Professor C. Cameron, Professor N. B. Capon, Dr. R. Cruickshank, Professor J. H. Gaddum, Dr. F. R. G. Heaf, Professor A. Bradford Hill, Dr. L. E. Houghton, Dr. J. Clifford Hoyle, Professor H. Raistrick, Dr. J. G. Scadding, Professor W. H. Tytler, Professor G. S. Wilson, and Dr. P. D'Arcy Hart (secretary). The centres at which the work was carried out and the specialists in charge of patients and pathological work were as follows:

Brompton Hospital, London.—Clinician: Dr. J. W. Crofton, Streptomycin Registrar (working under the direction of the honorary staff of Brompton Hospital); Pathologists: Dr. J. W. Clegg, Dr. D. A. Mitchison.

Colindale Hospital (L.C.C.), London.—Clinicians: Dr. J. V. Hurford, Dr. B. J. Douglas Smith, Dr. W. E. Snell; Pathologists (Central Public Health Laboratory): Dr. G. B. Forbes, Dr. H. D. Holt.

Harefield Hospital (M.C.C.), Harefield, Middlesex.—Clinicians: Dr. R. H. Brent, Dr. L. E. Houghton; Pathologist: Dr. E. Nassau.

Bangour Hospital, Bangour, West Lothian.—Clinician: Dr. I. D. Ross; Pathologist: Dr. Isabella Purdie.

Killingbeck Hospital and Sanatorium, Leeds.—Clinicians: Dr. W. Santon Gilmour, Dr. A. M. Reevie; Pathologist: Professor J. W. McLeod.

Northern Hospital (L.C.C.), Winchmore Hill, London.—Clinicians: Dr. F. A. Nash, Dr. R. Shoulman; Pathologists: Dr. J. M. Alston, Dr. A. Mohun.

Sully Hospital, Sully, Glam.—Clinicians: Dr. D. M. E. Thomas, Dr. L. R. West; Pathologist: Professor W. H. Tytler.

The clinicians of the centres met periodically as a working subcommittee under the chairmanship of Dr. Geoffrey Marshall; so also did the pathologists under the chairmanship of Dr. R. Cruickshank. Dr. Marc Daniels, of the Council's scientific staff, was responsible for the clinical co-ordination of the trials, and he also prepared the report for the Committee, with assistance from Dr. D. A. Mitchison on the analysis of laboratory results. For the purpose of final analysis the radiological findings were assessed by a panel composed of Dr. L. G. Blair, Dr. Peter Kerley, and Dr. Geoffrey S. Todd.

“If your experiment needs statistics, then you ought to have done a better experiment” Ernest Rutherford (1871-1937)

Therein lies the rub….

With the notable exception of psychology, and to a lesser extent economics, in the social sciences experimentation is often not routinely possible

(e.g. we cannot randomise people to ethnic and gender groups, social housing, schools, local authorities etc. etc.)

The Social World is Complex!

In the non-experimental social sciences when using genuine data we must use more comprehensive statistical methods which better help us to identify, and then quantify, the multifaceted relationships that characterise contemporary social life

Part 9 Hypotheses

Hypothesis

Hypothesis - A proposition that is advanced for testing, or appraising, a generalization regarding the empirical social world

Null Hypothesis H_0

- There is no structured, or systematic, difference between Group A and Group B in our experiment

Alternate Hypothesis H_1

- There is a structured, or systematic, difference between Group A and Group B in our experiment

Null Hypothesis H_0

Mean score in a statistics exam **is not** systematically different for those taught by instructor A and instructor B

Alternate Hypothesis H_1

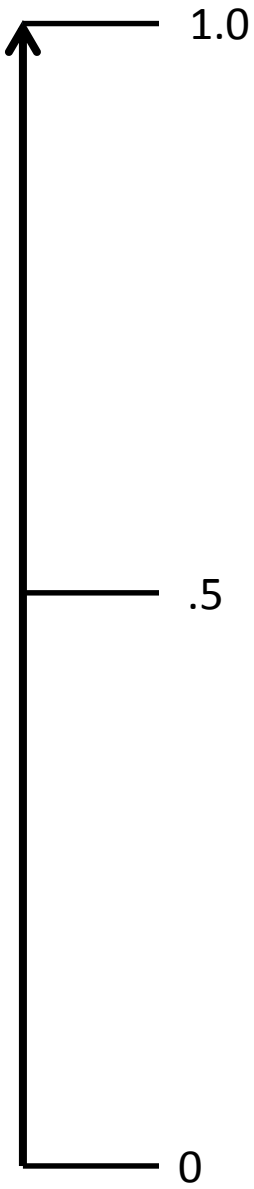
Mean score in a statistics exam **is** systematically different for those taught by instructor A and instructor B

Null Hypothesis H_0

$$\overline{X}_A = \overline{X}_B$$

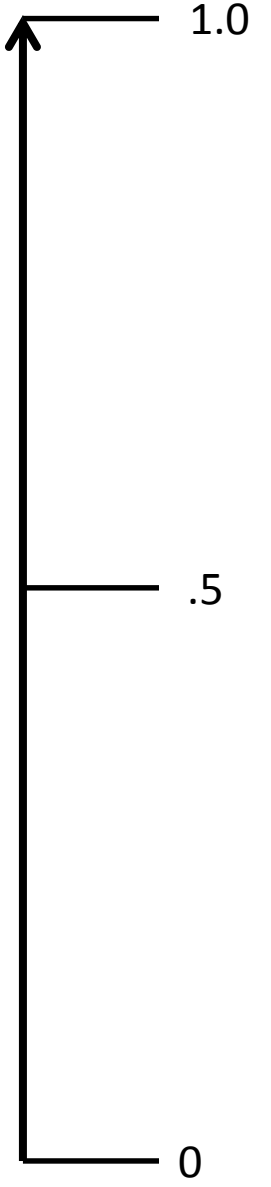
Alternate Hypothesis H_1

$$\overline{X}_A \neq \overline{X}_B$$



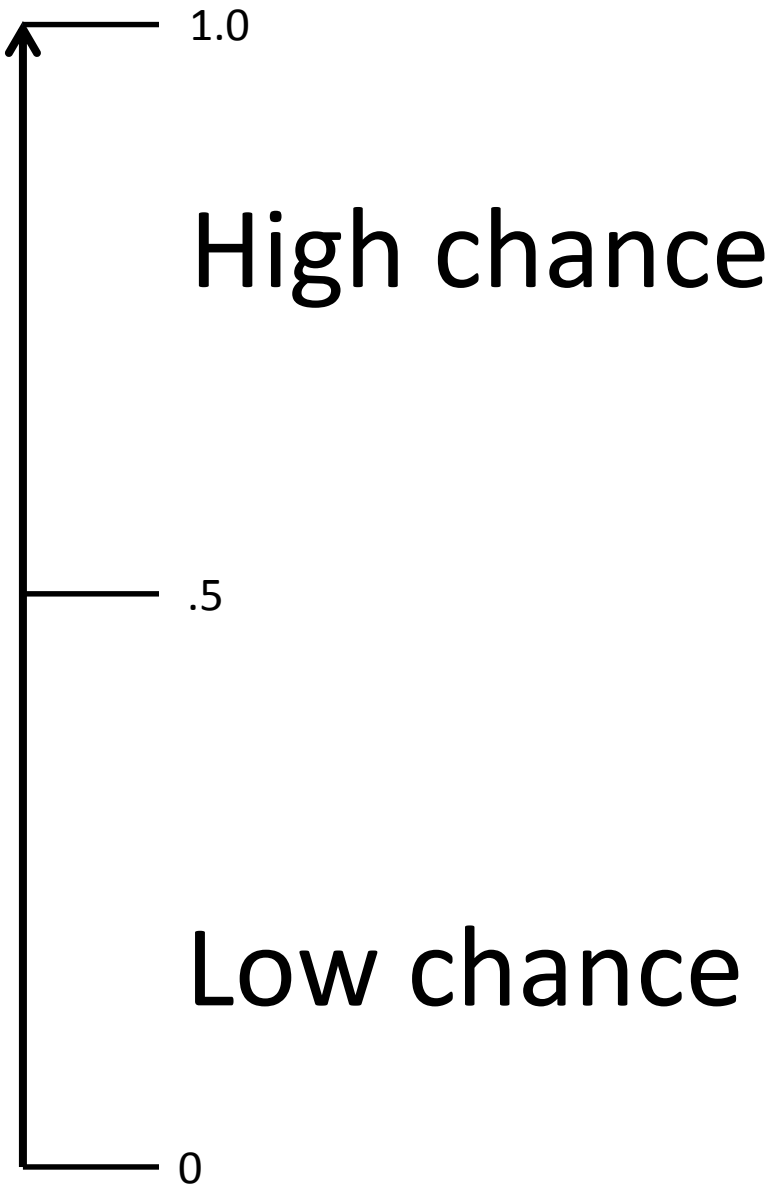
Probabilities take on values between 0 and 1

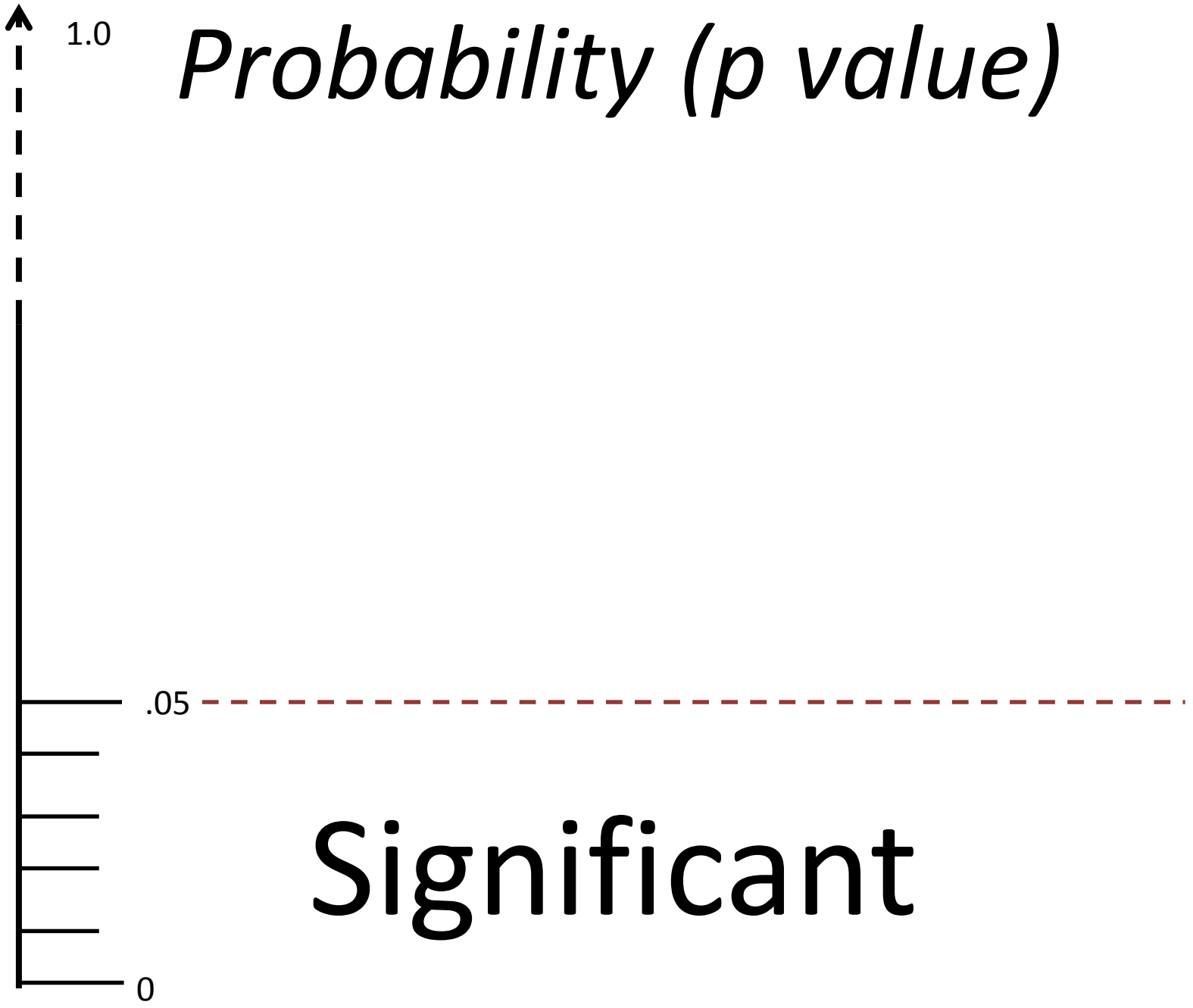
Denoted a “p”



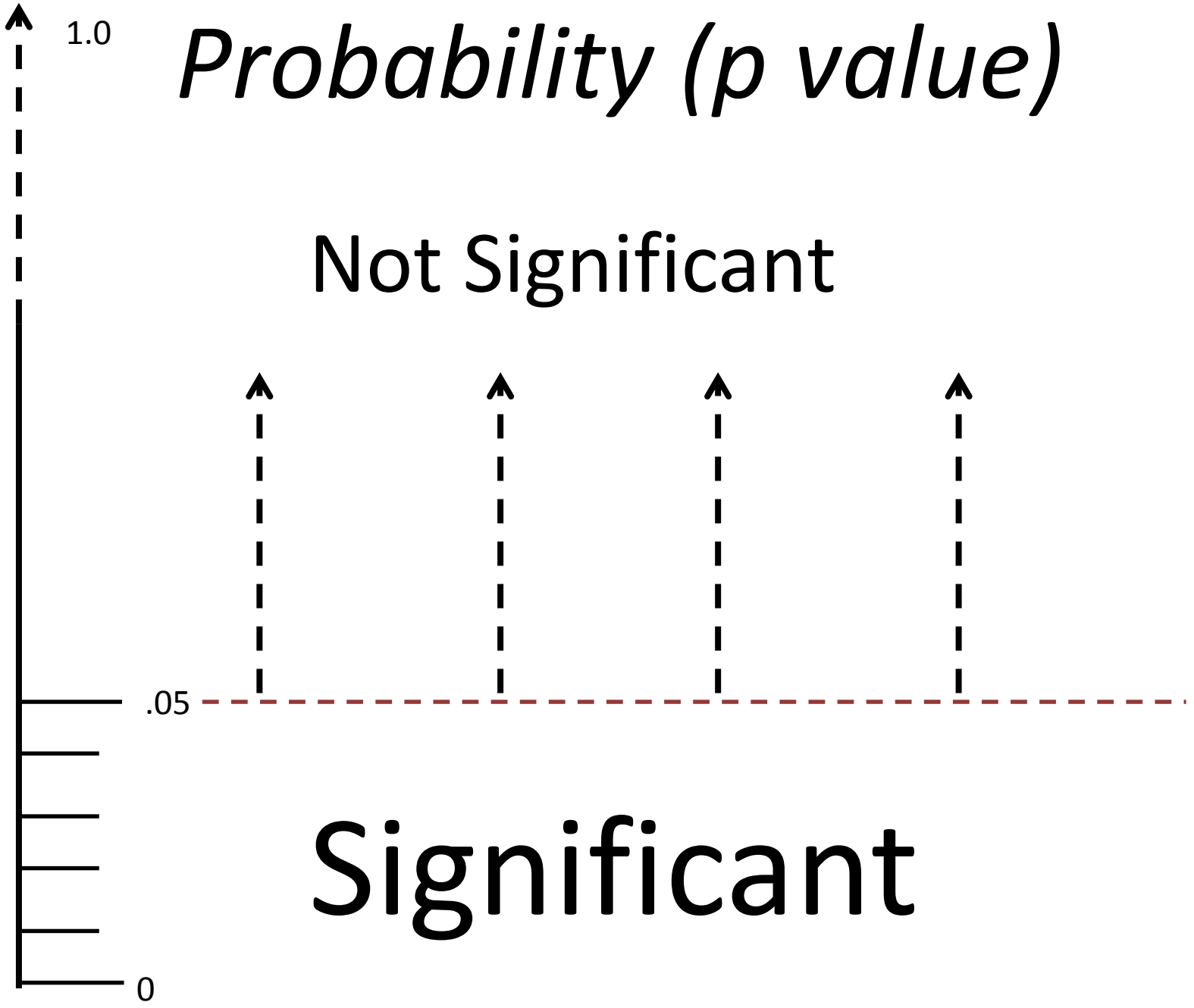
Event will definitely occur

Event will definitely **not** occur

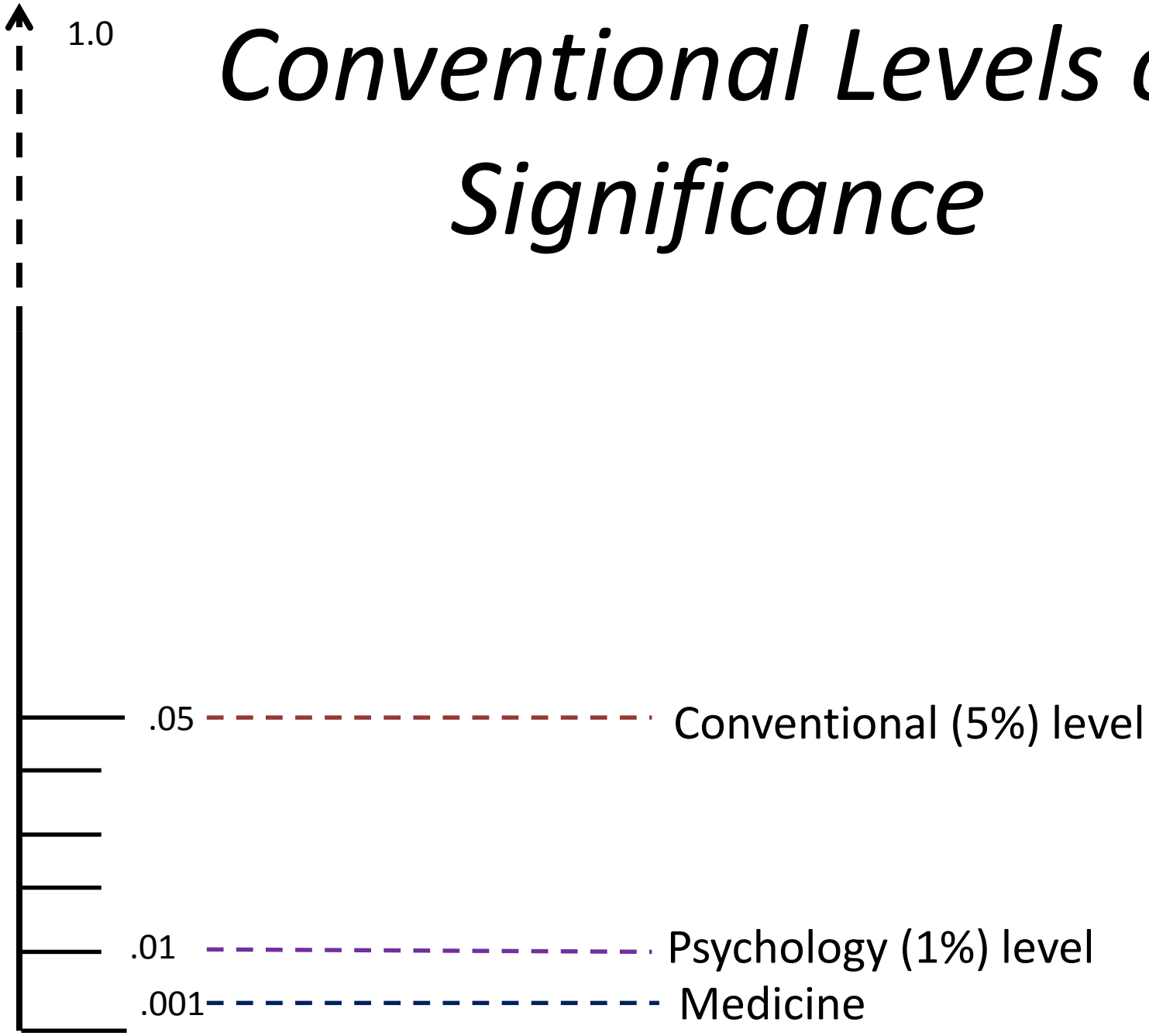

1.0
High chance
.5
Low chance
0


1.0
Probability (p value)
.05
Significant
0

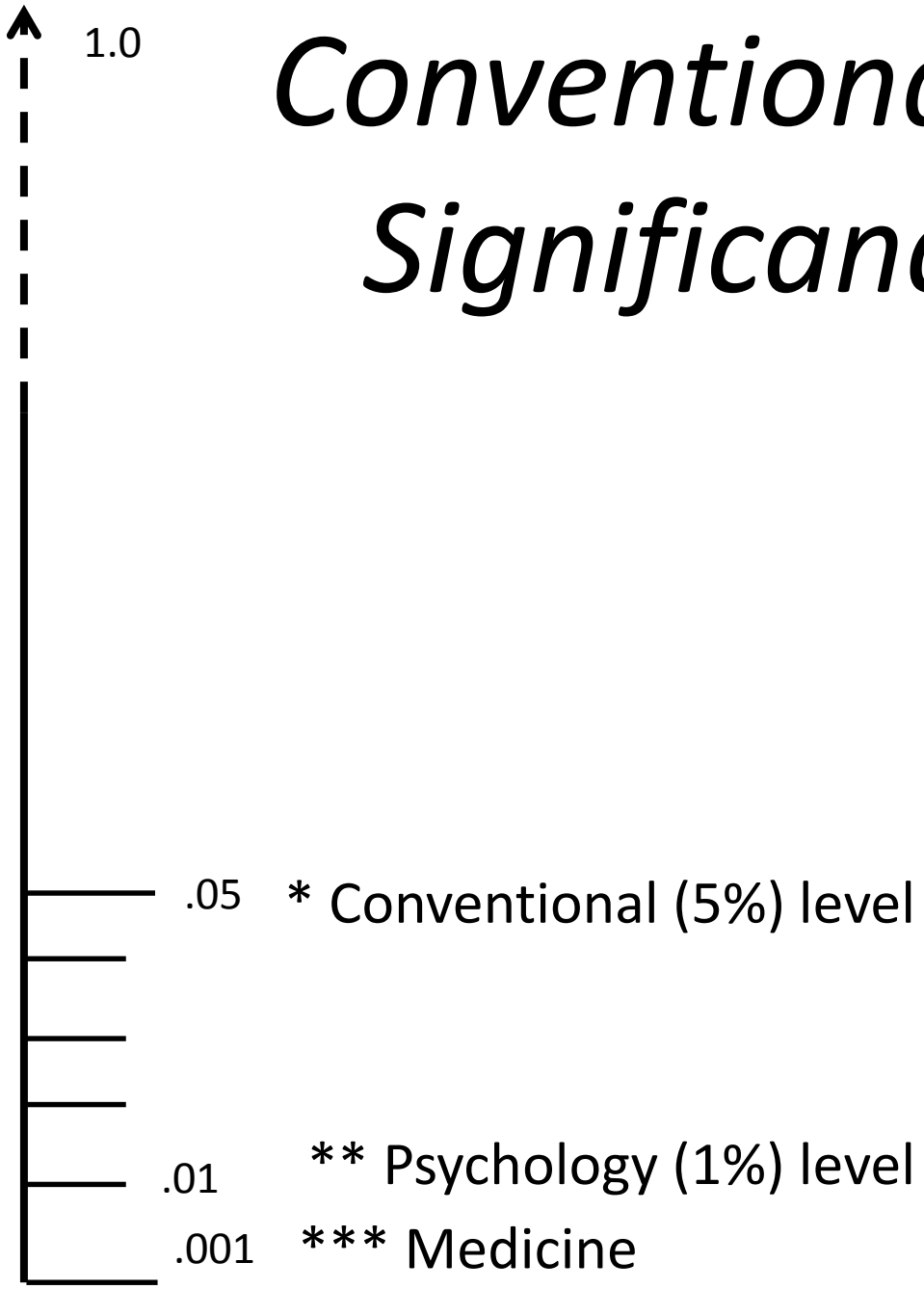
Probability (p value)



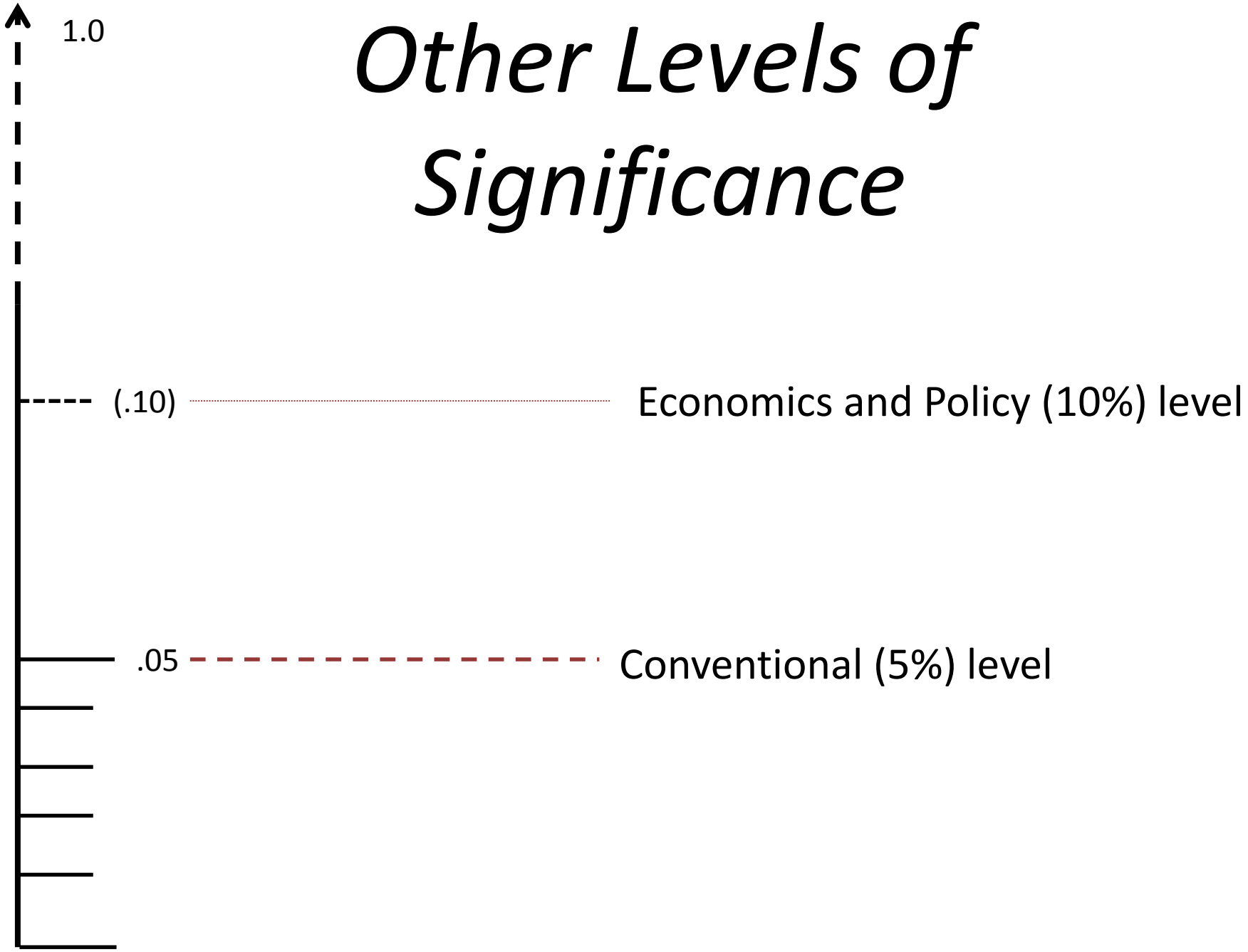
Conventional Levels of Significance



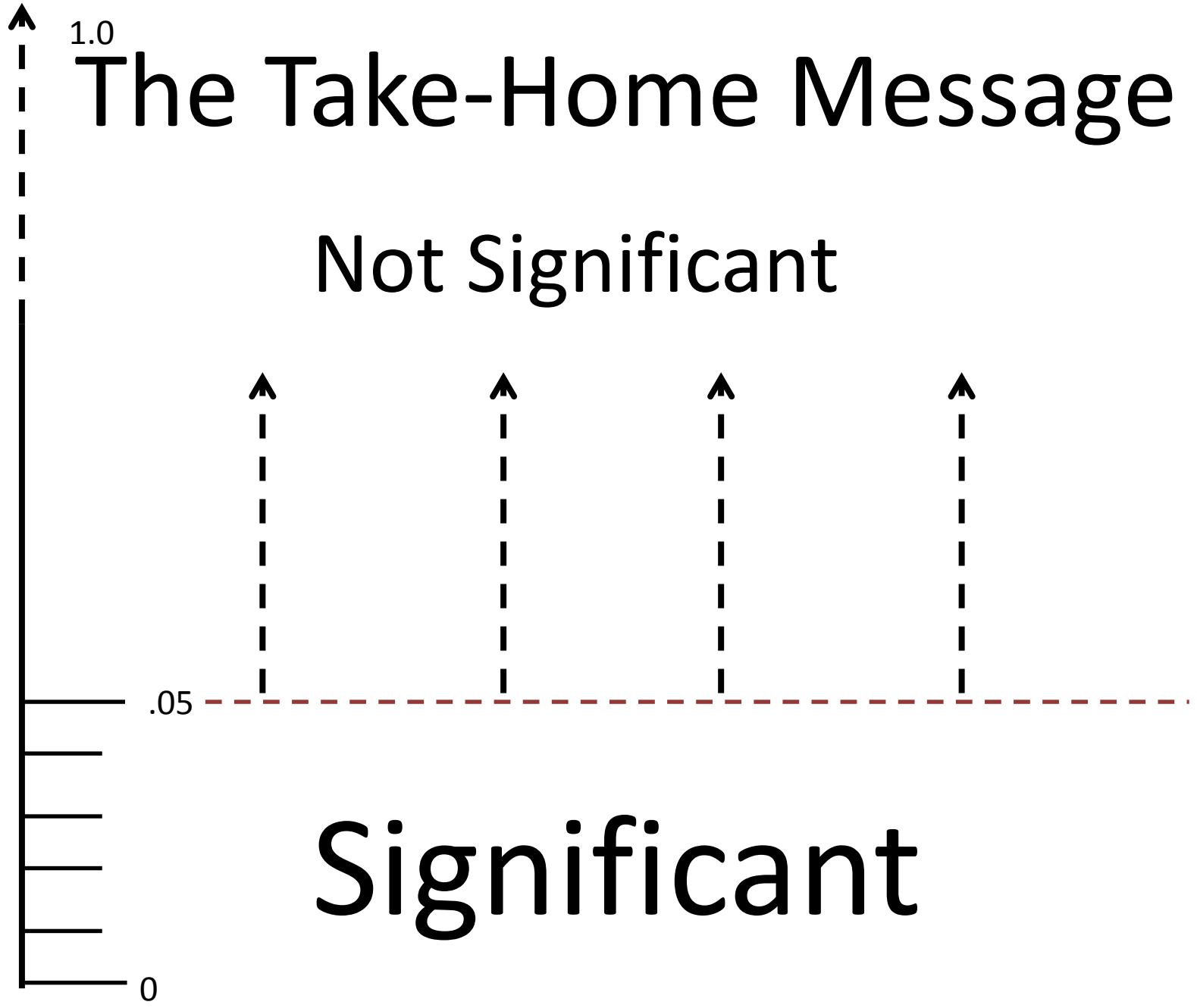
Conventional Levels of Significance (Stars)



Other Levels of Significance



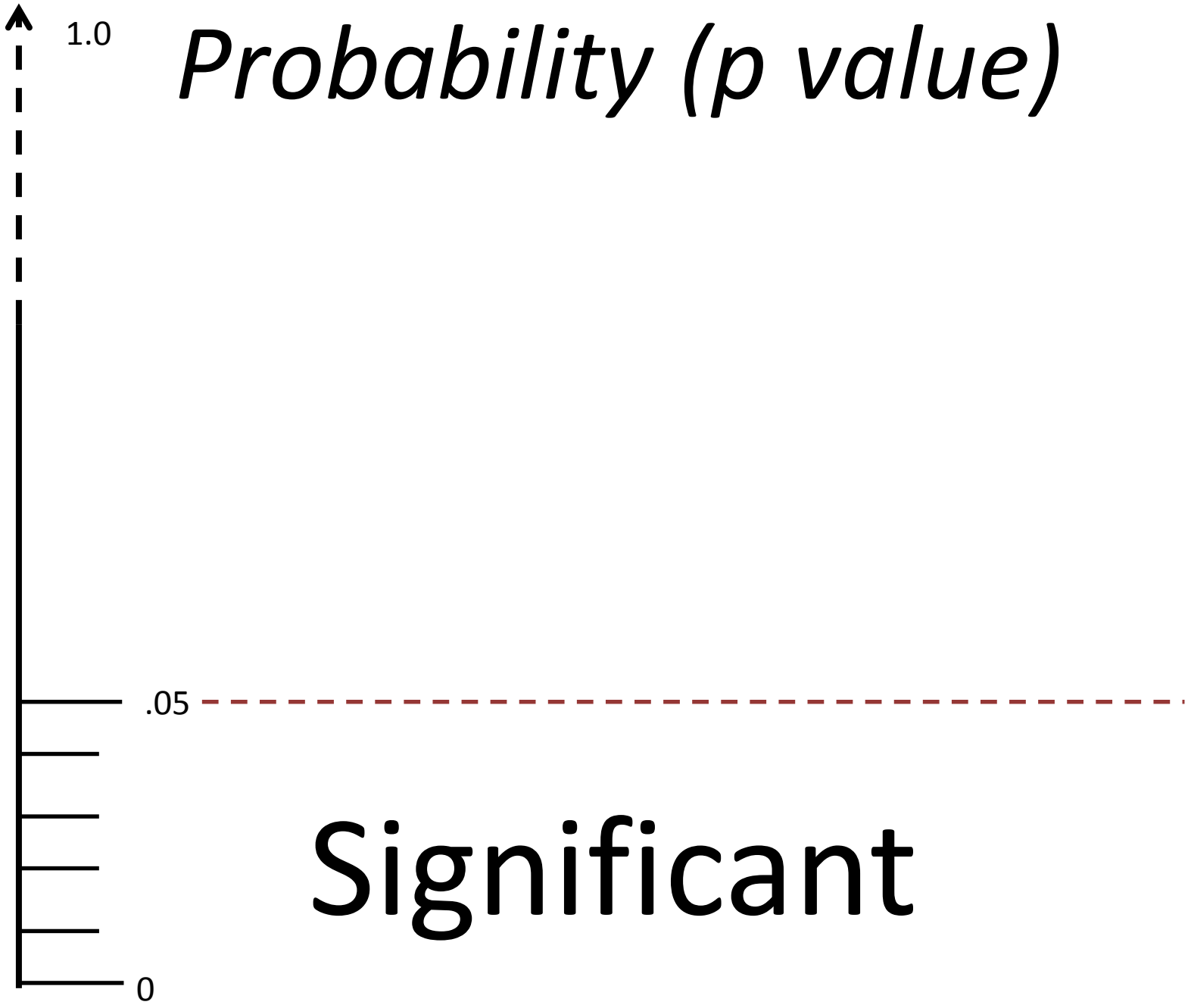
The Take-Home Message



HR computer test

	Mean	n	s.d.
Overall	11.38	50	5.26
West Dungyle	7.64	25	3.82
East Dungyle	15.12	25	5.87

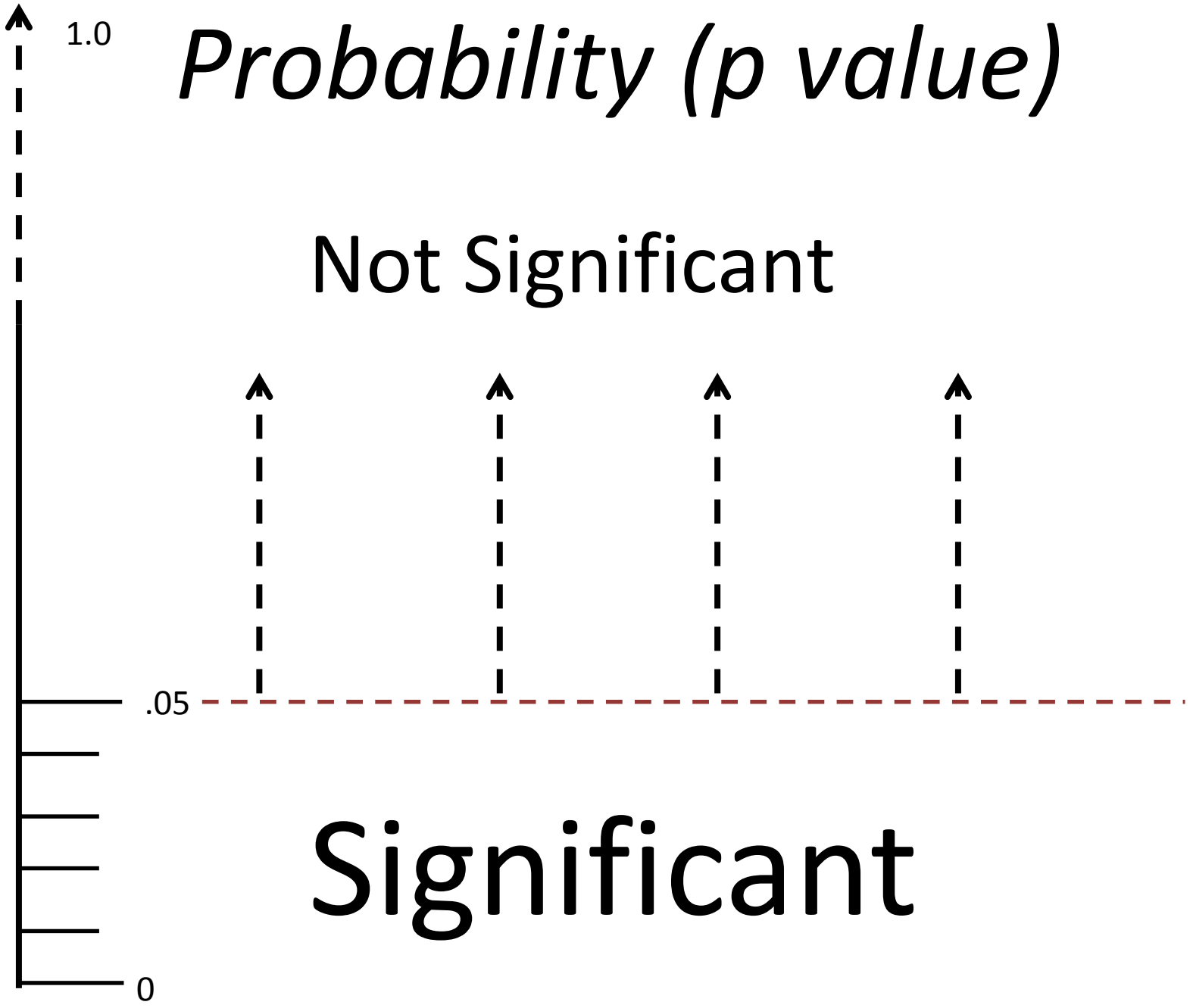
$p<.001$


1.0
Probability (p value)
.05
Significant
0

HR law test

	Mean	n	s.d.
Overall	7.56	50	3.69
West Dungyle	7.72	25	2.93
East Dungyle	7.40	25	3.30

p=.74


1.0
Probability (p value)
Not Significant
.05
Significant
0

Testing Hypotheses

Some extra information on the intuition behind significance tests

Designed for keen students!

HR computer test

Null Hypothesis H_0

Mean score for the computer test **is not** systematically different for those in West Dungyle and East Dungyle

Alternate Hypothesis H_1

Mean score for the computer test **is** systematically different for those in West Dungyle and East Dungyle

West Dungyle (Group A) and

East Dungyle (Group B)

$$\overline{X}_A = 7.64$$

$$\overline{X}_B = 15.12$$

What is the probability that *Ho* is correct?

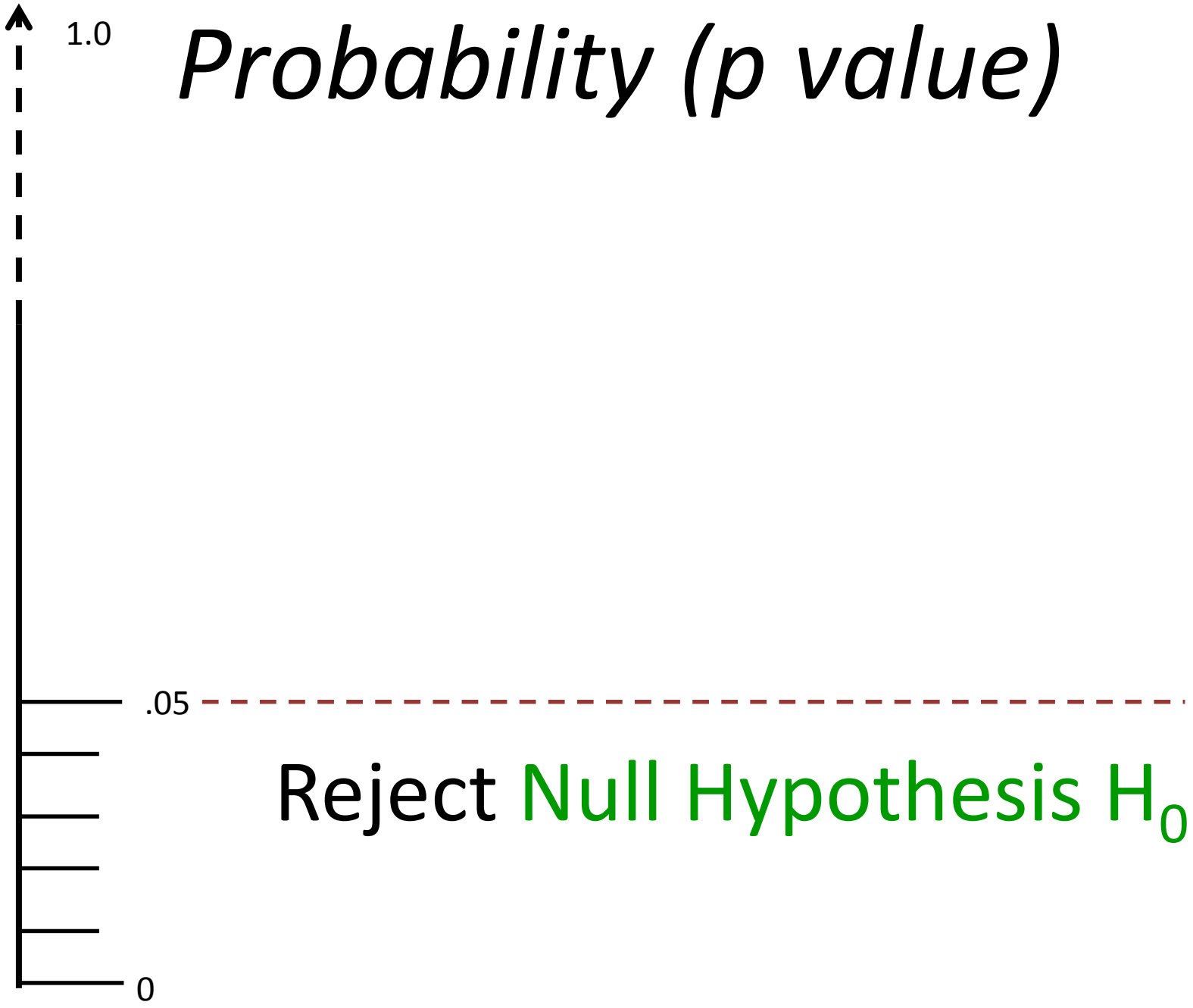
HR computer test

Null Hypothesis H_0

Mean score for the computer test **is not** systematically different for those in West Dungyle and East Dungyle

Probability that H_0 is correct is $p<.001$

Can we reject the Null Hypothesis?


1.0
Probability (p value)
.05
Reject Null Hypothesis H_0
0

HR law test

	Mean	n	s.d.
Overall	7.56	50	3.69
West Dungyle	7.72	25	2.93
East Dungyle	7.40	25	3.30

p=.74

HR law test

1. Write out the Null and the Alternate Hypotheses
2. Write out the overall mean
3. Write out the mean for West Dungyle and the mean for East Dungyle
4. Consider the p value (.74), can you reject the Null Hypothesis?
5. Is the difference between the two mean law test scores significant?

Part 10 Testing Differences in Means

The *t* Test

- A test of group differences (means)
- Metric Y variable
- Categorical (group) X variable
- William Sealy Gosset
  - Guinness
  - pen name ‘Student’

The *t* Test

Example car accidents in two areas (2010)

n= 74

y = Mean cost of cars accident

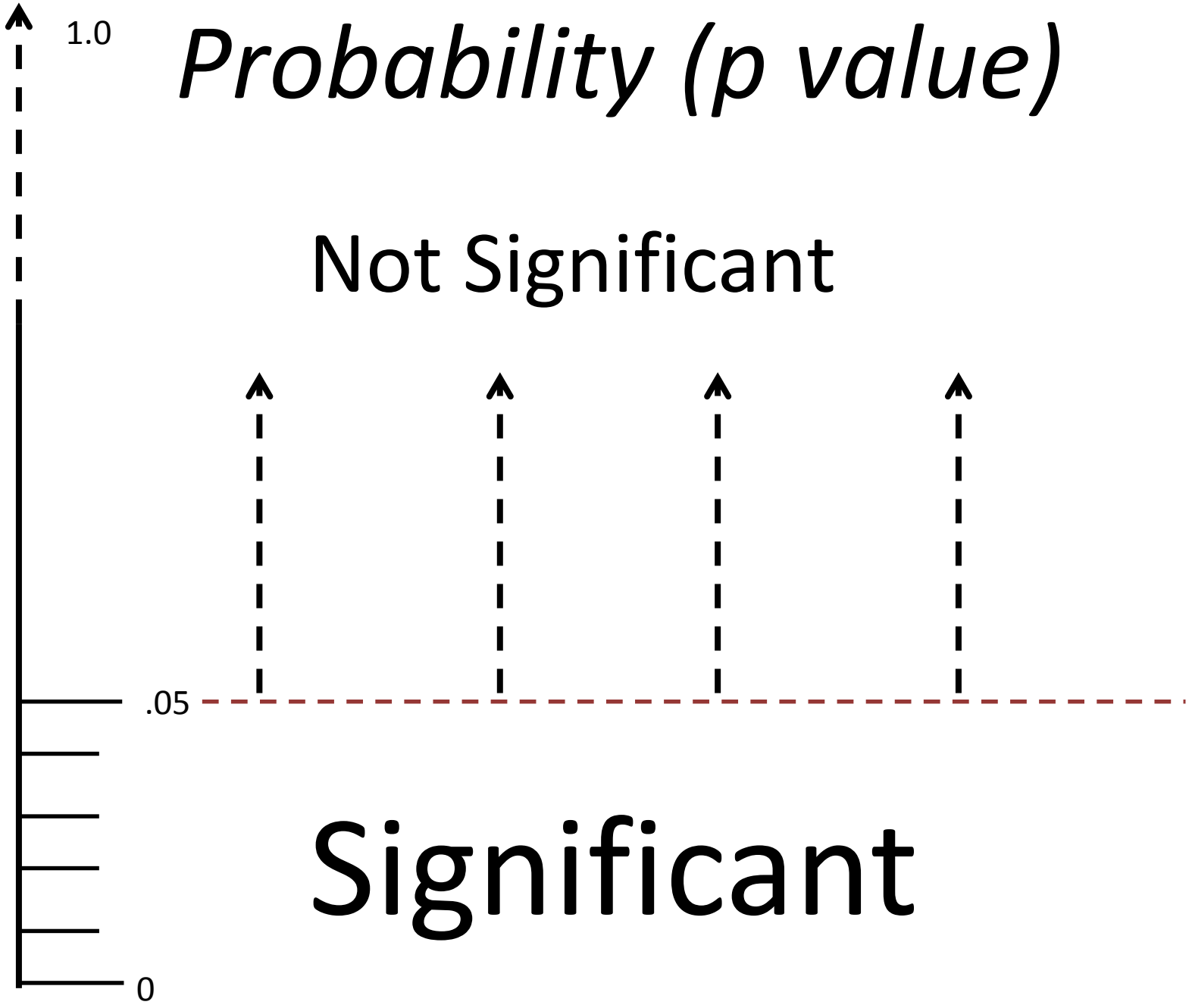
X_1 Area (Area A; Area B)

Group	Obs	Mean
Area A	52	6072
Area B	22	6385

t = -0.41

p < .68

Probability (p value)



Car accidents 2010

1. Write out the Null and the Alternate Hypotheses
2. Write out the mean for Area A and the mean for Area B

3. Consider the p value (.68)

4. Can you reject the Null Hypothesis?
5. Is their a significant difference between the two areas?

Take a look at the confidence intervals

Do they overlap or is there nice clear blue water separating them?

Group	Obs	Mean	Std. Err.	[95% Conf.	Interval]
Area A	52	6072	429	**5210**	**6935**
Area B	22	6385	559	**5222**	**7547**

The *t* Test

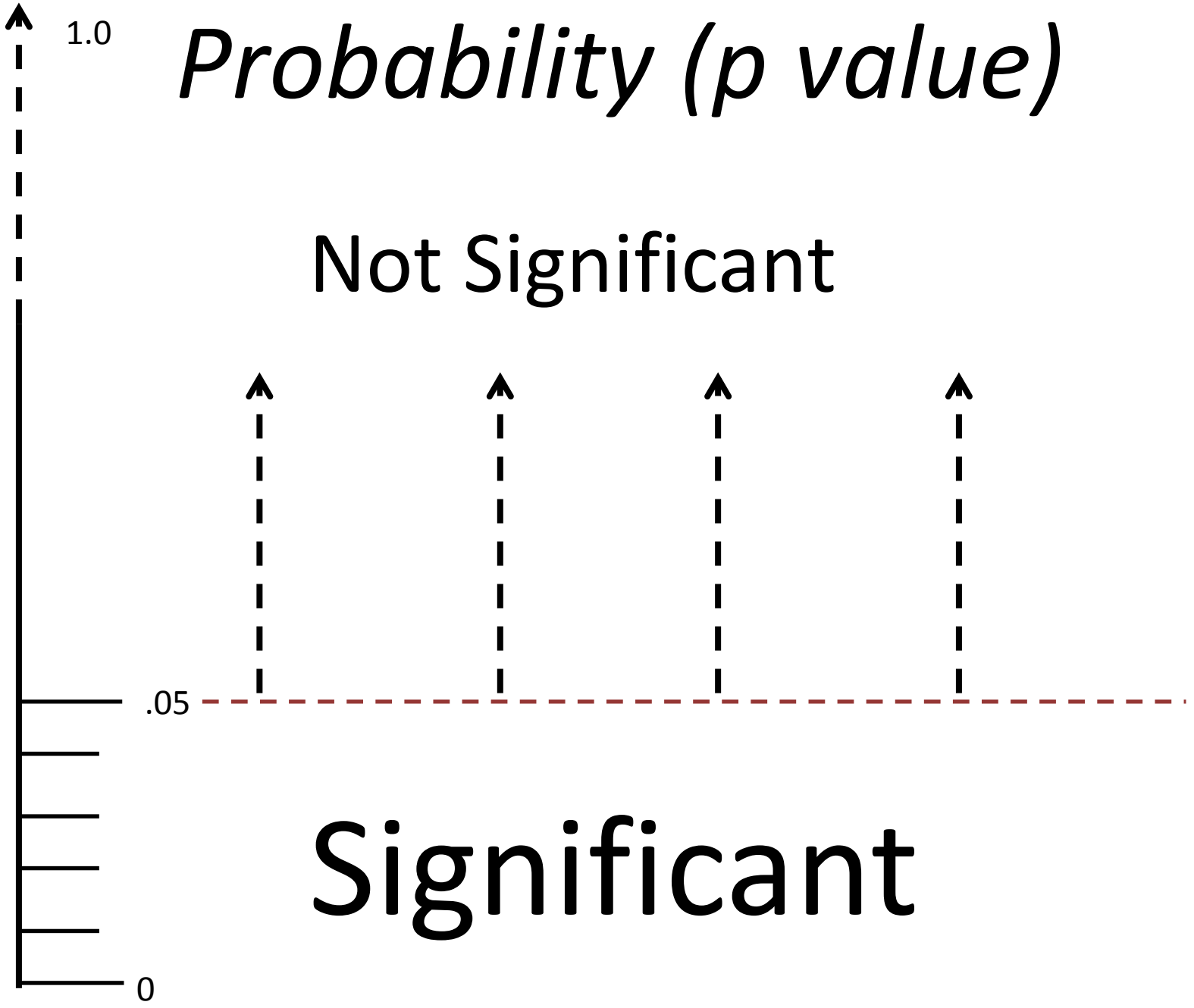
Example car accidents in two areas (2013)

Group	Obs	Mean
Area A	52	3317
Area B	22	2316

t = 6.25

p < .01

Probability (p value)



Take a look at the confidence intervals

Do they overlap or is there nice clear blue water separating them?

Group	Obs	Mean	Std. Err.	[95% Conf.	Int]
Area A	52	3317	96	**3124**	**3511**
Area B	22	2316	92	**2124**	**2508**

Part 12 Data in Tables

Frequencies in tables (BHPS 1999)

	Own / buy home	Social housing	Private renters	Total
Cons				
Labour				
Liberal				
Total	7187	2101	774	**10072**

Frequencies in tables (BHPS 1999)

	Own / buy home	Social housing	Private rent	Total
Cons	2432	327	191	2950
Labour	3653	1551	433	5637
Liberal	1102	223	150	1475
Total	7187	2101	774	**10072**

Percentages in tables (BHPS 1999)

- **column** {and total} percents for cells

	Own / buy home	Social housing	Private renters	Total
Cons	**34** 2432	**16** 327	**25** 191	**29** 2950
Labour	**51** 3653	**74** 1551	**56** 433	**56** 5637
Liberal	**15** 1102	**11** 223	**19** 150	**15** 1475
Total	**100** 7187	**100** 2101	**100** 774	 **10072**

Percentages in tables (BHPS 1999)

- ***Row*** {and total} percents for cells

	Own / buy home	Social housing	Private renters	Total
Cons	***83***	***11***	***7***	***100***
	2432	327	191	2950
Labour	***65***	***28***	***8***	***100***
	3653	1551	433	5637
Liberal	***75***	***15***	***10***	***100***
	1102	223	150	1475
Total	***71***	***21***	***8***	
	7187	2101	774	**10072**

Driving Tests

	Males	Females
Failed	100	0
Passed	0	100

Driving Tests

	Males	Females
Failed	50	50
Passed	50	50

Driving Tests

	Males	Females
Failed	31	45
Passed	69	55

Driving Tests

1. What is the outcome variable?

2. What are the two groups in the explanatory variable?

3. Write out the Null and the Alternate Hypotheses

Driving Tests

	Males	Females
Failed	31	45
Passed	69	55

p=.04

Driving Tests West Dungyle

	Males	Females
Failed	30	40
Passed	70	60

p=.14

Driving Tests East Dungyle

	Males	Females
Failed	31	45
Passed	69	55

p=.04

Part 13 Tests for tables

	Care less than 1 month	Care more than 1 month	
Died within month	20	6	
Survived at least one month	373	316	
	393	*322*	

Bishop, Y.M., 1969. Full contingency tables, logits, and split contingency tables. *Biometrics*, pp.383-399.

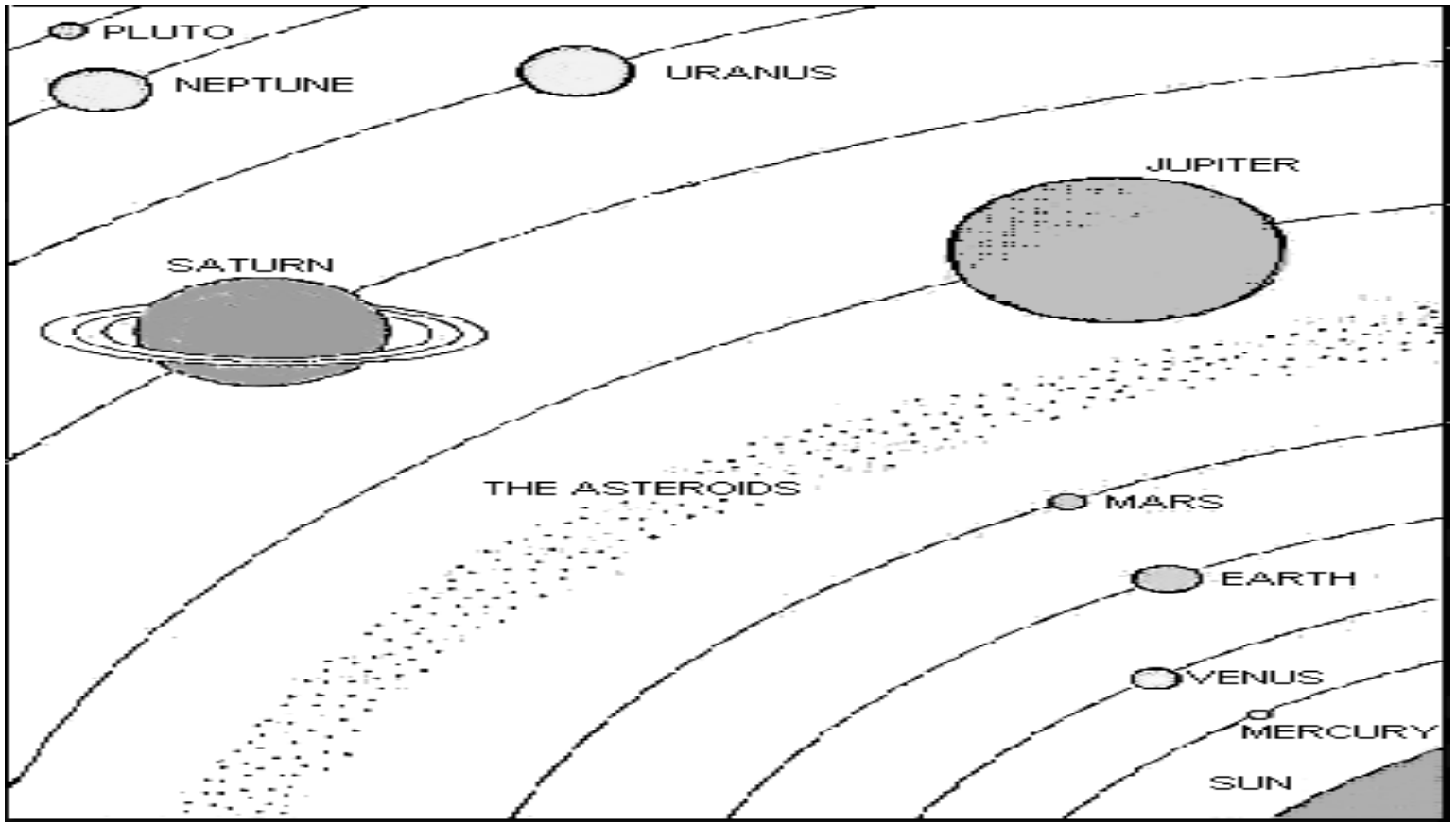
	Care less than 1 month	Care more than 1 month	
Died within month	5.09%	1.86%	
Survived at least one month	94.91%	98.14%	
	100%	*100%*	

Bishop, Y.M., 1969. Full contingency tables, logits, and split contingency tables. *Biometrics*, pp.383-399.

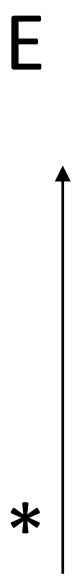
	Care less than 1 month	Care more than 1 month	
Died within month	20	6	
Survived at least one month	373	316	
	393	*322*	

Bishop, Y.M., 1969. Full contingency tables, logits, and split contingency tables. *Biometrics*, pp.383-399.

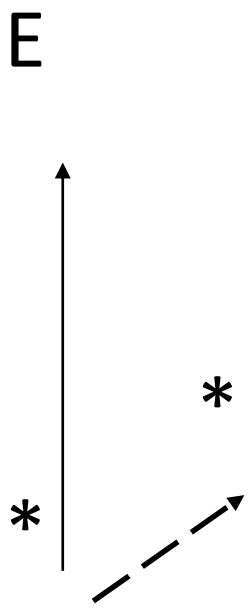
Heavenly Bodies…



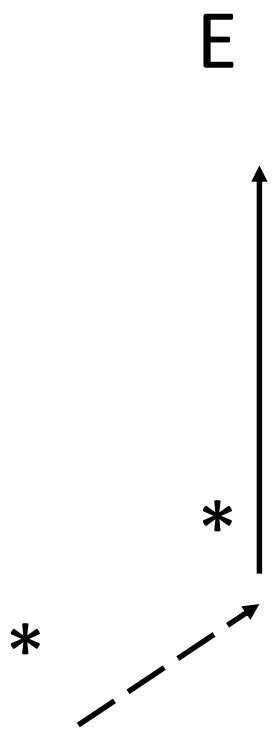
Observations



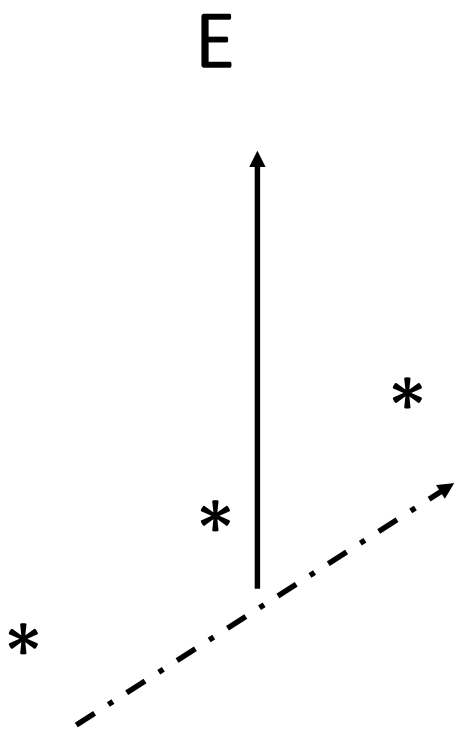
Observations



Observations



Observations



Observations

	Care less than 1 month	Care more than 1 month	
Died within month	20	6	
Survived at least one month	373	316	
	393	*322*	

Bishop, Y.M., 1969. Full contingency tables, logits, and split contingency tables. *Biometrics*, pp.383-399.

Calculating Expected Frequencies

	Care less than 1 month	Care more than 1 month	
Died within month	? →		*26*
Survived at least one month	↓		*689*
	393	*322*	*715*

	Care less than 1 month	Care more than 1 month	
Died within month	(26*393)/715		*26*
Survived at least one month			*689*
	393	*322*	*715*

	Care less than 1 month	Care more than 1 month	
Died within month	(26*393)/715 = 14.3		*26*
Survived at least one month			*689*
	393	*322*	*715*

Expected Frequencies

	Care less than 1 month	Care more than 1 month	
Died within month	14.3	11.7	*26*
Survived at least one month	378.7	310.3	*689*
	393	*322*	*715*

	Care less than 1 month	Care more than 1 month
Died within month	20	6
Survived at least one month	373	316
	393	*322*

OBSERVED

EXPECTED

	Care less than 1 month	Care more than 1 month
Died within month	14.3	11.7
Survived at least one month	378.7	310.3
	393	*322*

	Care less than 1 month	Care more than 1 month
Died within month	20	6
Survived at least one month	373	316
	393	*322*

OBSERVED

EXPECTED

	Care less than 1 month	Care more than 1 month
Died within month	14.3	11.7
Survived at least one month	378.7	310.3
	393	*322*

The Chi-Square Test

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

The Chi-Square Test

O	E	O - E	$(O – E)^2$	$(O – E)^2 / E$
20	14.3	5.7	32.49	2.3
6	11.7	-5.7	32.49	2.8
373	378.7	-5.7	32.49	0.1
316	310.3	5.7	32.49	0.1
				5.3

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Chi-Square Test

- Big chi-square big difference between observed and expected
- Big chi-Square = Big Difference!
- How big is a big chi-square?

Degrees of Freedom df –

size of the table calculated by

(No. of rows – 1) x (No. of columns –1)

The Chi-Square Test

O	E	O - E	$(O – E)^2$	$(O – E)^2 / E$
20	14.3	5.7	32.49	2.3
6	11.7	-5.7	32.49	2.8
373	378.7	-5.7	32.49	0.1
316	310.3	5.7	32.49	0.1
				5.3

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

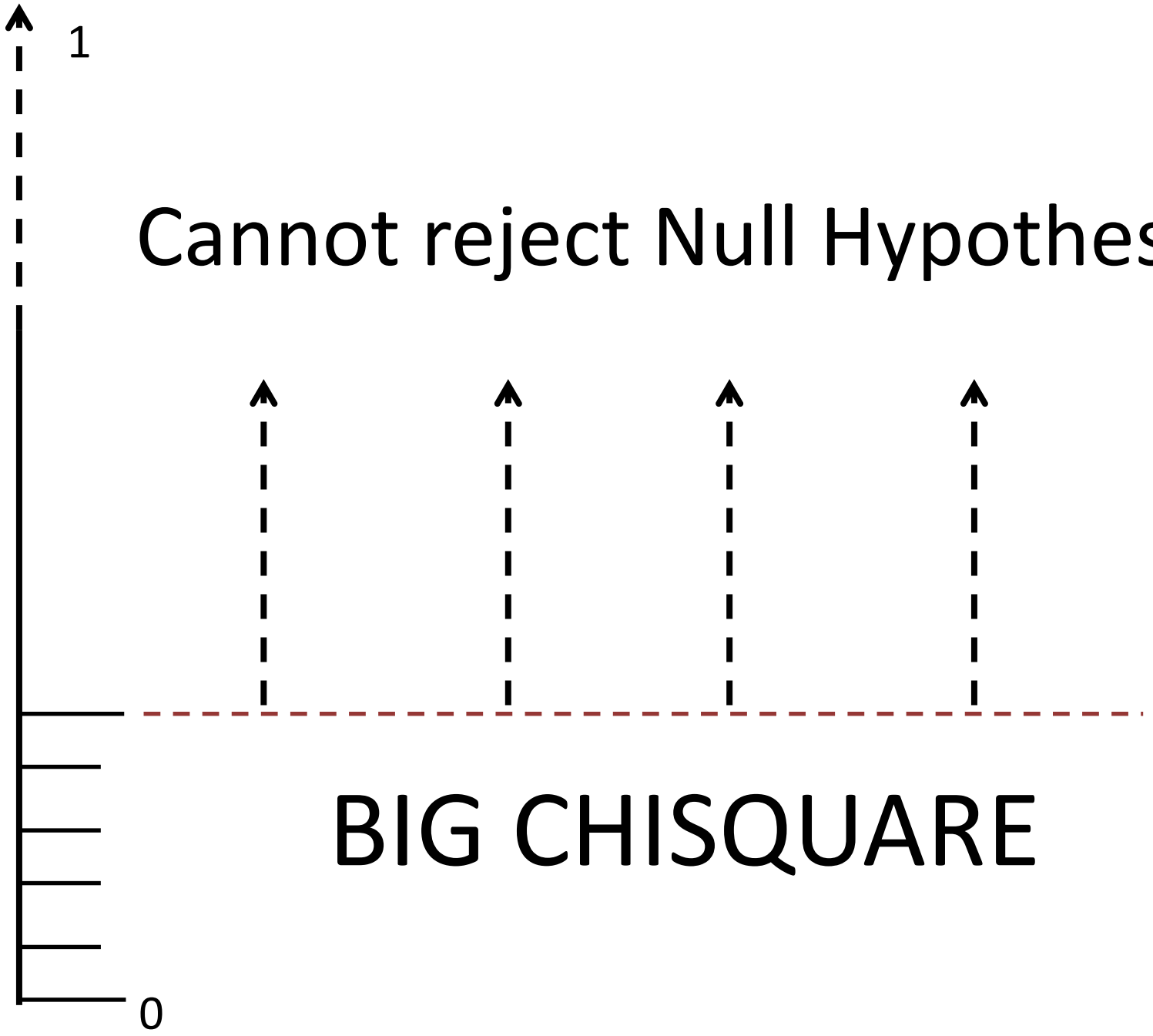
Chi-Square Test

Bigger than Critical Value (CV) at a certain number of degrees of freedom

$p<.05$ if chi-square greater (or equal) to 3.84 @ 1 d.f.

$p<.01$ if chi-square greater (or equal) to 6.64 @ 1 d.f.

Historically, we have got these values from statistical tables


1
Cannot reject Null Hypothesis
BIG CHISQUARE
0

Cramer's V

- Cramer's V is a test of association
  - Denoted as V
  - Values between 0 and 1
  - Weak or strong associations
  - Developed by Harald Cramér

Part 14 Statistical Models

Statistical models augment our ability to investigate this complicated world

Sir Francis Galton (1822-1911)

- Darwin's cousin
- Developed finger printing
- First weather map (Times 1st April 1875)
- Cutting a Round Cake on Scientific Principles (Nature 1906)
- Strawberry Cure for Gout (Nature 1899)
- On Spectacles for Divers
- Beauty Map of Britain (*I found London to rank highest for beauty: Aberdeen lowest, Memoire p.153)*

A Statistical Model - AKA

Simplest statistical model

- A regression model
- Multiple regression
- Linear regression model
- General linear model
- Vanilla regression

A (slightly drunk) statistician once said to me “Vernon, if we didn’t have so many confusing terms we couldn’t charge high consultancy fees”

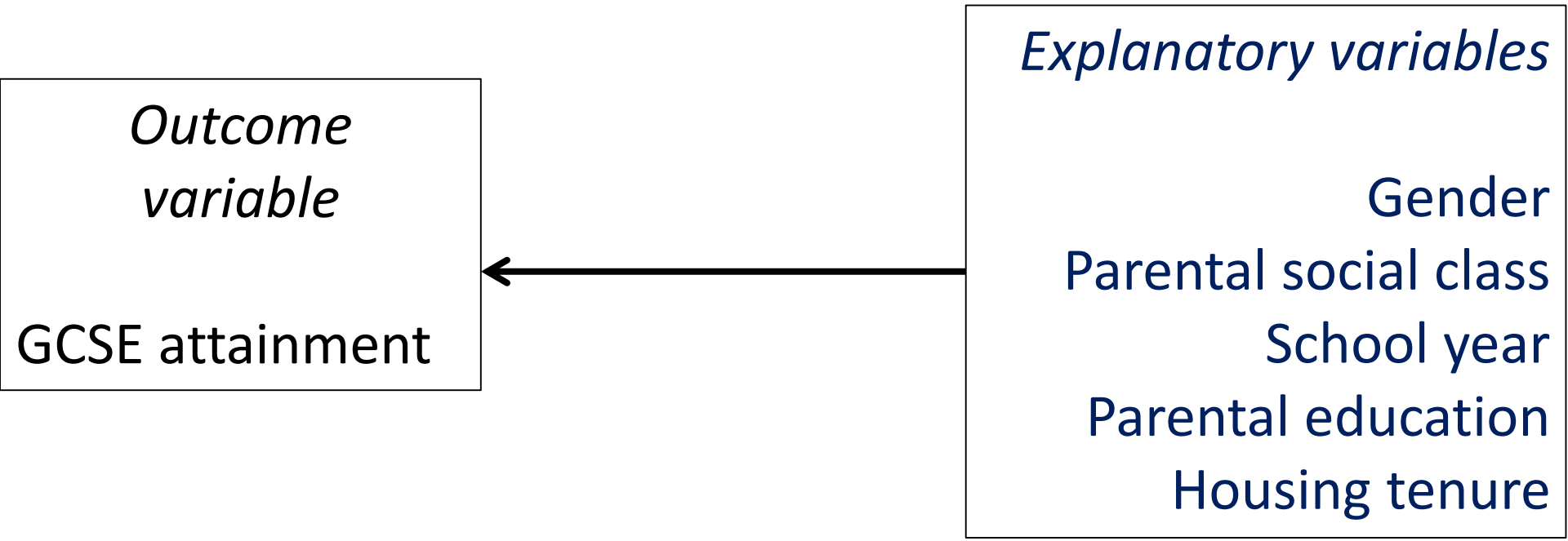
Take Home Message

A standard statistical model has ONE OUTCOME variable

and

MULTIPLE EXPLANATORY variables

Model in a paper



Connelly, Murray and Gayle (2013) *Sociological Research Online*

Take Home Message

A statistical model will usually tell you two things

1. Which variables are important (i.e. significant)

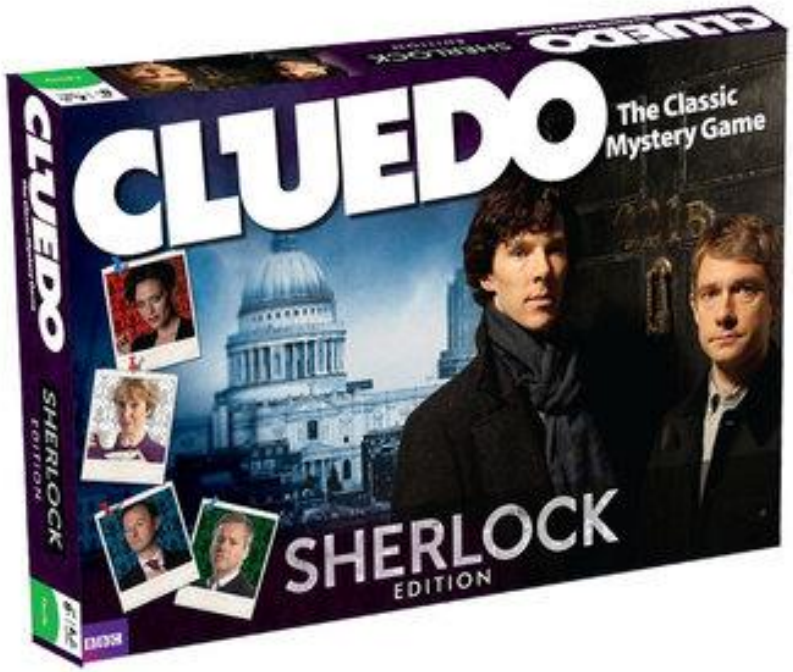
2. The effect of the important variables
   (i.e. the strength and the direction)

Model Building

- John Tukey – Exploratory Data Analysis

- X variables should have the means, the motive and the opportunity to commit the crime of changing the Y variable –

  Robert Luskin, U. of Texas



Part 15 Concluding Remarks

Why More Complex Data Analysis?

My view (although it might be controversial)…

In reality it is unlikely that a bivariate (two explanatory variable) explanation will capture the complexity of the real social world

Therefore there is no choice other than to develop more sophisticated analyses which include more explanatory variables (i.e. statistical models)

Final Thought…

If applied research was easy, theorists would do it. But it is not as hard as the dense pages of Econometrica might lead you to believe. Avoid embarrassment by being your own best sceptic. And, especially, Don't Panic!

Angrist and Pischke (2008) *Mostly Harmless Econometrics*

Introduction to Data Science

An Introduction to Statistical Concepts for Data Analysis

Professor Vernon Gayle

vernon.gayle@ed.ac.uk
@Profbigvern
github.com/vernongayle

AQMEN

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

THE UNIVERSITY of EDINBURGH