Quantitative Methods

Week 2: Using statistics to summarize datasets (aka: “summary statistics”)

AY 2023-24

Department of Political Economy

Instructor: Daniele Girardi

Week 2 – Using statistics to summarize datasets

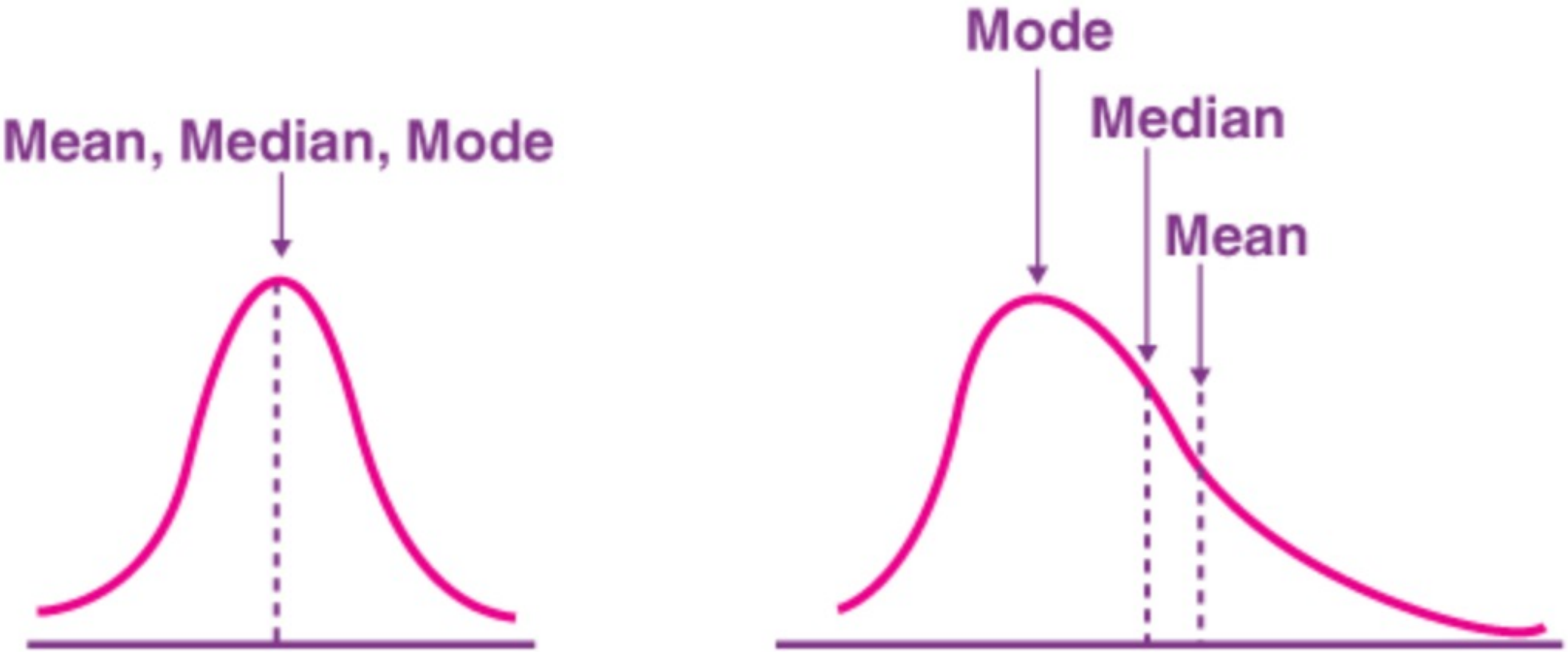
The plan

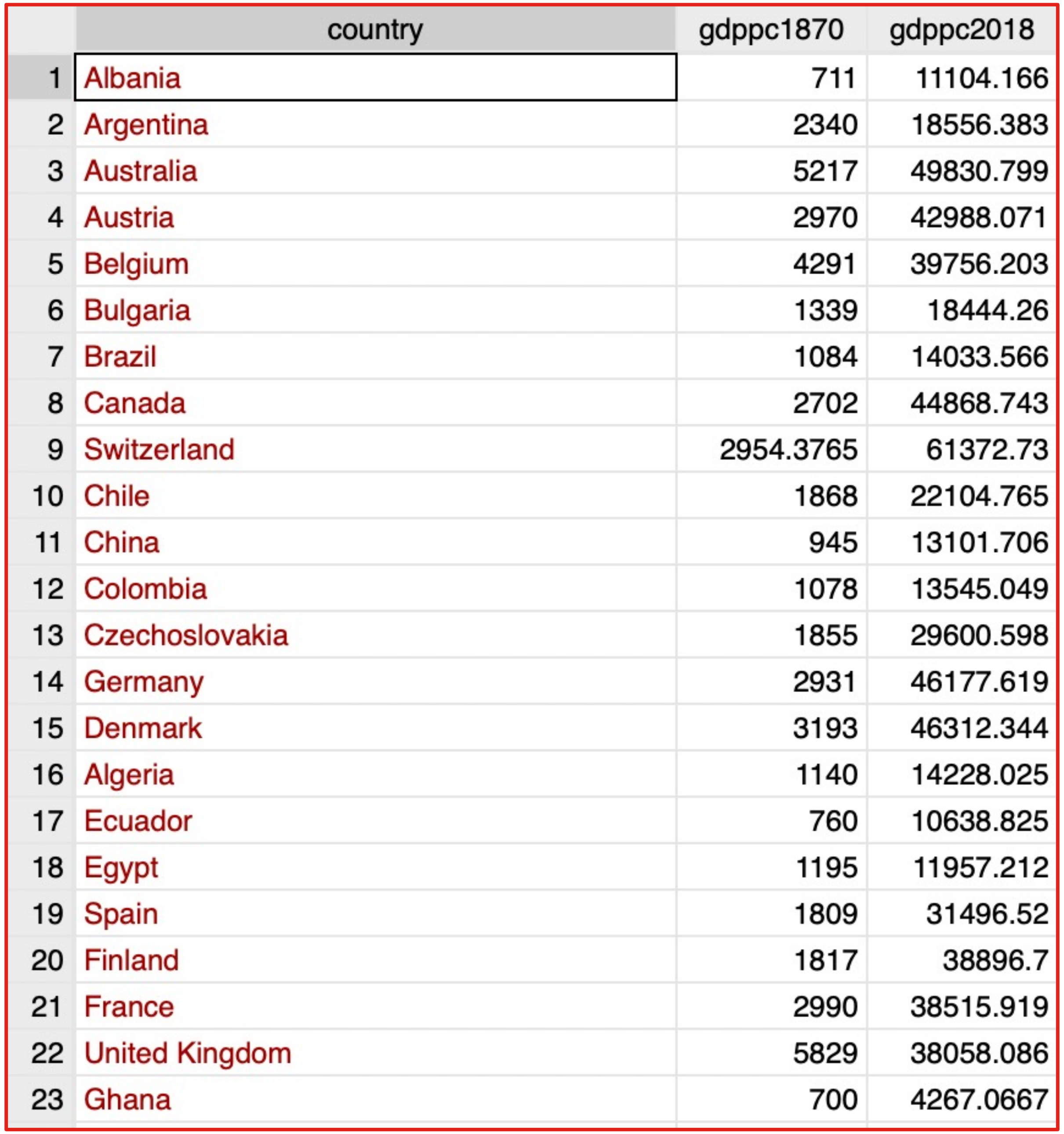
1. Sample mean, mode & median
2. Quartiles & Percentiles
3. Sample variance & standard deviation
4. Two variables: the sample correlation coefficient
5. “Normal” variables

Summary statistics

- Graphs are great, but sometimes we want a *numerical* summary.
- *Statistic* = a numerical quantity computed from a dataset.
- *Summary statistics* describe relevant features of the data.

1 - Sample mean, mode & median





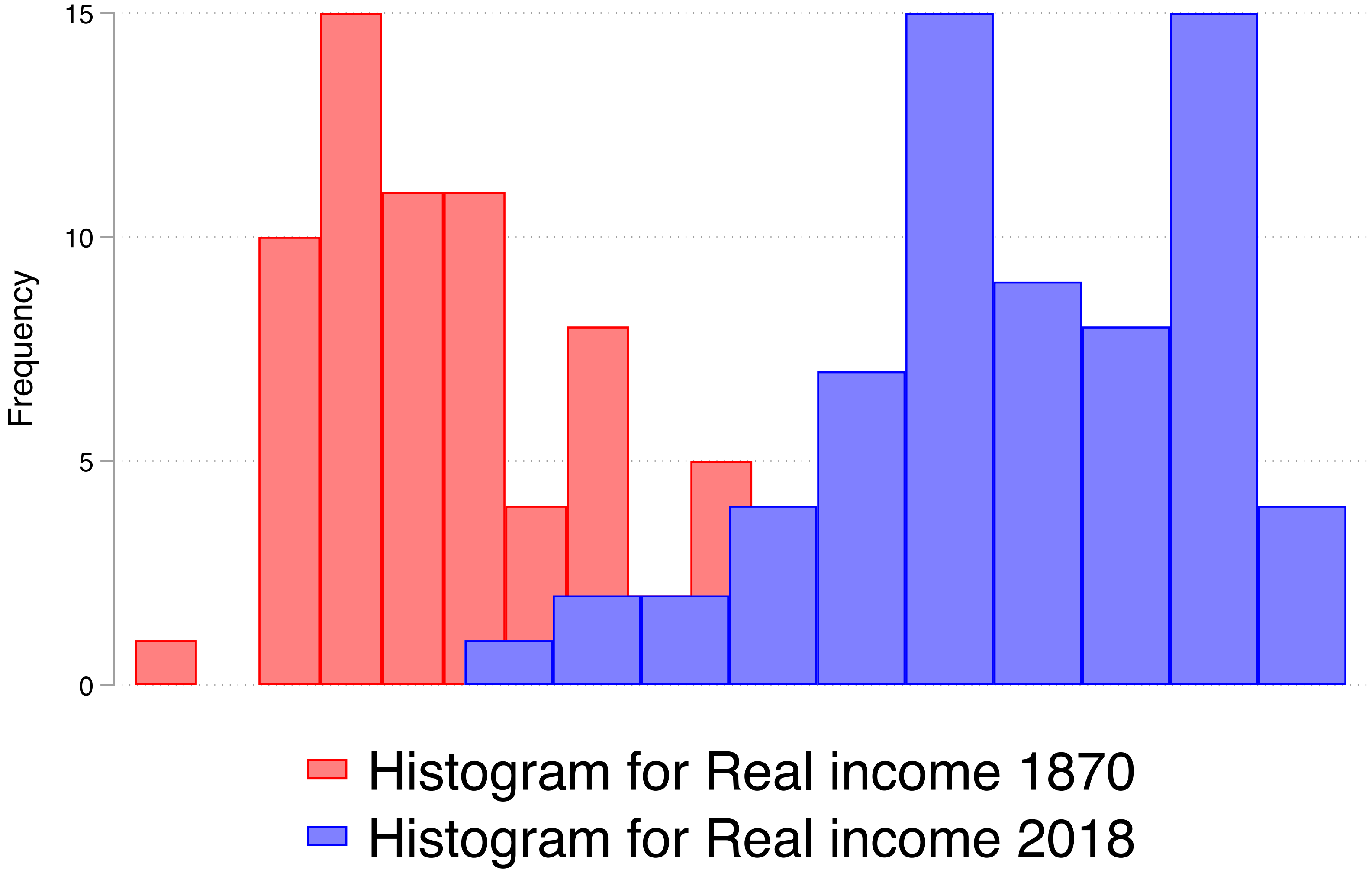
	country	gdppc1870	gdppc2018
1	Albania	711	11104.166
2	Argentina	2340	18556.383
3	Australia	5217	49830.799
4	Austria	2970	42988.071
5	Belgium	4291	39756.203
6	Bulgaria	1339	18444.26
7	Brazil	1084	14033.566
8	Canada	2702	44868.743
9	Switzerland	2954.3765	61372.73
10	Chile	1868	22104.765
11	China	945	13101.706
12	Colombia	1078	13545.049
13	Czechoslovakia	1855	29600.598
14	Germany	2931	46177.619
15	Denmark	3193	46312.344
16	Algeria	1140	14228.025
17	Ecuador	760	10638.825
18	Egypt	1195	11957.212
19	Spain	1809	31496.52
20	Finland	1817	38896.7
21	France	2990	38515.919
22	United Kingdom	5829	38058.086
23	Ghana	700	4267.0667

Maddison Project Dataset

(https://www.rug.nl/ggdc/historicaldevelopment/maddison/releases/maddison-project-database-2020)

- Historical income data for a sample of world countries.
- Real income measured in 2011 $.

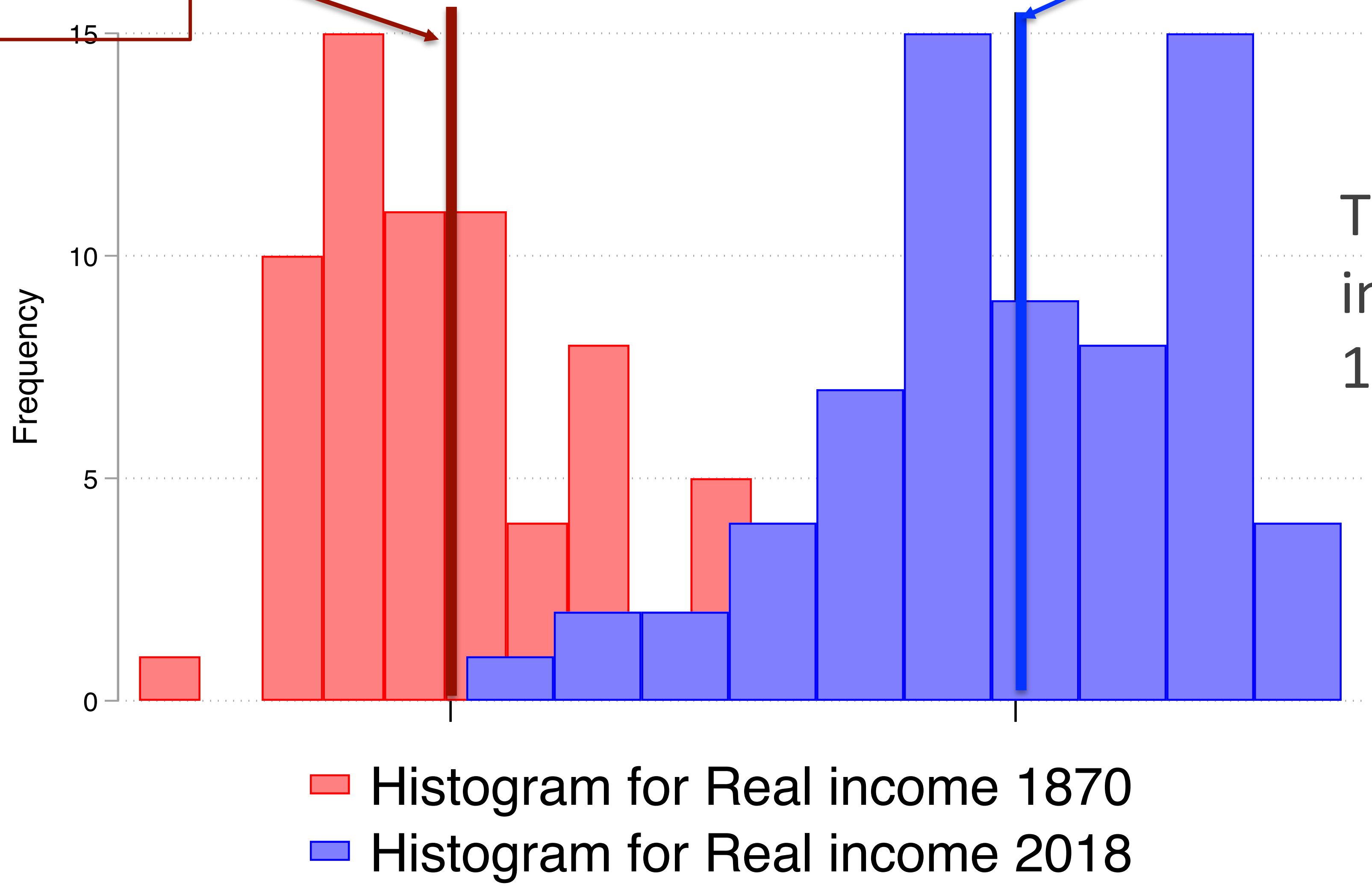
Overall, did countries get richer between 1870 and 2018? How much?



Note: Here the data has been transformed in its natural logarithm for better visualisation…but you don't need to worry about that for now.

Mean of Real Income 1870: $ 1,796

Mean of Real Income 2018: $ 25,908



The average increased by 13 times!

Sample mean, mode & median

- Alternative measures of the *center* of the data
- **Sample mean**: the *center of gravity.*
- **Sample median**: the *midpoint*.
- **Sample mode**: the *most frequent* outcome.

Sample mean (or average)

- The *center of gravity* of the dataset.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- **Deviation** from the mean for observation i:

$$x_i - \bar{x}$$

- **Property:** *The sum of the positive deviations exactly balances the sum of negative deviations.*

Sample mean & frequency table

Value	Frequency
3	2
4	1
5	3

$$\text{Mean} =$$

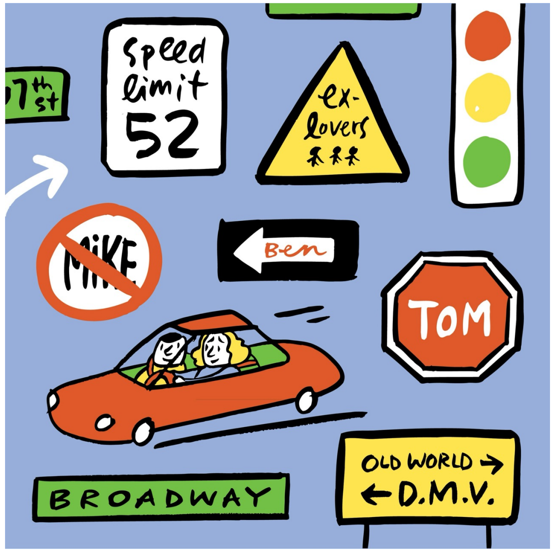
$$= \frac{(3\text{x}2) + (4\text{x}1) + (5\text{x}3)}{2 + 1 + 3}$$

$$= \frac{6 + 4 + 15}{6} = \frac{25}{6} = 4.17$$

Sample mean & outliers

- **Dataset:** Number of weeks to obtain driver's license after learn-to-drive course (n=7):

2, 110, 5, 7, 6, 7, 3



- Your turn: compute the sample mean.
- Mean = 140/7= 20
- With outliers, the sample mean can be misleading!
- In such cases, we want another way to measure the center of the distribution, less affected by extreme values (or outliers).

Sample median

- The *midpoint* of the dataset.
- The number such that half the observations are smaller and the other half are larger.
- Computing the median:
  1. Order the data values from smallest to largest
  2. If n is odd, median is the middle value
     - the [(n + 1)/2]-th observation
  3. If n is even, median = average of the 2 middle values

Sample median

Dataset: Number of weeks to obtain driver's license after learn-to-drive course (n=7):

2, 110, 5, 7, 6, 7, 3

Your turn: compute the sample median.

1. Order the observations from smallest to largest
   - 2, 3, 5, 6, 7, 7, 110
2. Middle value = 4th observation = **6**

- Often more sensible than the mean with big outliers!

Sample mode

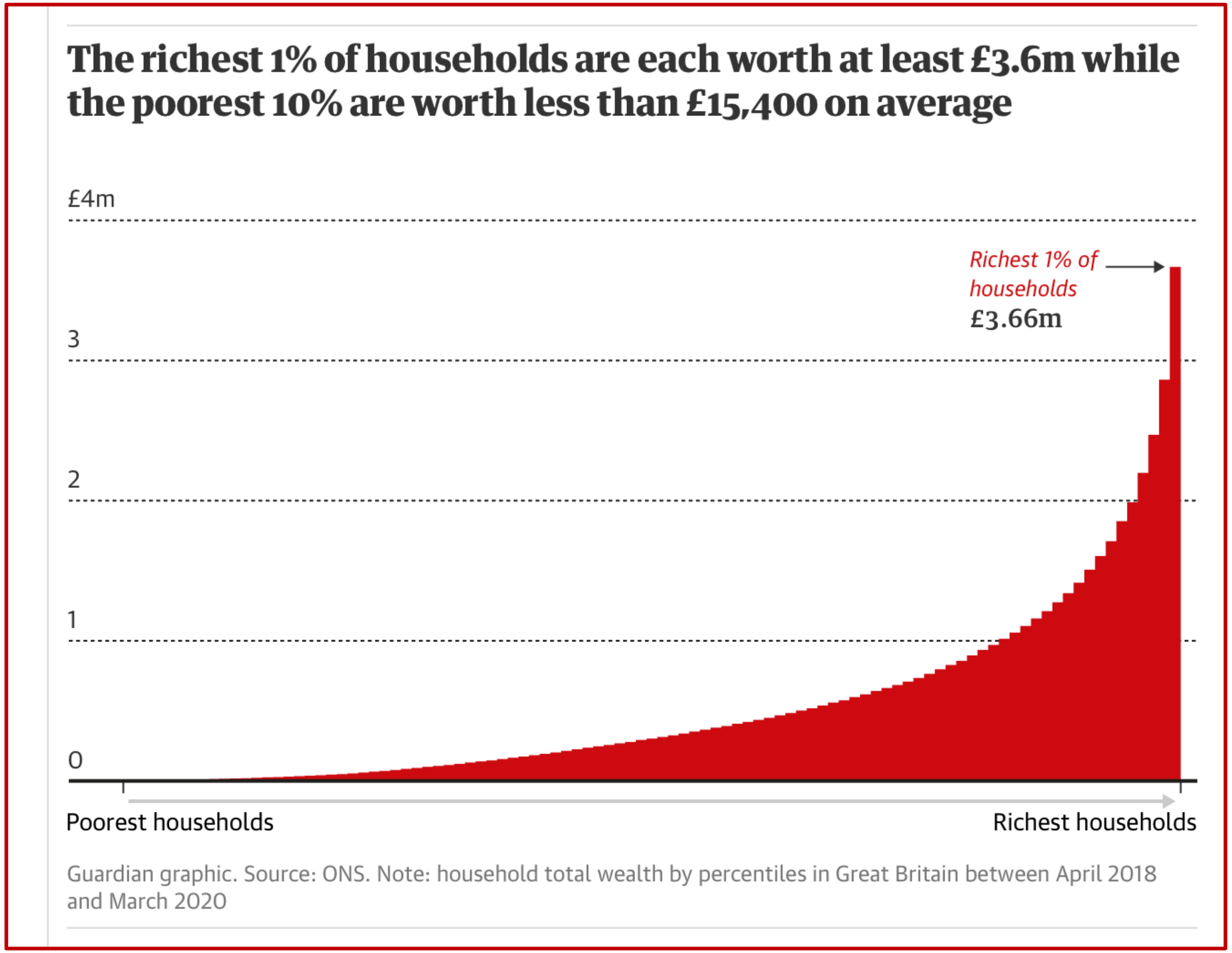
- **Sample mode**: the *most frequent* value.
- Sometimes there is not a unique sample mode, but two or more *modal values*.
- There isn't much to add! ☺

Learn-to-drive dataset again:

2, 110, 5, 7, 6, 7, 3

What is the sample mode?

2 - Quartiles & Percentiles



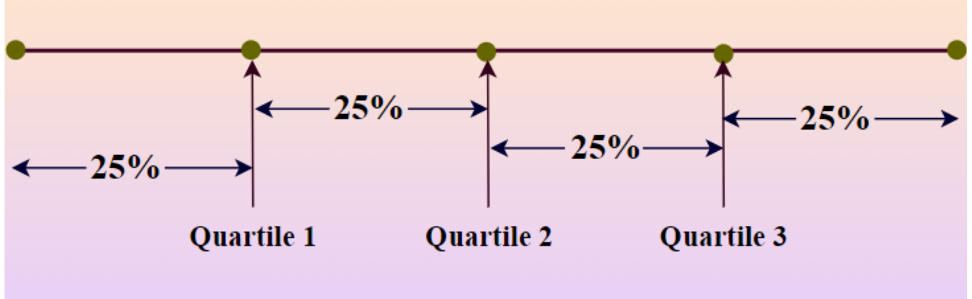
The richest 1% of households are each worth at least £3.6m while the poorest 10% are worth less than £15,400 on average

Guardian graphic. Source: ONS. Note: household total wealth by percentiles in Great Britain between April 2018 and March 2020

Source: https://www.theguardian.com/money/2022/jan/07/richest-uk-households-worth-at-least-36m-each

Quartiles

- A 3-numbers summary of the data.
- Order data from smallest to largest.



- Slice in 4 equal blocks.

  1st quartile: larger than 25% of the observations.

  2nd quartile: larger than 50% of the obs [= *median*].

  3rd quartile: larger than 75% of the obs.

- We often add min & max (a *5-numbers summary*)

Quartiles

Learn-to-drive dataset again:

2, 110, 5, 7, 6, 7, 3

Your turn: compute the quartiles.

1. Order from smallest to largest
   - 2, **3**, 5, **6**, 7, **7**, 110

1st quartile: 3

2nd quartile (=median):6

3rd quartile: 7

A five-numbers summary

Min: 2
1st quartile: 3
2nd quartile: 6
3rd quartile: 7
Max: 110

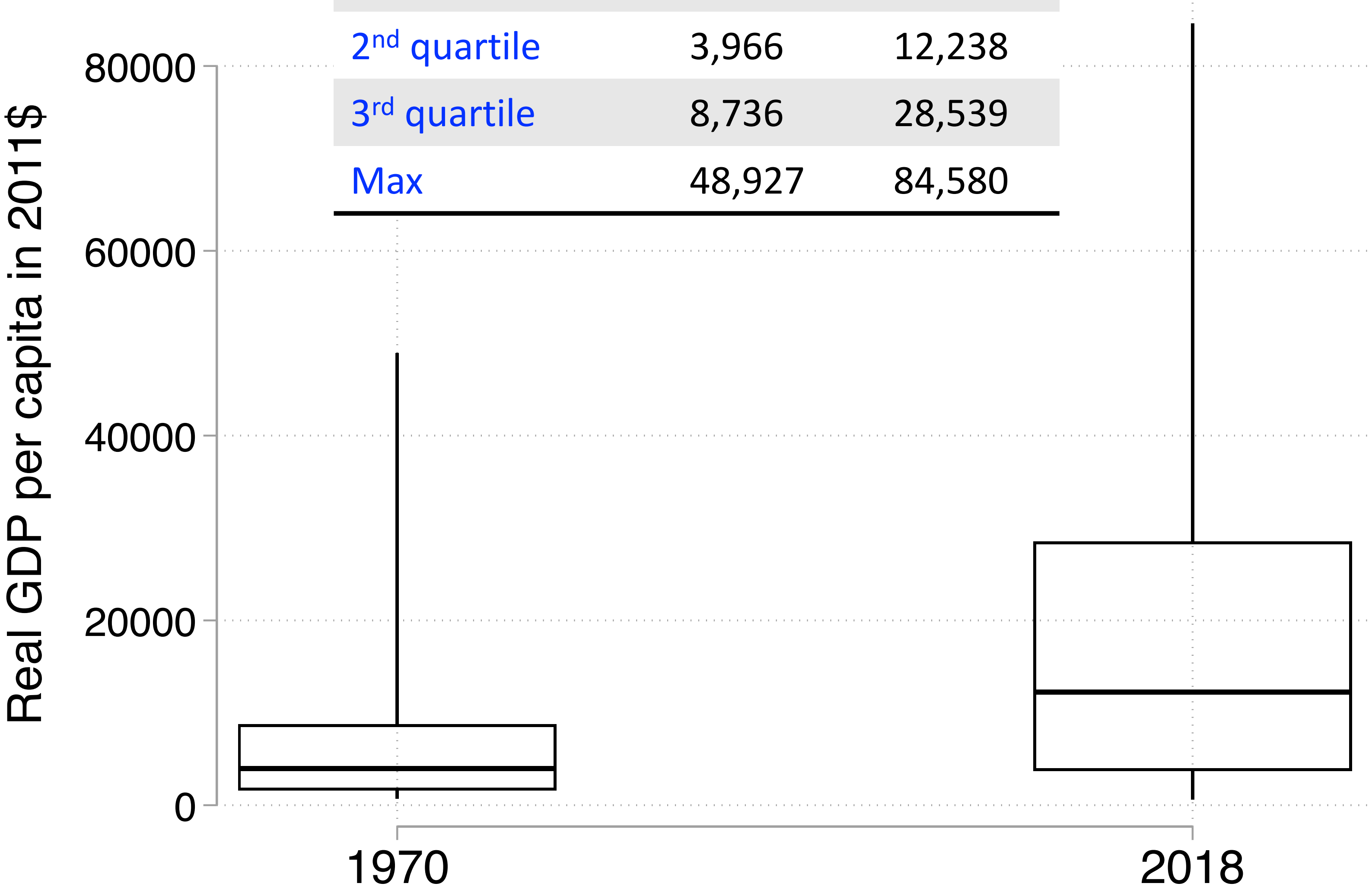
(in this case it's almost all values! But we usually have much larger datasets)

Boxplot

a visualization of the “five-numbers summary”

Real GDP per capita across countries

	1970	2018
Min	713	623
1st quartile	1,599	3,722
2nd quartile	3,966	12,238
3rd quartile	8,736	28,539
Max	48,927	84,580



Percentiles

- Order from smallest to largest.
- Slice data in 100 equal blocks

1st percentile: larger than 1% of observations

2nd percentile: larger than 2% of the obs.

…

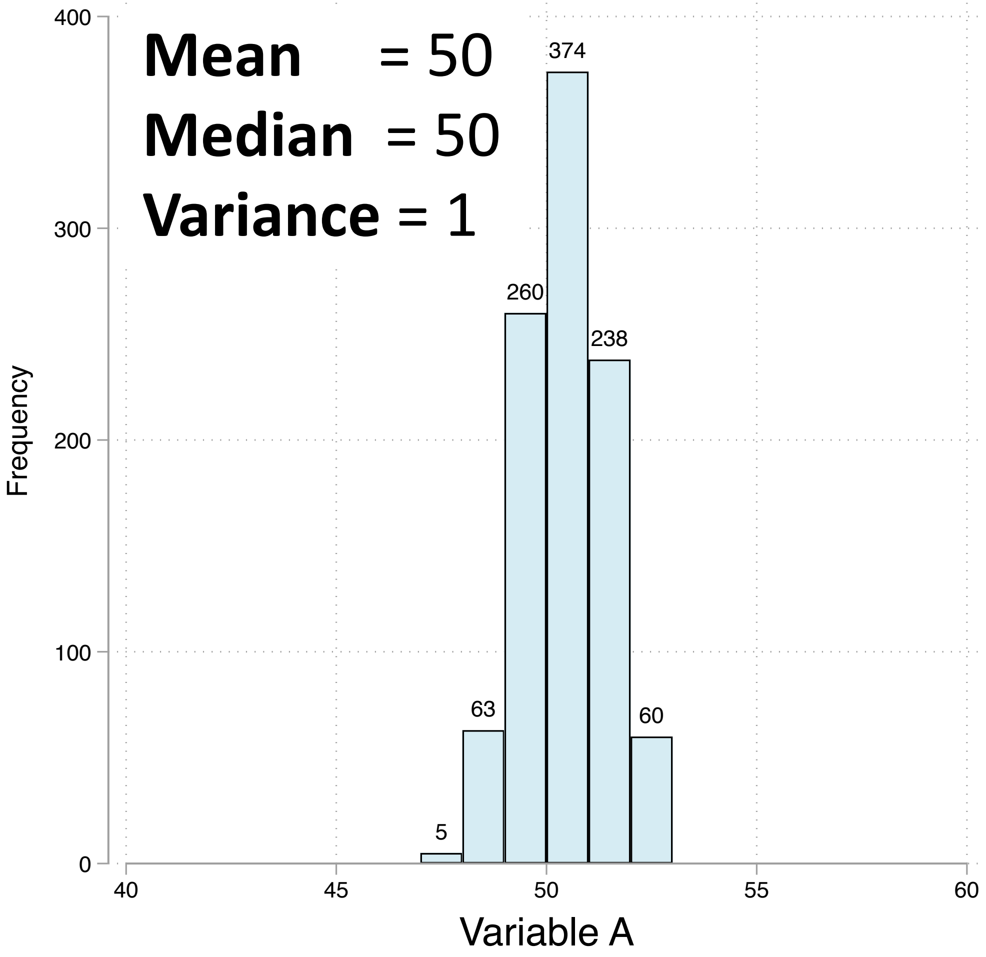
98th percentile: larger than 98% of the obs.

99th percentile: larger than 99% of the obs.

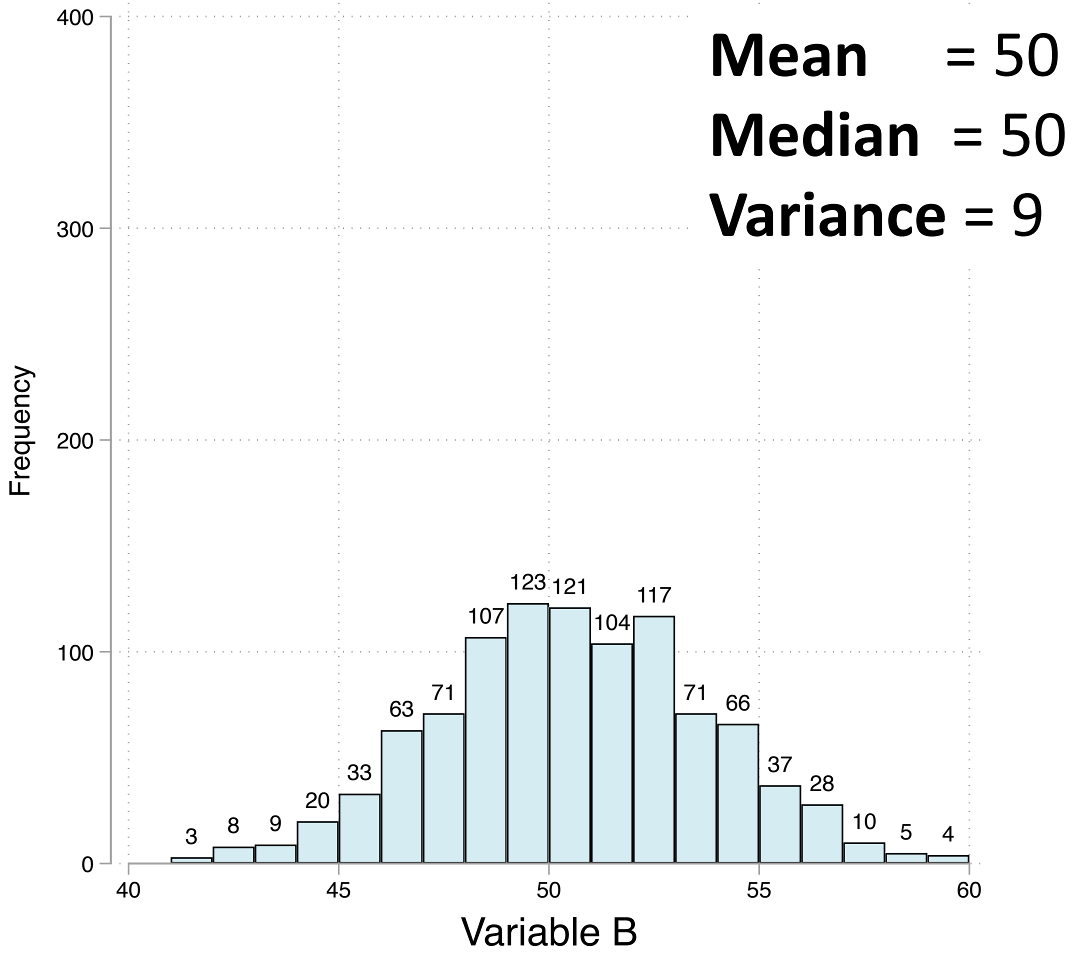
Quartiles & Percentiles

- The quartiles are nothing but 3 selected percentiles!
- 1st quartile ⟺ 25th percentile
- 2nd quartile ⟺ 50th percentile (aka median)
- 3rd quartile ⟺ 75th percentile
- Can also do *quintiles* (5 slices), *deciles* (10 slices), etc..

3 - Sample variance & S.D.



Mean = 50
Median = 50
Variance = 1



Mean = 50
Median = 50
Variance = 9

Sample variance

- Measures *variability* (or dispersion, or “spread”).
- *Mean squared deviation* of variable x from its sample mean $\bar{x}$
  - ○ Deviation for observation i: $x_i - \bar{x}$
  - ○ Squared deviation: $(x_i - \bar{x})^2$
  - ○ Variance = the average of this in the sample (kind of)

$$\boxed{s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n - 1}}$$

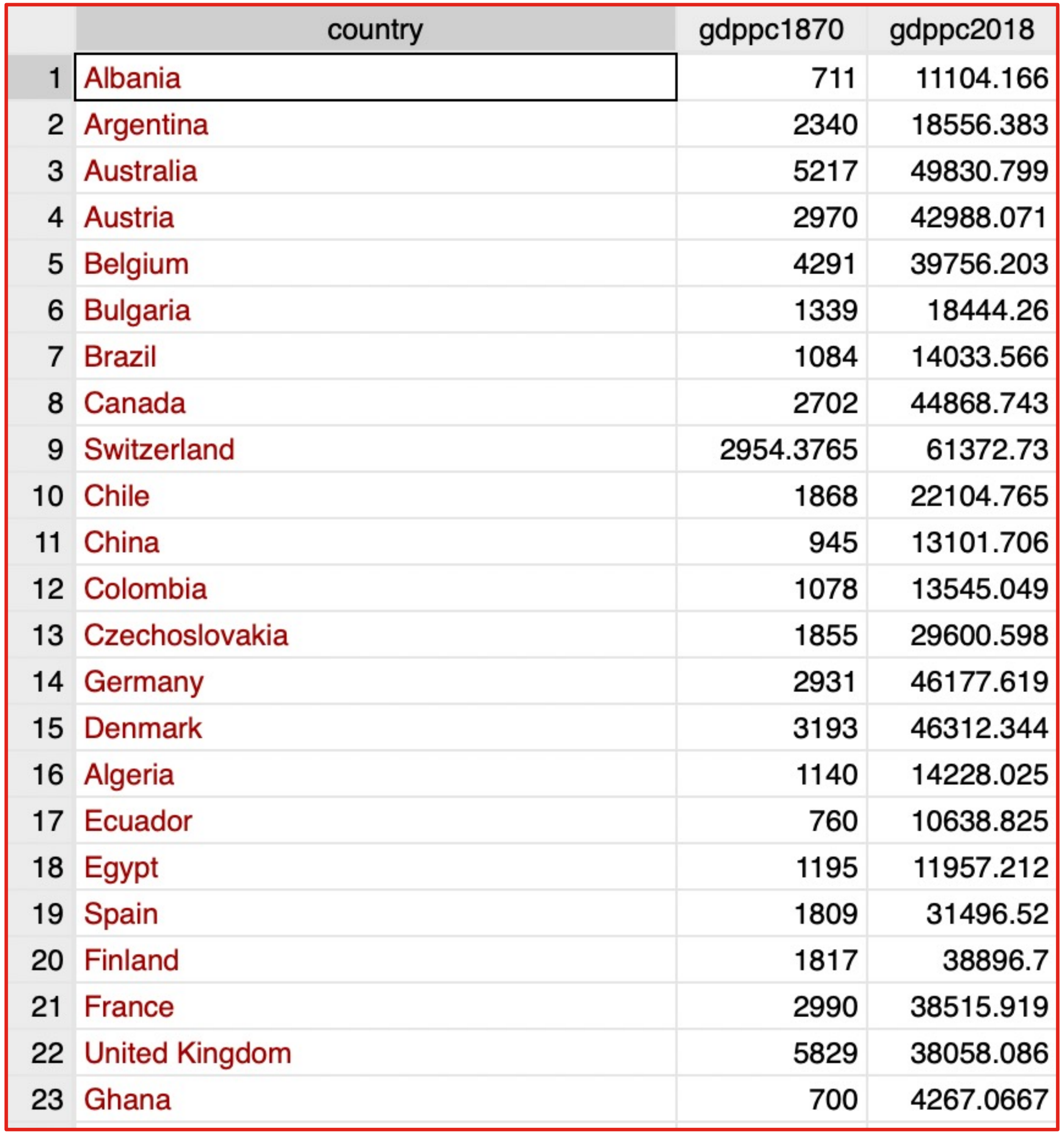
Standard Deviation (SD)

- Variance is measured in squared units of x
  - hard to interpret! (*squared USD*? *squared hours*?)

- SD: Square root of the variance.

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

- SD has the same unit of measure of the underlying variable



	country	gdppc1870	gdppc2018
1	Albania	711	11104.166
2	Argentina	2340	18556.383
3	Australia	5217	49830.799
4	Austria	2970	42988.071
5	Belgium	4291	39756.203
6	Bulgaria	1339	18444.26
7	Brazil	1084	14033.566
8	Canada	2702	44868.743
9	Switzerland	2954.3765	61372.73
10	Chile	1868	22104.765
11	China	945	13101.706
12	Colombia	1078	13545.049
13	Czechoslovakia	1855	29600.598
14	Germany	2931	46177.619
15	Denmark	3193	46312.344
16	Algeria	1140	14228.025
17	Ecuador	760	10638.825
18	Egypt	1195	11957.212
19	Spain	1809	31496.52
20	Finland	1817	38896.7
21	France	2990	38515.919
22	United Kingdom	5829	38058.086
23	Ghana	700	4267.0667

GDP per capita 1870

Mean: 1,796 (2011 USD)

SD: 1,238 (2011 USD)

GDP per capita 2018

Mean: 25,908 (2011 USD)

SD: 18,570 (2011 USD)

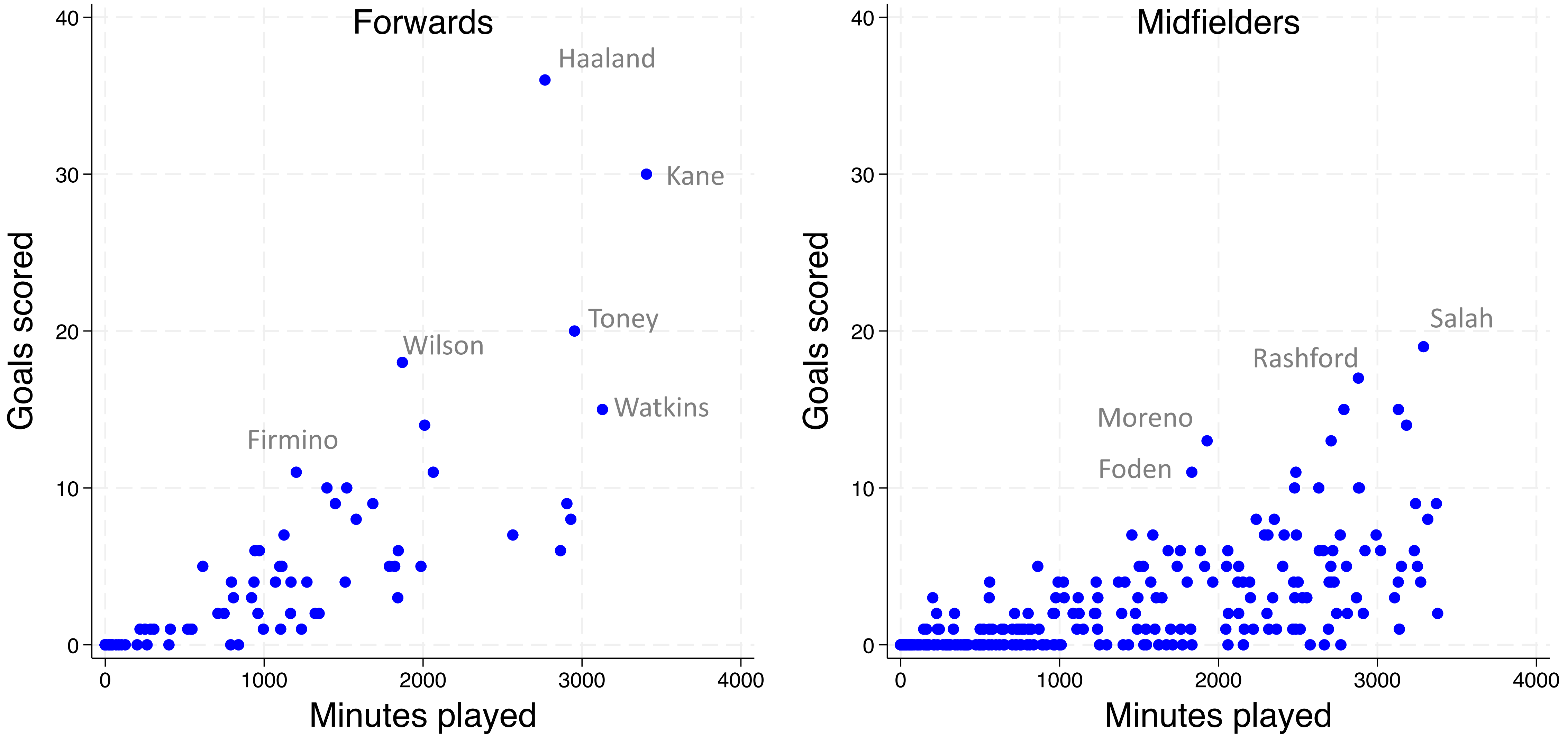
SD measures by how much observations tend to deviate from the mean.

4 - Sample correlation coefficient

- Do forwards that play more minutes score more goals?
- Does national income in 1870 predict national income today?
- Do Conservative Party voters tend to be richer than Labour Party voters?

- These Qs involve the relationship between two variables.

2020-2021 Premier League data



Sample coviariance

- How much do variables *x* and *y* move together in our sample?
- Sample covariance:

$$\boxed{s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}$$

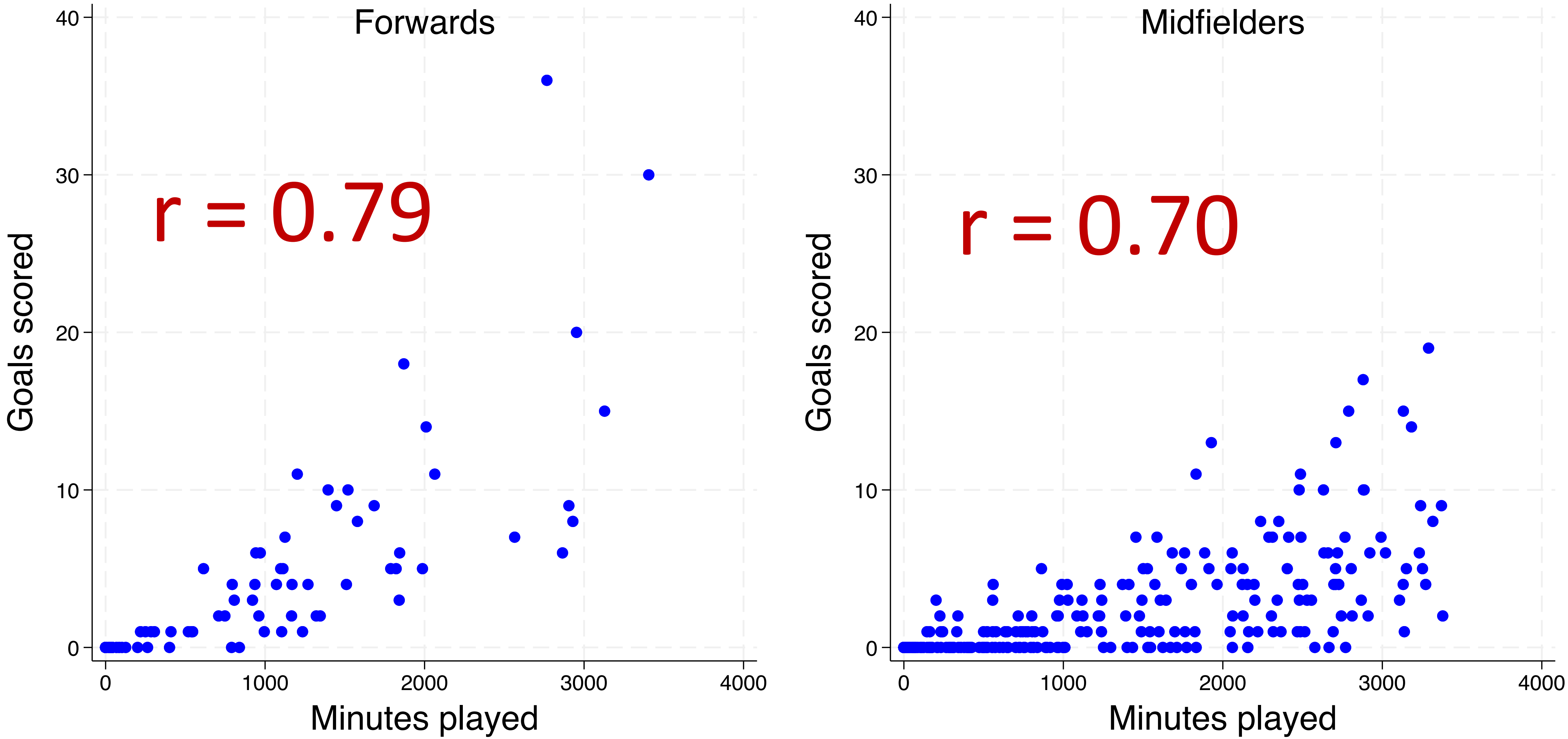
Sample correlation coefficient

- The units of covariance are awkward (units of x * units of y).

- Sample correlation coefficient:

$$\boxed{r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)s_x s_y}}$$

- Unit-free and always between -1 and +1.


Forwards
Midfielders
Goals scored
Minutes played
0
10
20
30
40
0
1000
2000
3000
4000

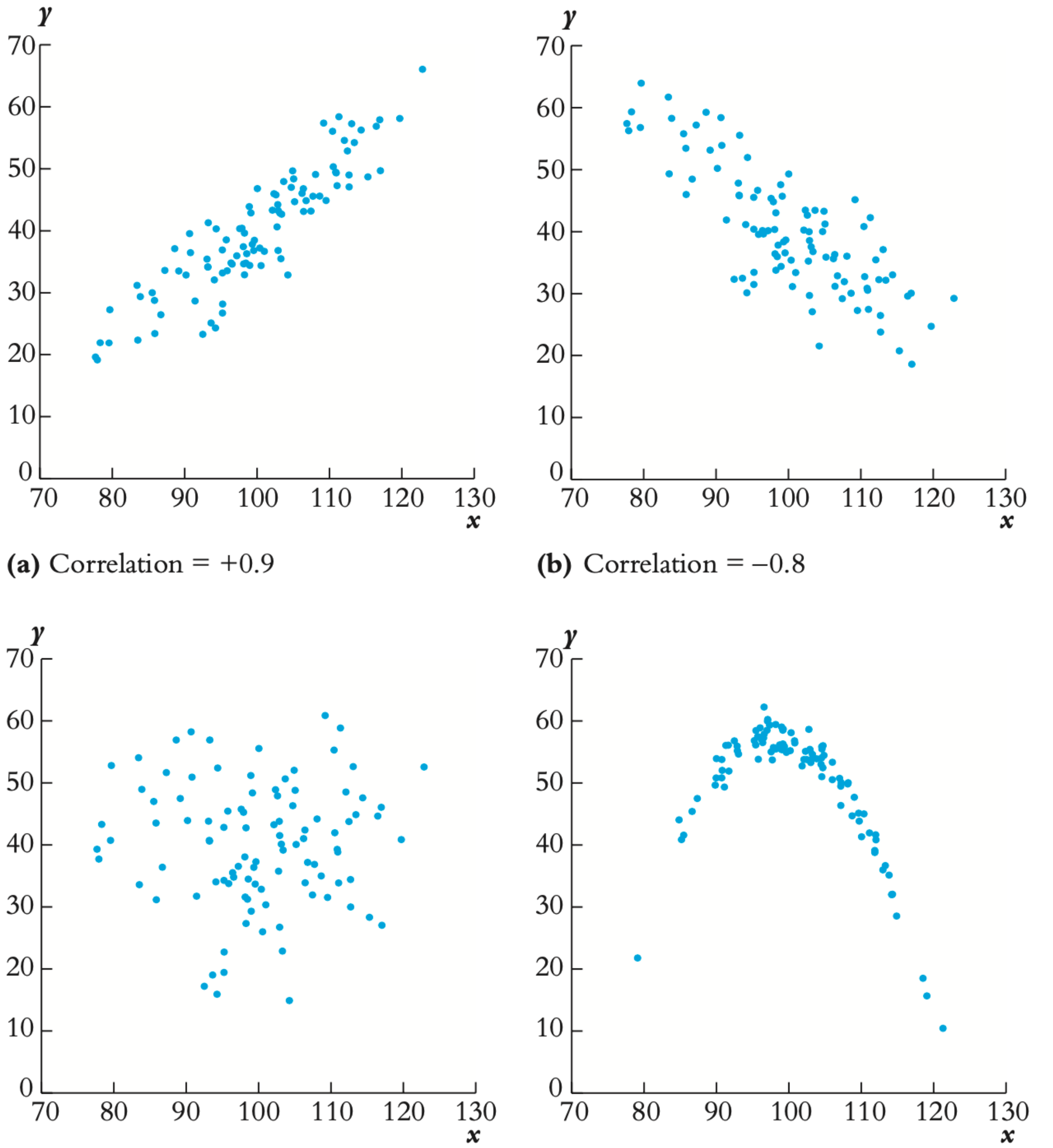
r = 0.79

r = 0.70

SCATTERPLOTS & CORRELATION COEFFICIENTS

- The correlation coefficient captures *linear* associations between variables, as in panels (a) & (b).

- It can miss non-linear ones, as in panel (d)



(a) Correlation = +0.9

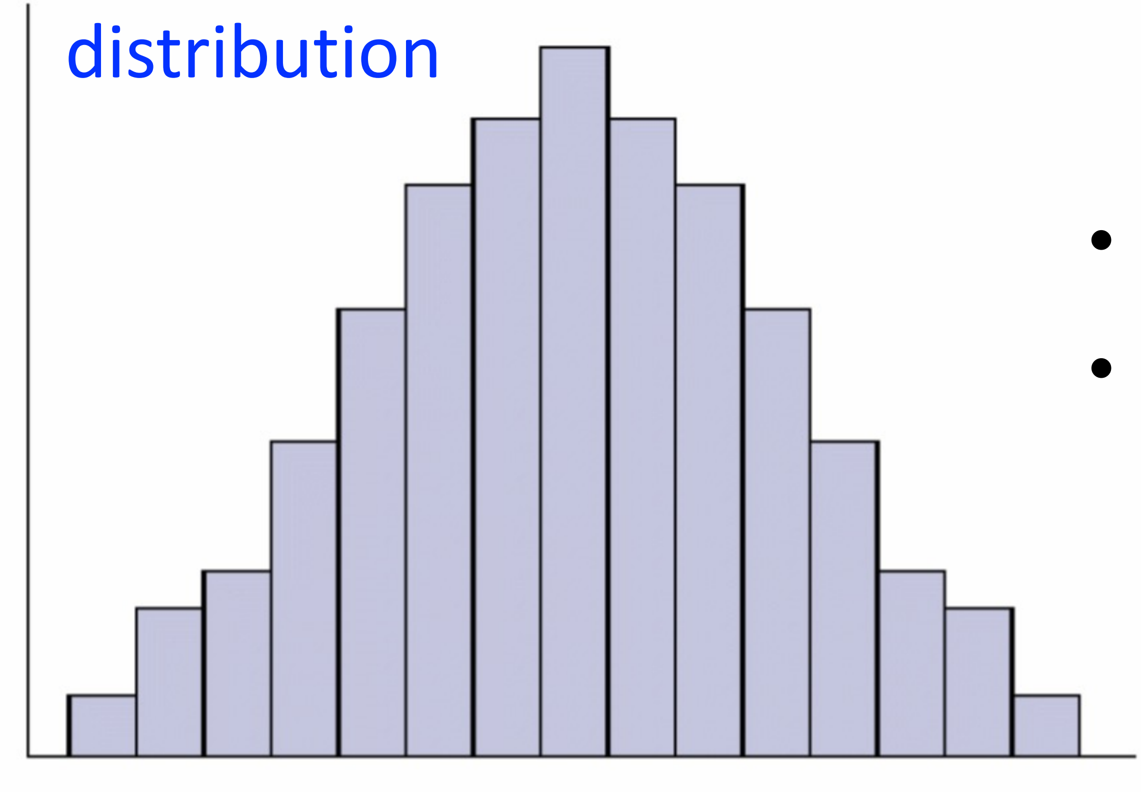
(b) Correlation = −0.8

(c) Correlation = 0.0

(d) Correlation = 0.0 (quadratic)

5 – "Normal" variables

Histogram of a variable with a normal distribution



- Relates to the shape of the histogram
  1. Highest in the middle
  2. Bell-shaped
  3. Symmetric
- mean=median.
- Moreover:
  - ~ 68% of data points less than a SD from mean.
  - ~ 95% less than 2 SDs from the mean.
  - ~ 99.7% less than 3SDs from the mean.

Thank you for your attention