Quantitative Methods

Week 1: Introduction to data analysis

AY 2023-24

Department of Political Economy

Instructor: Daniele Girardi

What to expect from this module

- Introduction to statistics, probability & econometrics.
- Analyze data to learn about the world.
  - Describe reality.
  - Test theories.
  - Evaluate the effect of policies.

Laptop Ban

The use of laptops, tablets, smartphones and similar devices is banned during this module's lectures

Research shows that the use of laptops in class harms learning and reduces students' grades.


THE NEW YORKER

News Culture Books Business & Tech Humor Cartoons Magazine Video Podcasts Archi

ANNALS OF TECHNOLOGY

THE CASE FOR BANNING LAPTOPS IN THE CLASSROOM

By Dan Rockmore June 6, 2014


SCIENTIFIC AMERICAN

Cart 0 Sign In

A Learning Secret: Don't Take Notes with a Laptop

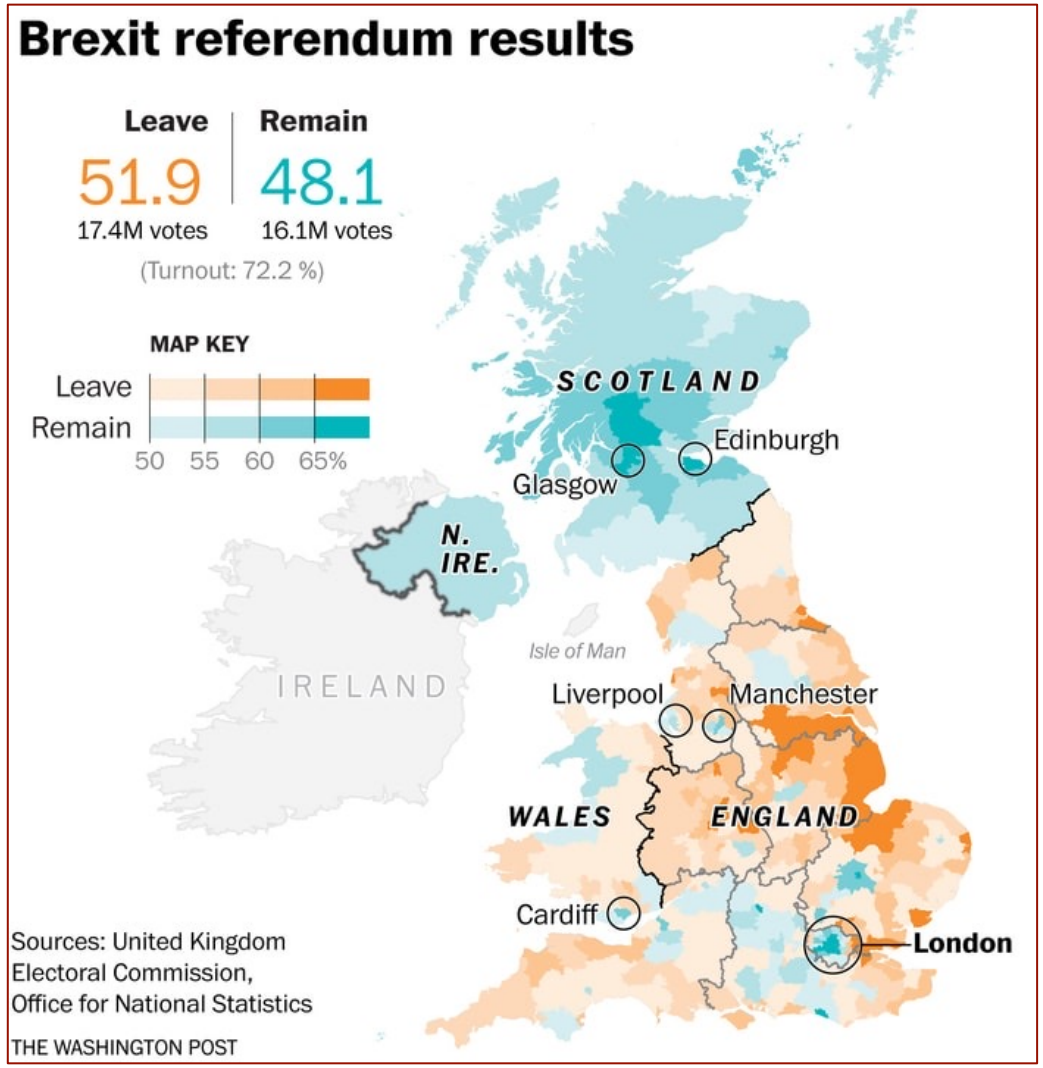
Students who used longhand remembered more and had a deeper understanding of the material


The New York Times

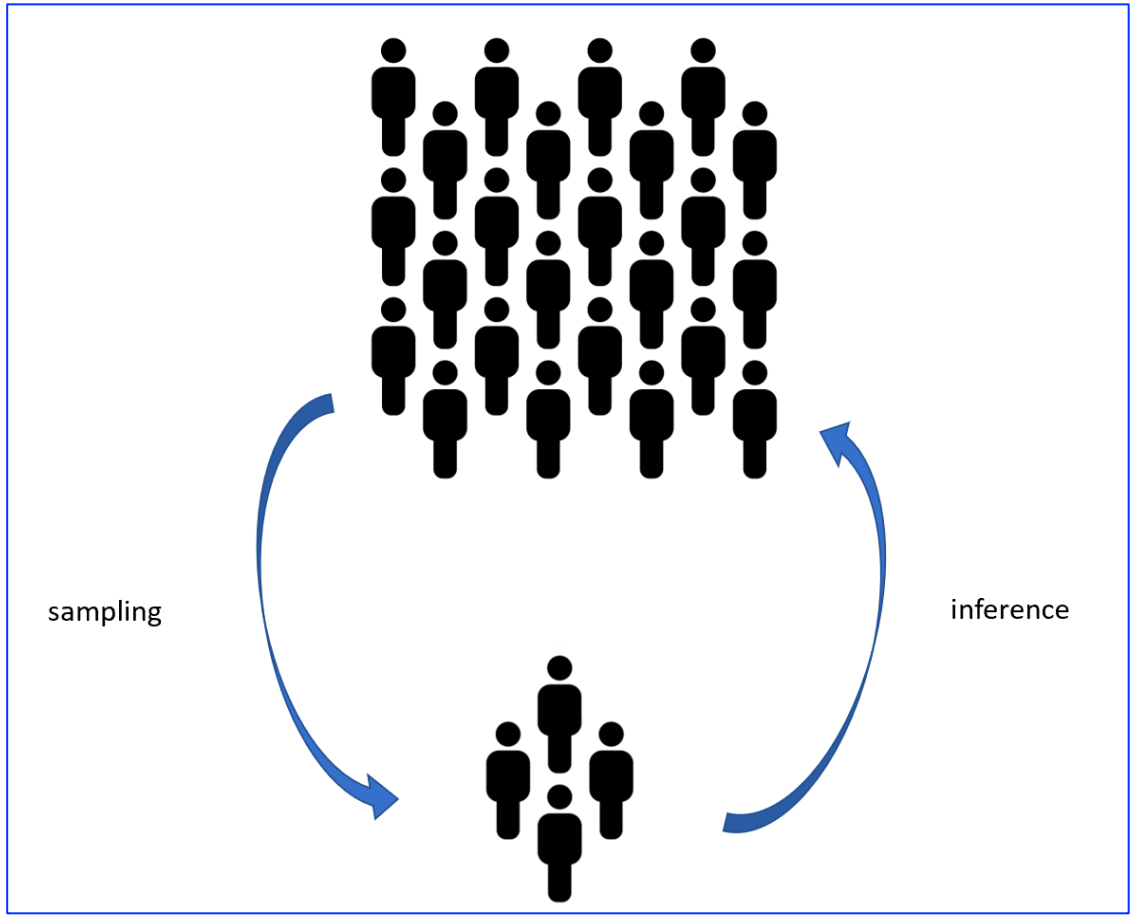
ECONOMIC VIEW

Laptops Are Great. But Not During a Lecture or a Meeting.

Descriptive statistics



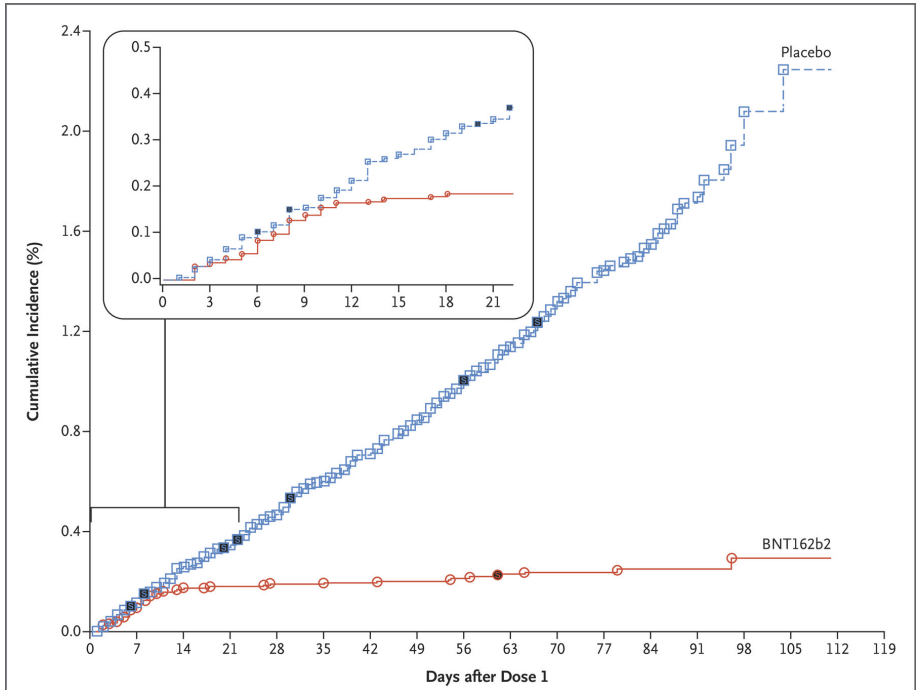
Statistical inference (or inferential statistics)



STATISTICAL INFERENCE: EVALUATING COVID VACCINES

Clinical trial for the Pfizer COVID vaccine

- 43,448 participants worldwide.
- ½ received vaccine, ½ a placebo.
- 9 infections among vaccinated vs. 172 in placebo group.
- *How likely is this difference to have arisen randomly, rather than reflecting an immunization effect of the vaccine?*
- *Statistical inference* answers this question, using the tools of *probability theory*.

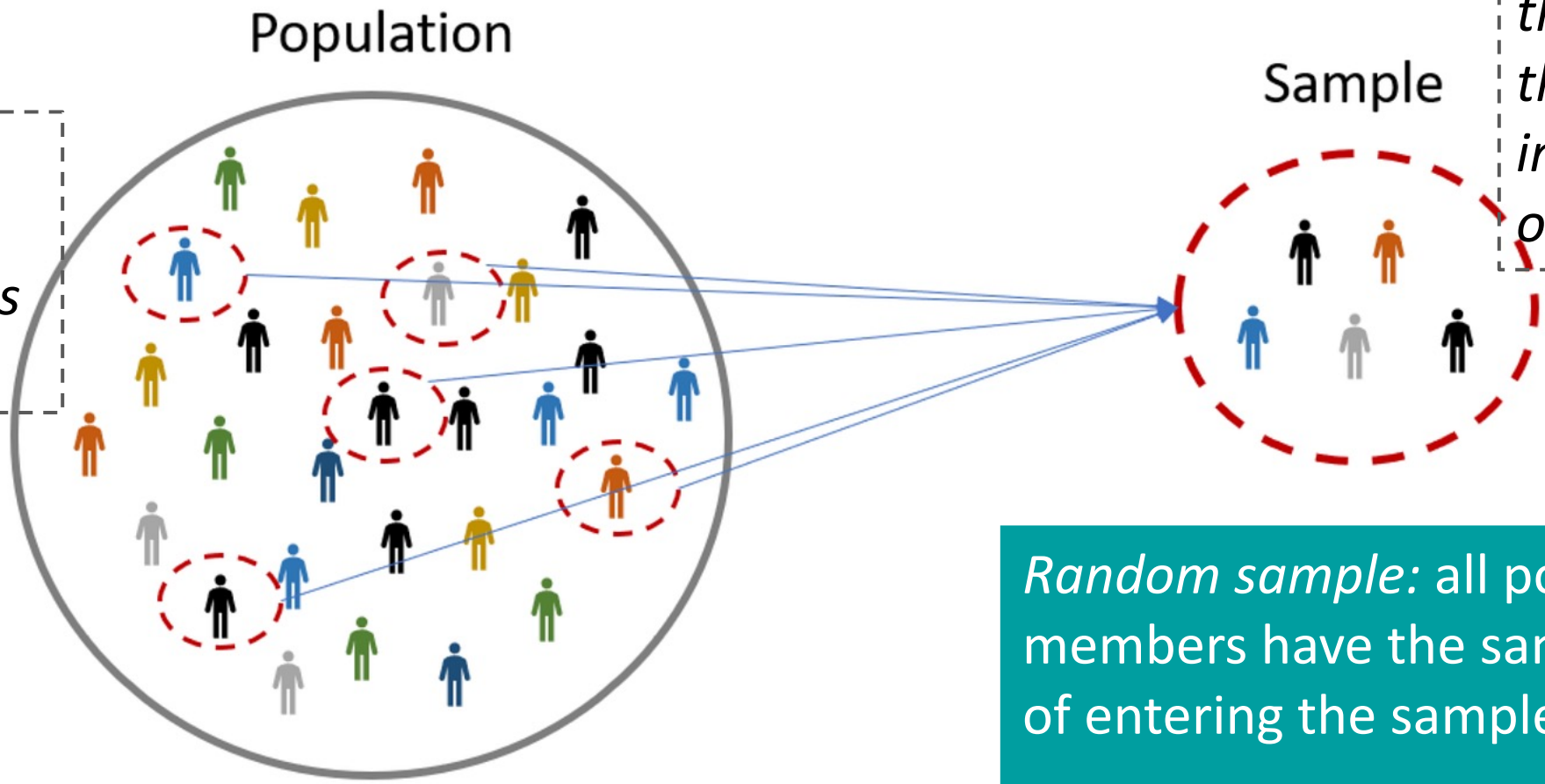


Efficacy End-Point Subgroup	BNT162b2, 30 μg (N=21,669)		Placebo (N=21,686)		VE (95% CI)
	No. of participants	Surveillance time person-yr (no. at risk)	No. of participants	Surveillance time person-yr (no. at risk)	percent
Covid-19 occurrence					
After dose 1	50	4.015 (21,314)	275	3.982 (21,258)	82.0 (75.6–86.9)
After dose 1 to before dose 2	39		82		52.4 (29.5–68.4)
Dose 2 to 7 days after dose 2	2		21		90.5 (61.0–98.9)
≥7 Days after dose 2	9		172		94.8 (89.8–97.6)

Source: Polack et al (2020) "Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine"

Statistics: population vs sample

Total collection of all individuals of interest.



A subset of the population that we have information on.

Random sample: all population members have the same chance of entering the sample.

Descriptive Statistics: the art of describing, summarizing & visualizing data

1. Frequency tables & graphs
2. Grouped data & histograms
3. Scatter diagrams (or scatterplots)

Datasets: basic structure

- **Observation**: one entry in the dataset.
- **Individuals** (or *units of observation*): the objects described by the dataset.
- **Variables**: characteristics that can take different values for different observations.
  - Typically, one 'special' variable indexes individuals.
  - Sometimes another 'special' variable represents time

(A subset of) The 2019 British Election Study (BES)

	finalserialno	b01	b02	b05
1	10102	No, did not vote	.	.
2	10103	No, did not vote	.	.
3	10105	Yes, voted	Labour Party	By post
4	10110	No, did not vote	.	.
5	10111	No, did not vote	.	.
6	10202	Yes, voted	Conservative Party	In person
7	10206	Yes, voted	Conservative Party	In person
8	10208	Yes, voted	Conservative Party	In person
9	10210	Yes, voted	Conservative Party	In person
10	10304	Yes, voted	Conservative Party	In person
11	10307	Yes, voted	Labour Party	In person
12	10309	Don't know	.	.
13	10310	Yes, voted	Conservative Party	In person
14	10402	Yes, voted	Liberal Democrats	In person
15	10407	Yes, voted	Labour Party	By post
16	10409	Don't know	.	.
17	10410	Yes, voted	Conservative Party	In person
18	10501	Yes, voted	Labour Party	In person
19	10503	Yes, voted	Labour Party	In person
20	10504	No, did not vote		

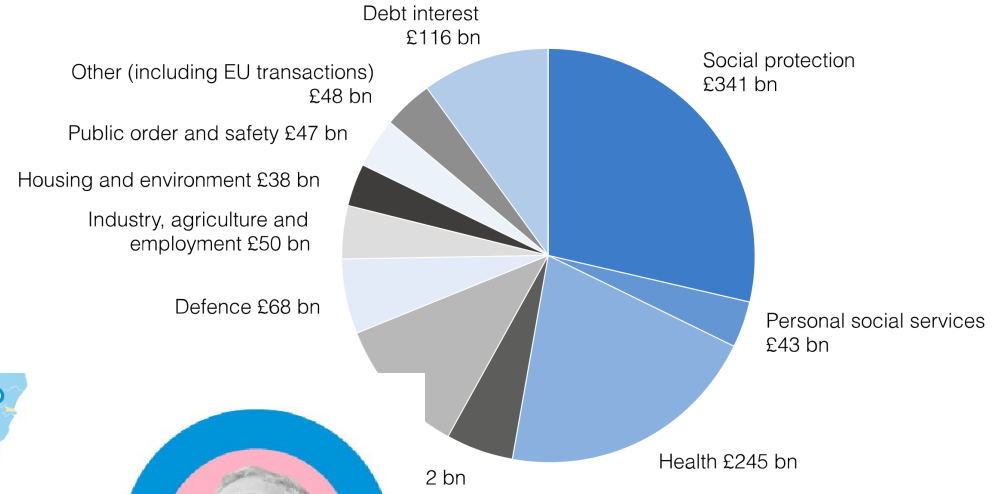
Which ones are the observations?

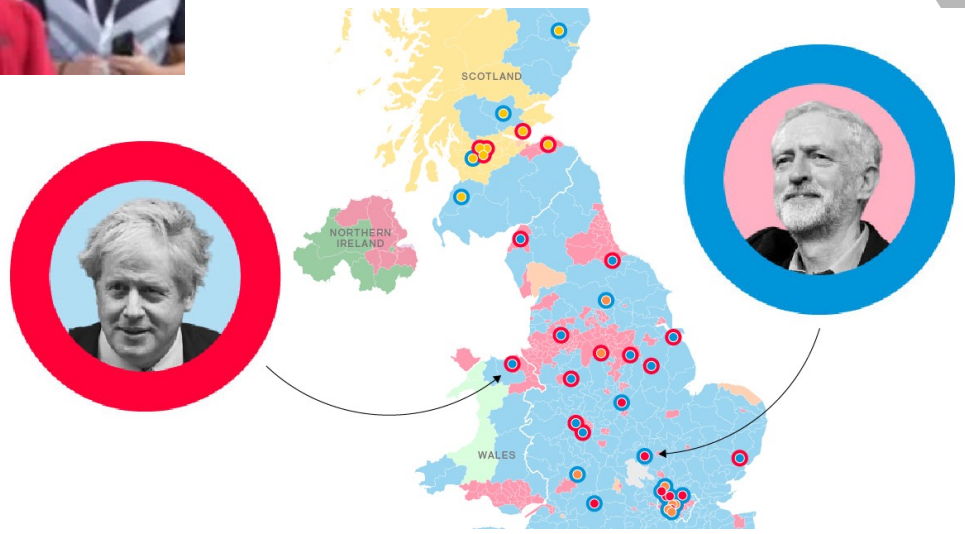
…the individuals?

…the variables?

https://www.britishelectionstudy.com/data-objects/cross-sectional-data/

1 – Frequency tables & graphs





FREQUENCY TABLE

Dataset: BES 2019

	finalserialno	b01	b02	b05
1	10102	No, did not vote	.	.
2	10103	No, did not vote	.	.
3	10105	Yes, voted	Labour Party	By post
4	10110	No, did not vote	.	.
5	10111	No, did not vote	.	.
6	10202	Yes, voted	Conservative Party	In person
7	10206	Yes, voted	Conservative Party	In person
8	10208	Yes, voted	Conservative Party	In person
9	10210	Yes, voted	Conservative Party	In person
10	10304	Yes, voted	Conservative Party	In person
11	10307	Yes, voted	Labour Party	In person
12	10309	Don't know	.	.
13	10310	Yes, voted	Conservative Party	In person
14	10402	Yes, voted	Liberal Democrats	In person
15	10407	Yes, voted	Labour Party	By post
16	10409	Don't know	.	.
17	10410	Yes, voted	Conservative Party	In person
18	10501	Yes, voted	Labour Party	In person
19	10503	Yes, voted	Labour Party	In person
20	10504	No, did not vote		

Frequency table for vote method:

b05 (vote method)	Frequency
In person	2,410
By post	689
By proxy	21
Prefer not to say	10
Missing value	816

in STATA: `tab b05, missing`

FREQUENCY & RELATIVE FREQUENCY

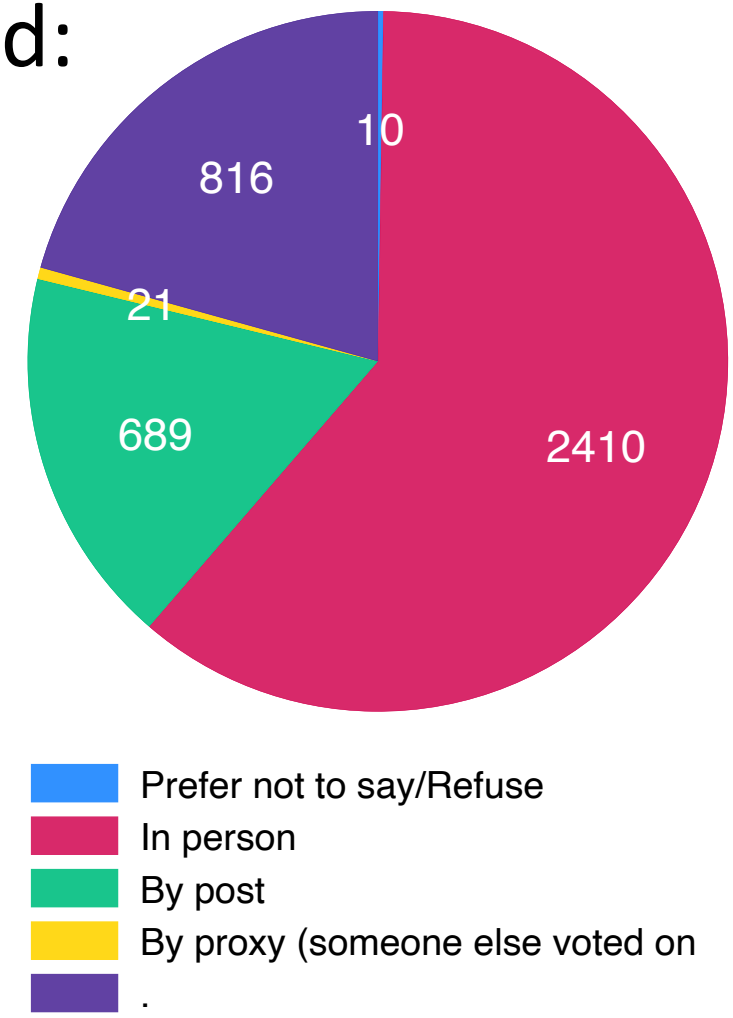
b05 (vote method)	Frequency	Relative frequency (=freq/3,946)
In person	2,410	61.1%
By post	689	17.4%
By proxy	21	0.5%
Prefer not to say	10	0.2%
Missing value	816	20.7%

THE PIE CHART

Frequency table for vote method:

b05 (vote method)	Frequency	Relative frequency
In person	2,410	61.1%
By post	689	17.4%
By proxy	21	0.5%
Prefer not to say	10	0.2%
Missing value	816	20.7%

Pie Chart for vote method:



MASTER GOLF TOURNAMENTS 1968-2004

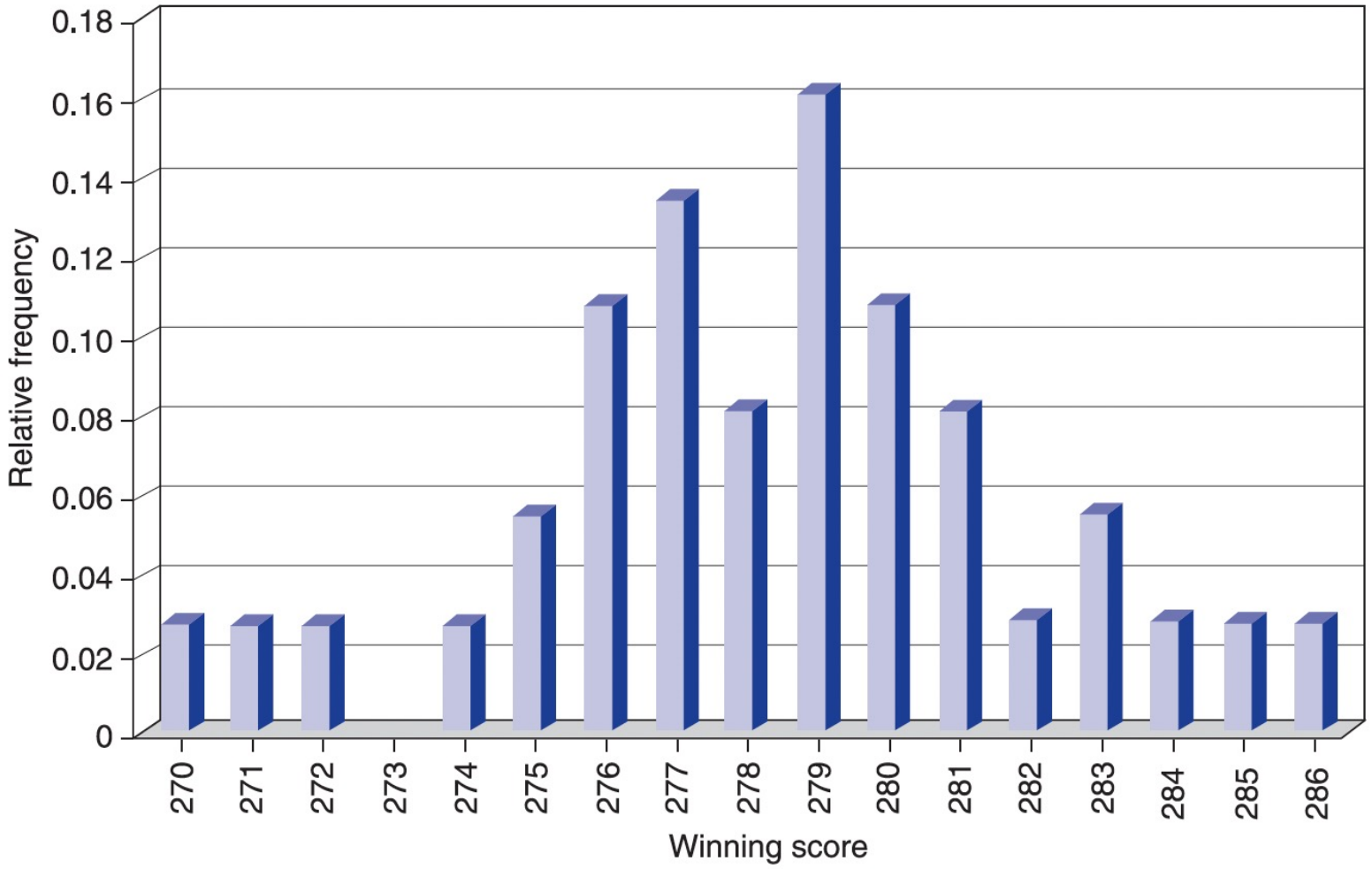
Dataset

Year	Winner	Score	Year	Winner	Score
1968	Bob Goalby	277	1987	Larry Mize	285
1969	George Archer	281	1988	Sandy Lyle	281
1970	Billy Casper	279	1989	Nick Faldo	283
1971	Charles Coody	279	1990	Nick Faldo	278
1972	Jack Nicklaus	286	1991	Ian Woosnam	277
1973	Tommy Aaron	283	1992	Fred Couples	275
1974	Gary Player	278	1993	Bernhard Langer	277
1975	Jack Nicklaus	276	1994	J.M. Olazabal	279
1976	Ray Floyd	271	1995	Ben Crenshaw	274
1977	Tom Watson	276	1996	Nick Faldo	276
1978	Gary Player	277	1997	Tiger Woods	270
1979	Fuzzy Zoeller	280	1998	Mark O'Meara	279
1980	Severiano Ballesteros	275	1999	J.M. Olazabal	280
1981	Tom Watson	280	2000	Vijay Singh	278
1982	Craig Stadler	284	2001	Tiger Woods	272
1983	Severiano Ballesteros	280	2002	Tiger Woods	276
1984	Ben Crenshaw	277	2003	Mike Weir	281
1985	Bernhard Langer	282	2004	Phil Mickelson	279
1986	Jack Nicklaus	279			

Frequency table for winning scores

Winning score	Frequency f	Relative frequency $f/37$
270	1	0.027
271	1	0.027
272	1	0.027
274	1	0.027
275	2	0.054
276	4	0.108
277	5	0.135
278	3	0.081
279	6	0.162
280	4	0.108
281	3	0.081
282	1	0.027
283	2	0.054
284	1	0.027
285	1	0.027
286	1	0.027

RELATIVE FREQUENCY BAR GRAPH



Shows the *distribution* of the variable

This distribution is quite *symmetric*

DATASET: UK government spending in 2023

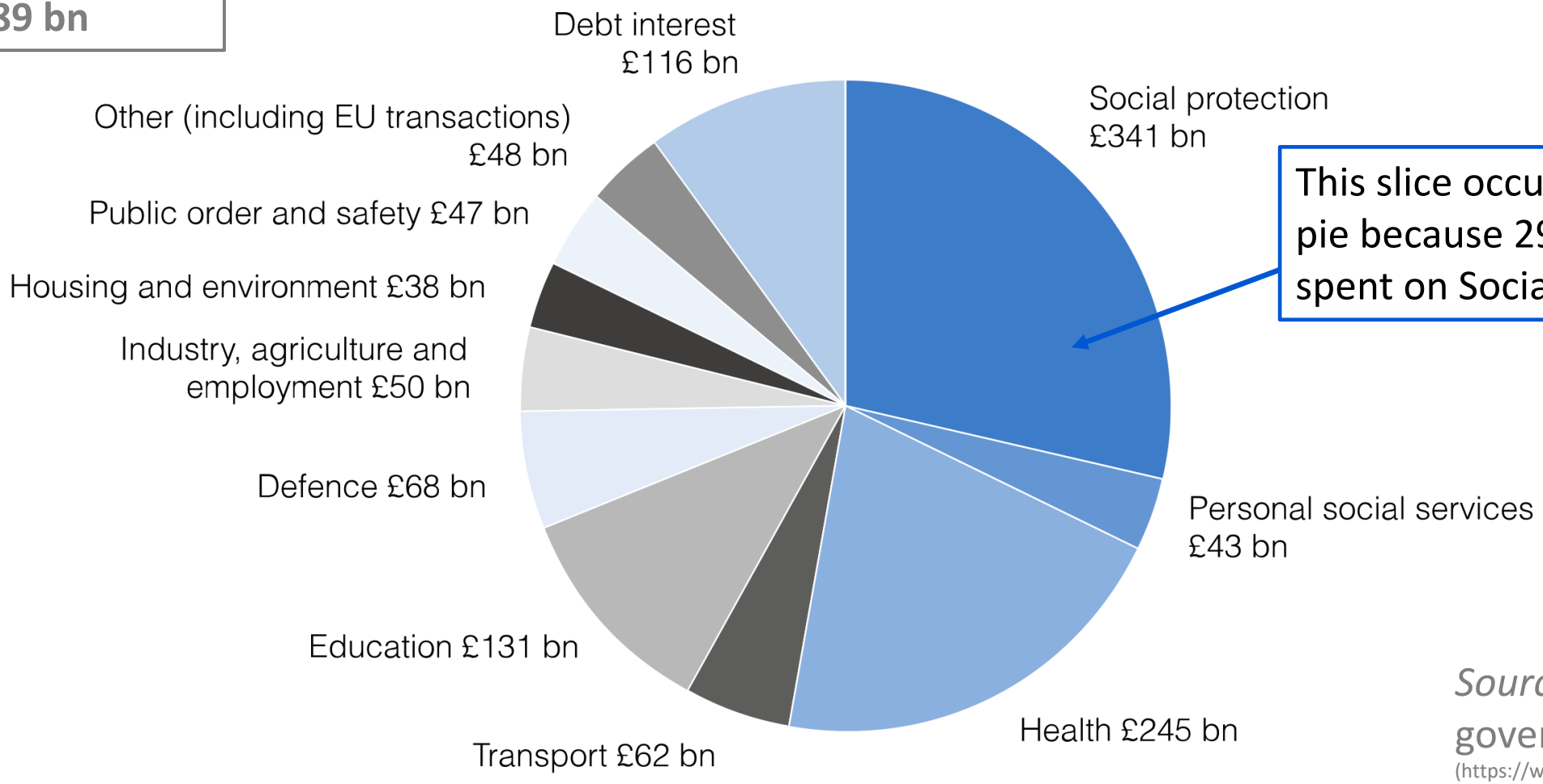
Government function	Spending (£bn)
Social protection	341
Health	245
Education	131
Debt interest	116
Defence	68
Transport	62
Industry, agriculture & employment	50
Public order & safety	47
Personal social services	43
Housing and environment	38
Other (including EU transactions)	48

- Not a frequency table…
- …but we can still use a pie chart to represent it!

Source: UK government (https://www.gov.uk/government/publications/spring-budget-2023/spring-budget-2023-html)

PIE CHART: UK government spending in 2023

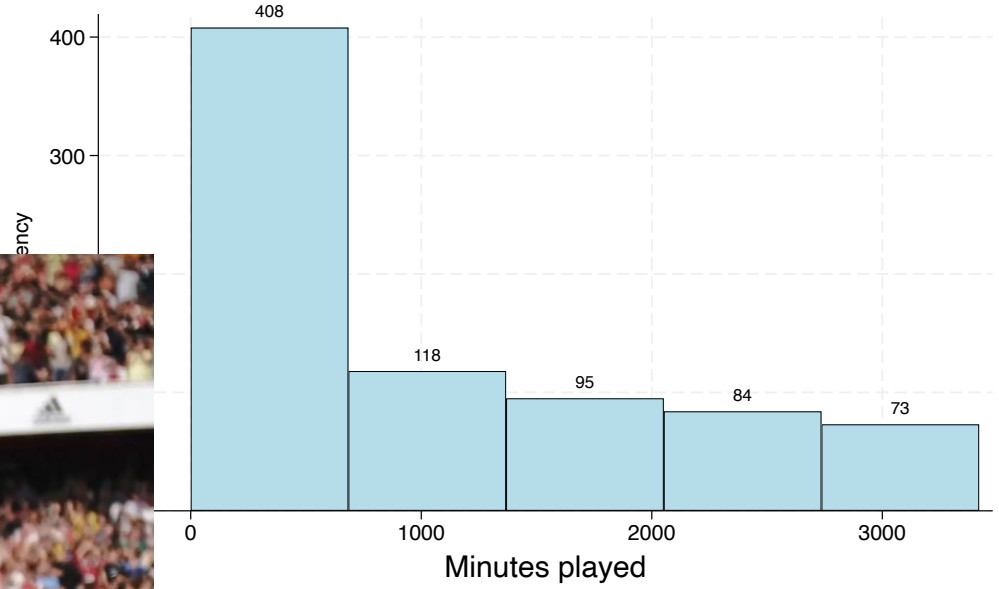
Total spending:
£1,189 bn



This slice occupies 29% of the pie because 29% of money was spent on Social Protection

Source: UK government (https://www.gov.uk/government/publications/spring-budget-2023/spring-budget-2023-html)

2 – Grouped data & histograms



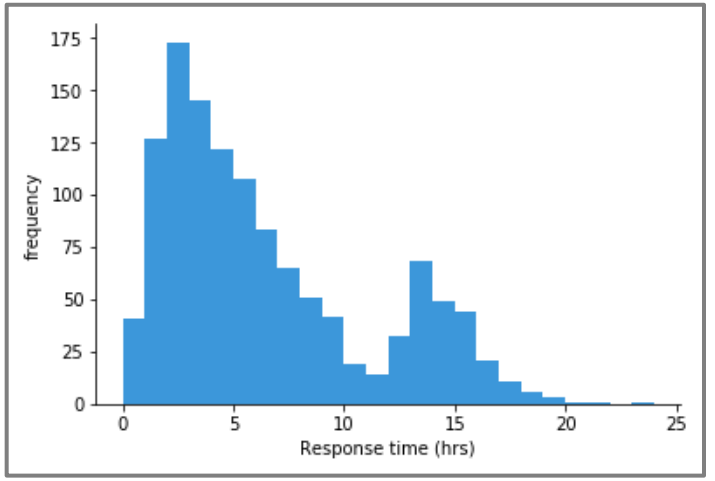
Dataset: 2020-21 PL player statistics

	first_name	second_name	position	goals_scored	assists	minutes	red_cards	yellow_…	t
1	Granit	Xhaka	MID	7	8	2992	0	4	
2	Mohamed	Elneny	MID	0	0	111			
3	Rob	Holding	DEF	1	0	562			
4	Thomas	Partey	MID	3	0	2480			
5	Martin	Ãdegaard	MID	15	8	3132			
6	Kieran	Tierney	DEF	0	1	774			
7	Nicolas	PÃ©pÃ©	MID	0	0	0			
8	Benjamin	White	DEF	2	5	3054			
9	Eddie	Nketiah	FWD	4	2	1071			
10	Emile	Smith Rowe	MID	0	2	161			
11	Bukayo	Saka	MID	14	12	3183			
12	Takehiro	Tomiyasu	DEF	0	1	651			
13	Aaron	Ramsdale	GK	0	0	3420			
14	Gabriel	dos Santos MagalhÃ£es	DEF	3	0	3409			
15	Nuno	Varela Tavares	DEF	0	0	0	0		
16	Gabriel	Martinelli Silva	MID	15	9	2789	0	3	
17	Pablo	MarÃ Villar	DEF	0	0	0	0	0	
18	Lucas	Torreira di Pascua	MID	0	0	0	0	0	
19	Reiss	Nelson	MID	3	3	202	0	0	
20	Matt	Turner	GK	0	0	0	0	0	

- 778 players
- We want to plot minutes played.
- Too many distinct values for a frequency table/graph!

Building a histogram: 3 steps

1. Choose classes of values (or *bins*)
   - Divide the range of the data into classes of equal width.
2. Count the observations in each bin
3. Draw the histogram
   - Like a bar chart
   - But each bar is a bin, not a single value.



Building a histogram of minutes played in the 2020-21 Premier League

STEP 1 – *Choose classes of values (or bins)*

- Group players in classes (or *bins*) based on minutes played.
- Minutes range from 0 to 3,420
- Can do five bins of width $3{,}420/5=684$:

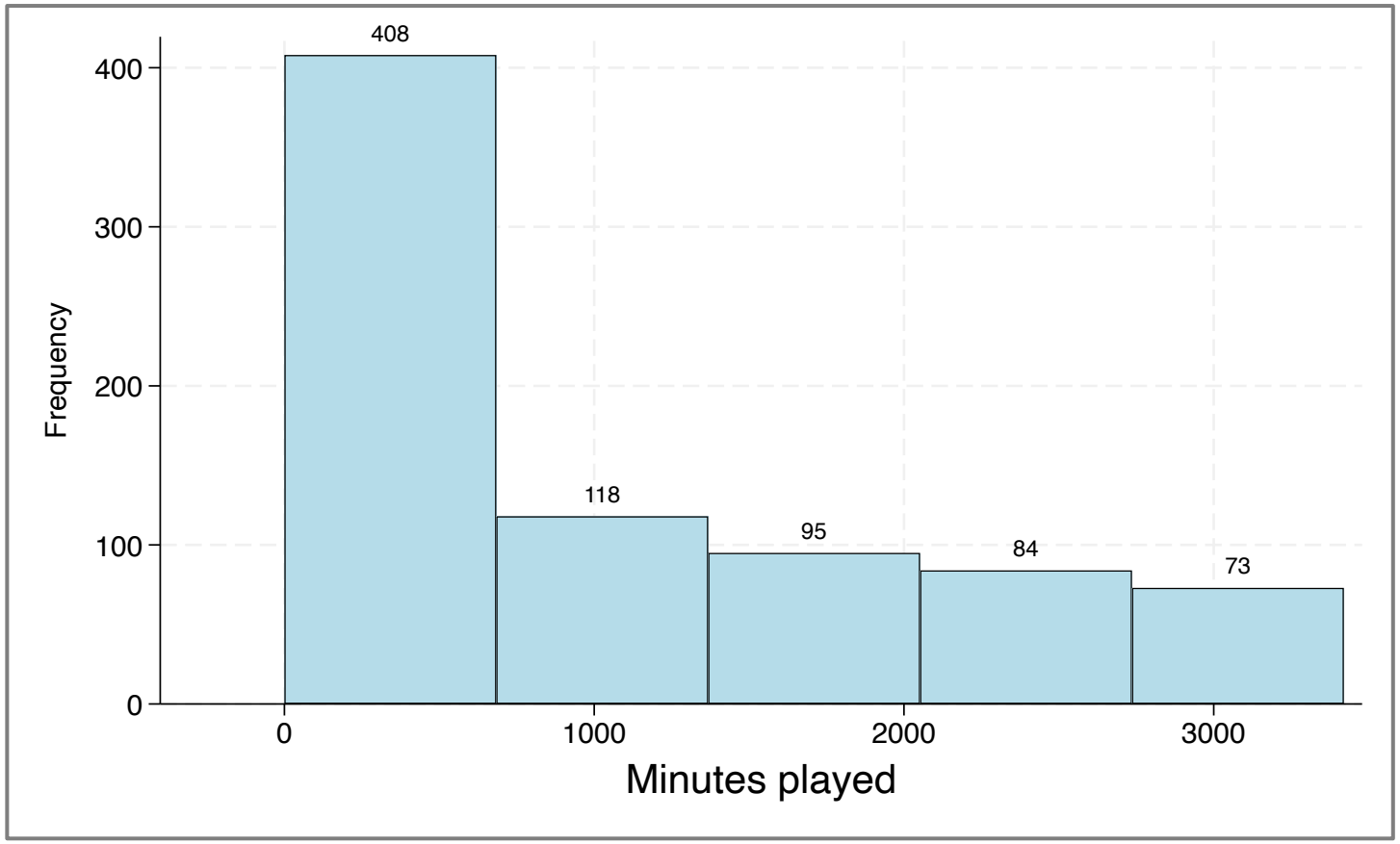
I. Less than 684 minutes [minutes < 684]
II. 684 to 1,368 minutes [$684 \leq$ minutes $< 1{,}368$]
III. 1,368 to 2,052 minutes [$1{,}368 \leq$ minutes $< 2{,}052$]
IV. 2,052 to 2,736 minutes [$2{,}052 \leq$ minutes $< 2{,}736$]
V. 2,736 or more minutes [minutes $\geq 2{,}736$]

STEP 2 – *Count the observations in each bin*

Bin	Range	Players
1	Less than 684 minutes	408
2	684 to 1,368	118
3	1,368 to 2,052	95
4	2,052 to 2,736	84
5	2,736 or more	73

STEP 3 – *Make the histogram*

Bin	Range	Players
1	Less than 684	408
2	684 to 1,368	118
3	1,368 to 2,052	95
4	2,052 to 2,736	84
5	More than 2,736	73



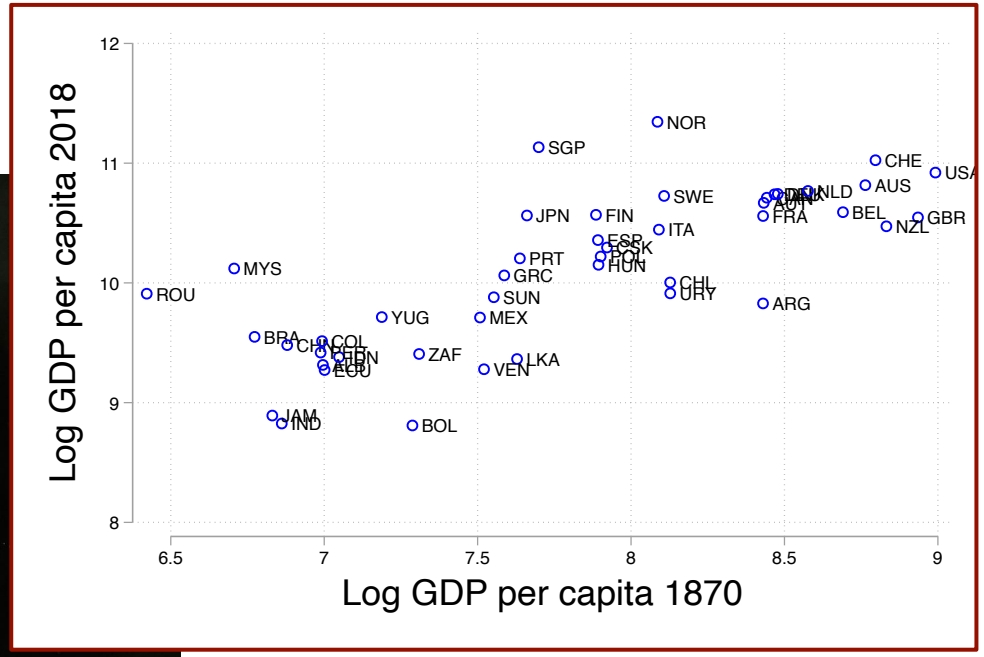
in STATA:

```
histogram minutes,
bins(5) frequency
```

Note: 2,736 = around 30 full games

684 = around 7 full games

3 – Scatter diagrams (or *scatterplots*)



Dataset: goals and minutes played for forwards in 2020-21 PL

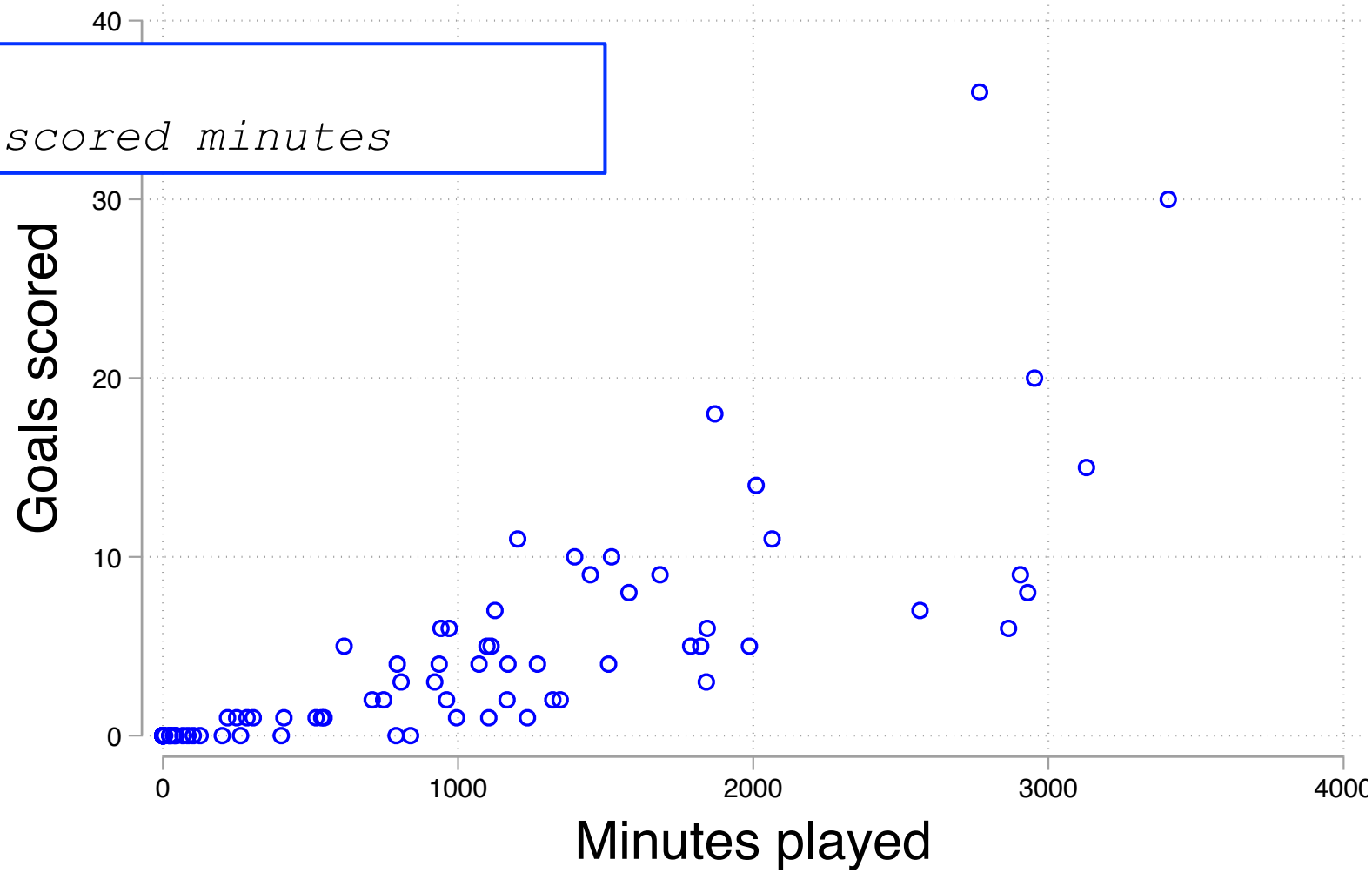
	first_name	second_name	goals_scored	minutes
1	Eddie	Nketiah	4	1071
2	Gabriel	Fernando de Jesus	11	2064
3	Nathan	Butler-Oyedeji	0	0
4	Ollie	Watkins	15	3129
5	Keinan	Davis	0	0
6	Cameron	Archer	0	43
7	Jhon	DurÃ¡n	0	126
8	Kieffer	Moore	4	1269
9	Dominic	Solanke	6	2865
10	Christian	Saydee	0	0
11	Daniel	Adu-Adjei	0	0
12	Dominic	Sadi	0	0
13	Euan	Pollock	0	0
14	Antoine	Semenyo	1	250
15	Ivan	Toney	20	2953
16	Bryan	Mbeumo	9	2905
17	Marcus	Forss	0	0
18	Halil	DerviÅoÄlu	0	6
19	Danny	Welbeck	6	1844
20	Deniz	Undav	5	614

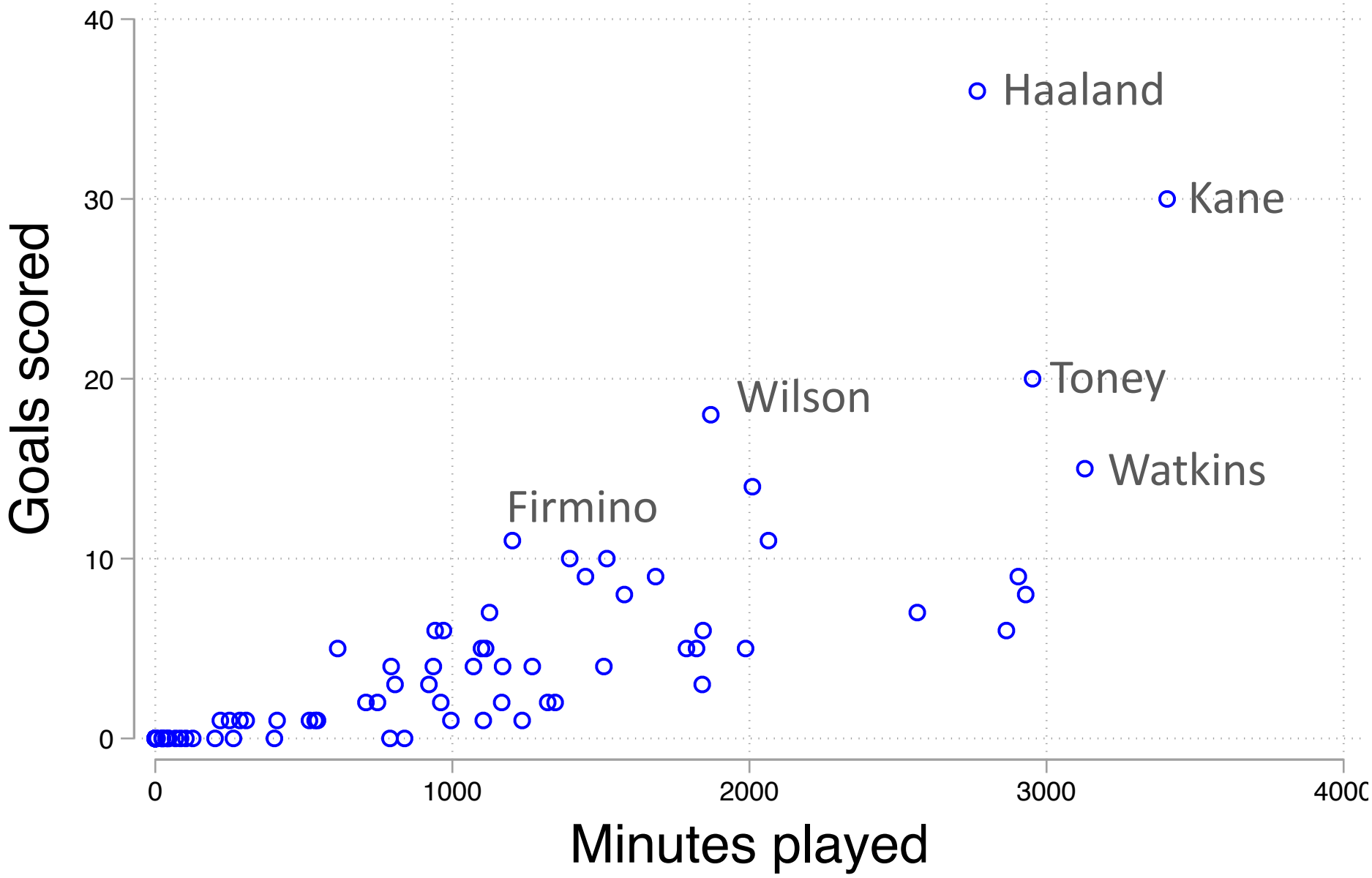
Do forwards that play more minutes score more goals?

SCATTERPLOT: Goals vs Minutes Played

in STATA:

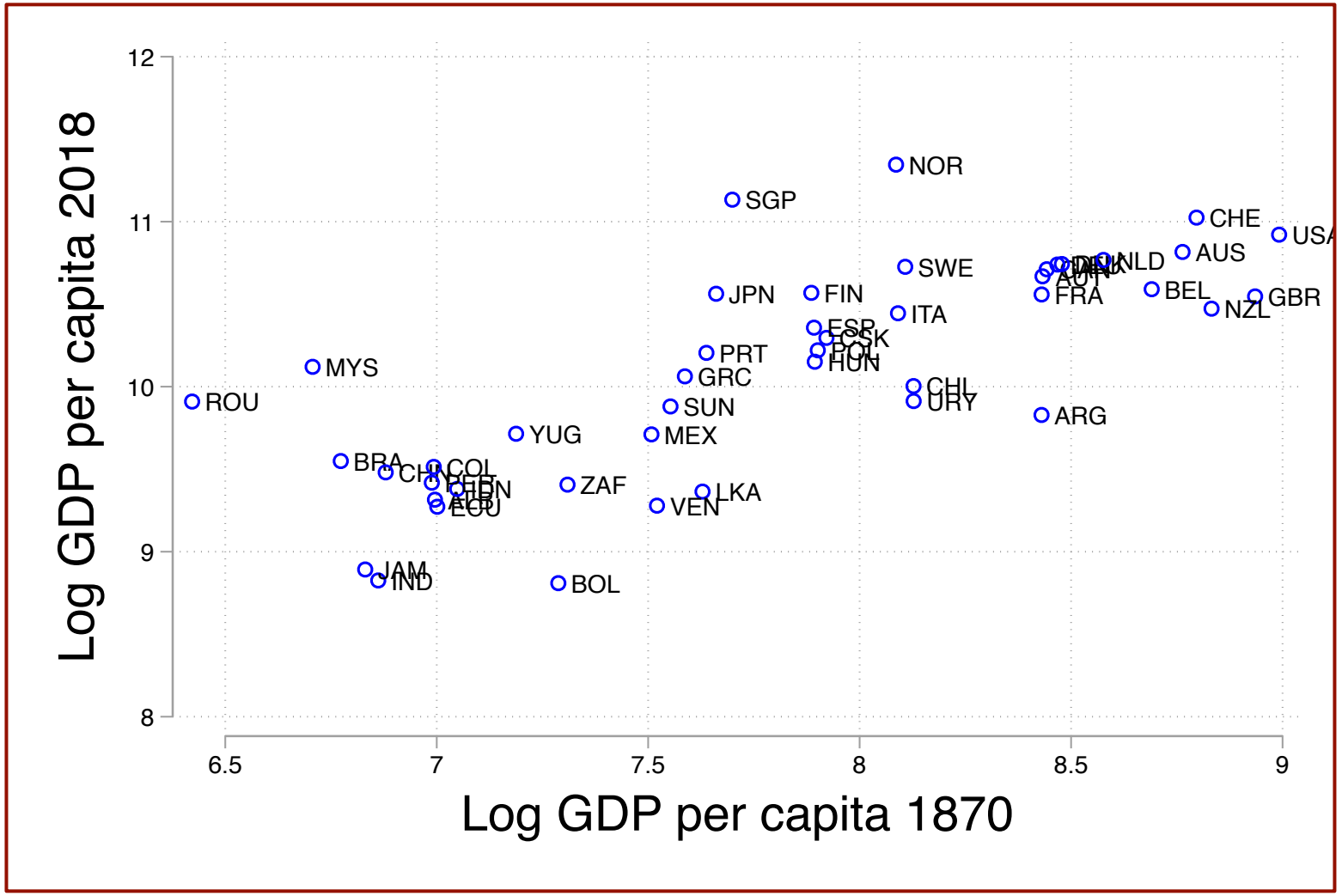
```
scatter goals_scored minutes
```




Haaland
Kane
Toney
Wilson
Watkins
Firmino
Goals scored
Minutes played
40
30
20
10
0
0
1000
2000
3000
4000

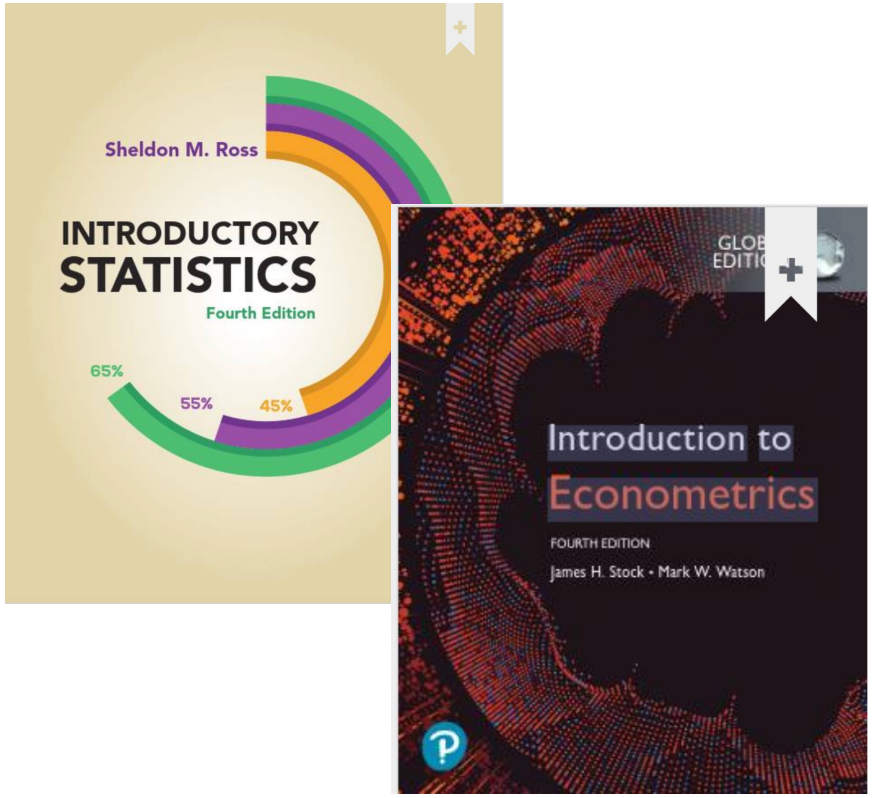
SCATTERPLOT: Persistence in GDP per capita

	country	gdppc1870	gdppc2018
3	Albania	6.5666724	9.3150757
5	Argentina	7.7579062	9.8285691
7	Australia	8.5596778	10.816389
8	Austria	7.9963172	10.668678
11	Belgium	8.3642751	10.590521
15	Bulgaria	7.1996783	9.8225085
20	Brazil	6.9884132	9.5492073
24	Canada	7.9017475	10.711497
25	Switzerland	7.9910429	11.024721
26	Chile	7.5326236	10.003548
27	China	6.8511849	9.4804978
32	Colombia	6.9828628	9.5137764
36	Czechoslovakia	7.52564	10.29555
40	Germany	7.9830989	10.740251
43	Denmark	8.0687162	10.743164
45	Algeria	7.0387835	9.5629689
46	Ecuador	6.6333184	9.2722653
47	Egypt	7.0859015	9.3890899
48	Spain	7.5005295	10.357632
51	Finland	7.5049421	10.568665
52	France	8.0030287	10.558827
54	United Kingdom	8.6706007	10.546869
56	Ghana	6.5510803	8.3586819



About the module

- **Topics**
  1. Introduction
  2. Using statistics to summarize datasets
  3. Probability
  4. Statistics
  5. Econometrics (Linear Regression)



- **Module assessment**
  1. Participation (10%) – this means coming to seminars & labs
  2. Coursework (30%) – an essay/coding project due on March 28
  3. Final Exam (60%)

Next reading

- Chapter 3 of the Sheldon Ross textbook
  - "Using statistics to summarize data sets"
- Read before next Monday's lecture

Thank you for your attention