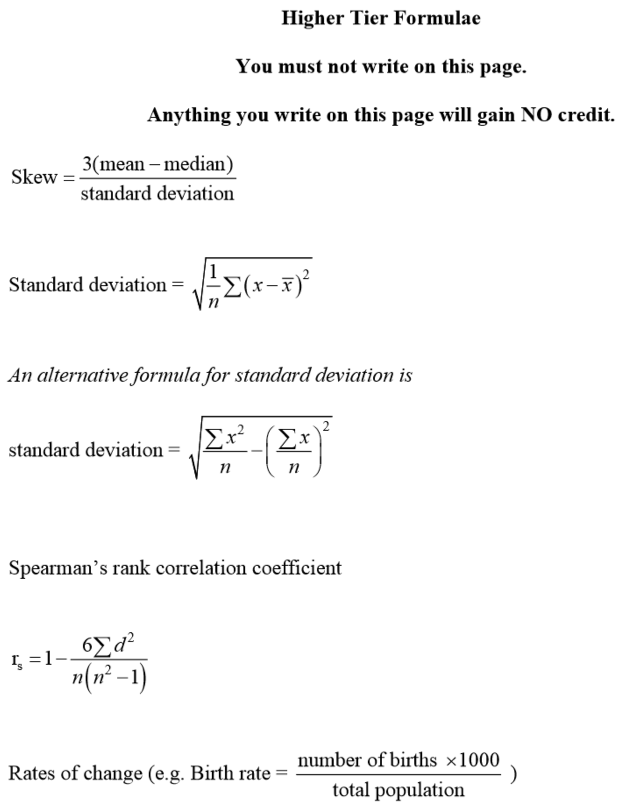
Edexcel GCSE Statistics Higher between paper revision

Topics:

- Normal Distribution
- Correlation
- Petersen Capture Recapture
- Index Numbers
- Risk
- Binomial Distribution
- Box plots plus outliers
- Cumulative Frequency Graphs
- Median by interpolation
- Random Response
- Control groups
- Matched pairs
- Stem and leaf diagrams
- Choropleth Maps
- Standard deviation
- Geometric and weighted mean



Higher Tier Formulae

You must not write on this page.

Anything you write on this page will gain NO credit.

$$\text{Skew} = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$$

$$\text{Standard deviation} = \sqrt{\frac{1}{n}\sum(x - \bar{x})^2}$$

An alternative formula for standard deviation is

$$\text{standard deviation} = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}$$

Spearman's rank correlation coefficient

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

Rates of change (e.g. Birth rate = $\frac{\text{number of births} \times 1000}{\text{total population}}$)

1.

For each of the following Normal distributions fill in the table for the range of values for which you find:

(a) The middle 68% of the population

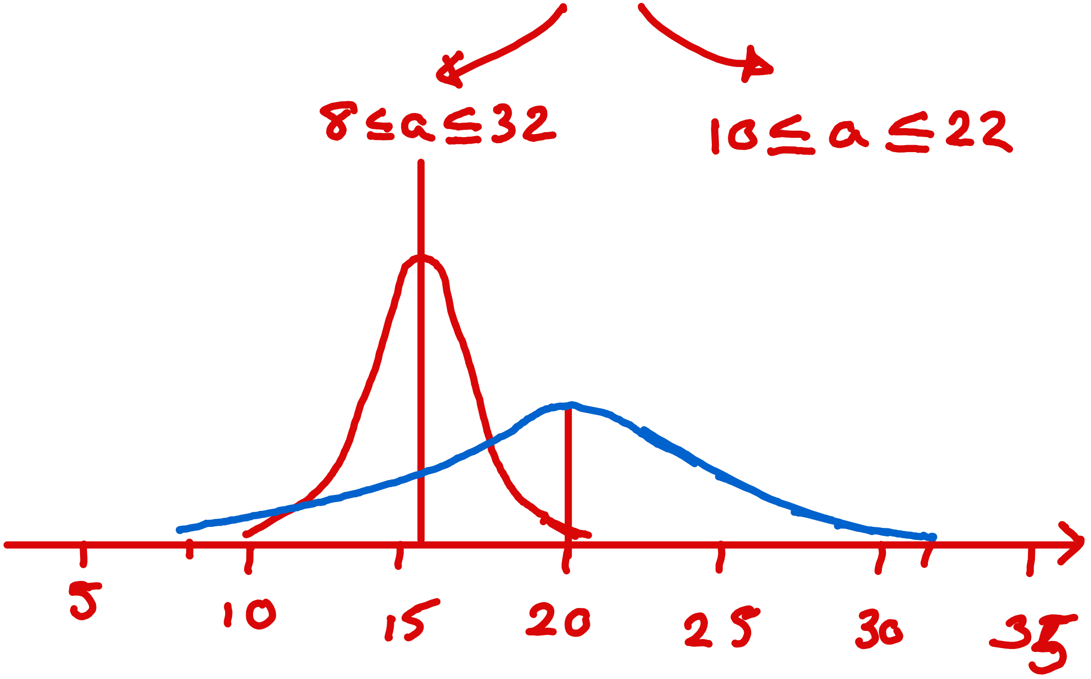
(b) The middle 95% of the population

(c) Almost all the population

Distribution	The middle 68% of the data	The middle 95% of the population	Almost all the population
$X \sim N(12,\ 2^2)$	$10 \le x \le 14$	$8 \le x \le 16$	$6 \le x \le 18$
$X \sim N(17,\ 36)$ 6^2	$11 \le x \le 23$	$5 \le x \le 29$	$-1 \le x \le 35$
$X \sim N(24.2,\ 0.16)$ $0 \cdot 4$	$23 \cdot 8 \le x \le 24 \cdot 6$	$23 \cdot 4 \le x \le 25$	$23 \le x \le 25 \cdot 4$

2. Sketch the following two normal distribution curves on the same axes, given below, showing the scale on the

	A	B
Mean	20	16
Standard Deviation	4	2



3. A company manufactures batteries with an average lifespan of 500 hours and a standard deviation of 50 hours.

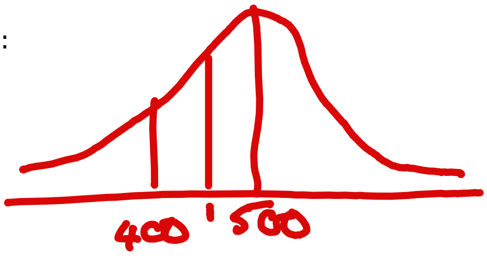
a)

i. Identify the type of probability distribution that can be used to represent the lifespan of these batteries.

Normal distribution $X \sim N(500, 50^2)$

ii. State one condition that must be met for this distribution to be a valid model.

b) Calculate the probability that a randomly selected battery will:



i. Operate between 400 hours and 500 hours.

$\frac{0.95}{2} = 0.475$

ii. Operate for less than 450 hours.

0.16

c) If 4000 batteries are tested, estimate how many of them would function for more than 550 hours.

$0.16 \times 4000 = 640$

4.

Below is the data for 8 runners' personal bests for a 5km race and a 10 km race

Runner	5 km time	10 km time			d	d^2
A	00:18:38	00:44:30	2	1	1	1
B	00:42:34	01:21:21	7	6	1	1
C	00:30:54	00:54:07	4	5	1	1
D	00:33:25	00:49:16	5	3	2	4
E	00:22:57	00:50:01	3	4	1	1
F	00:37:32	01:25:55	6	7	1	1
G	00:17:05	00:47:49	1	2	1	1
H	00:50:03	01:56:48	8	8	0	0

10

(a) Calculate Spearman's rank correlation coefficient for this data.

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{60}{(8)(63)} = 0.881$$

(b) Interpret this value of the Spearman's rank correlation coefficient.

There is strong positive correlation.

Runners performances in 5km and 10km races are similar

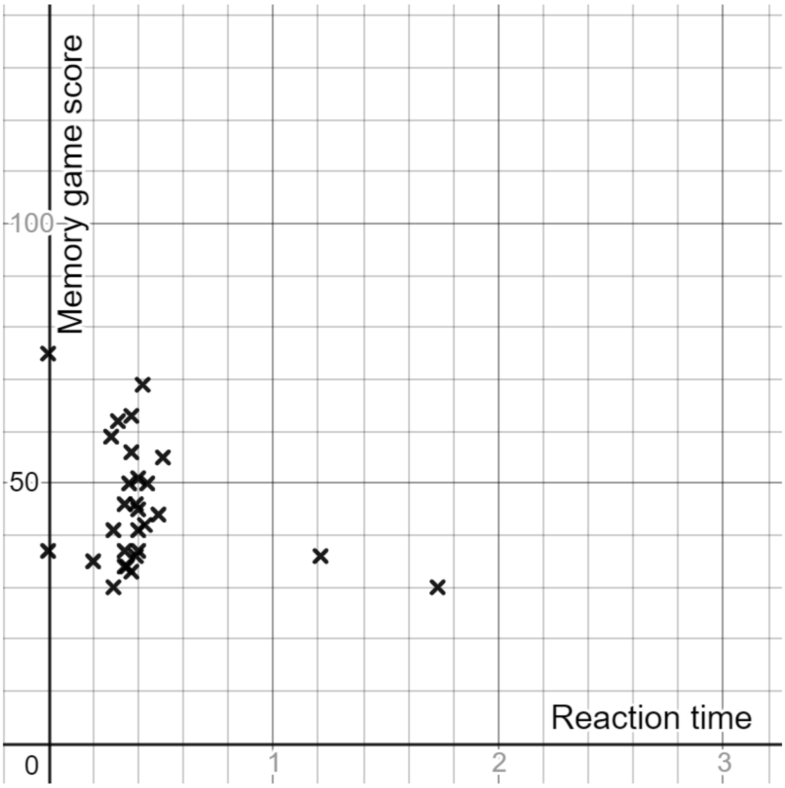
(c) The 10km time for runner D has been recorded incorrectly. It should be 00:59:16

Without any further calculation, explain whether the value of the Spearman's rank correlation coefficient will be larger, smaller or stay the same.

It will become smaller since the ranking order for C & D will now

5.

Below is the scatter graph of data showing students reaction time against their memory game score.



(a) Explain how you can tell this data has not been cleaned?

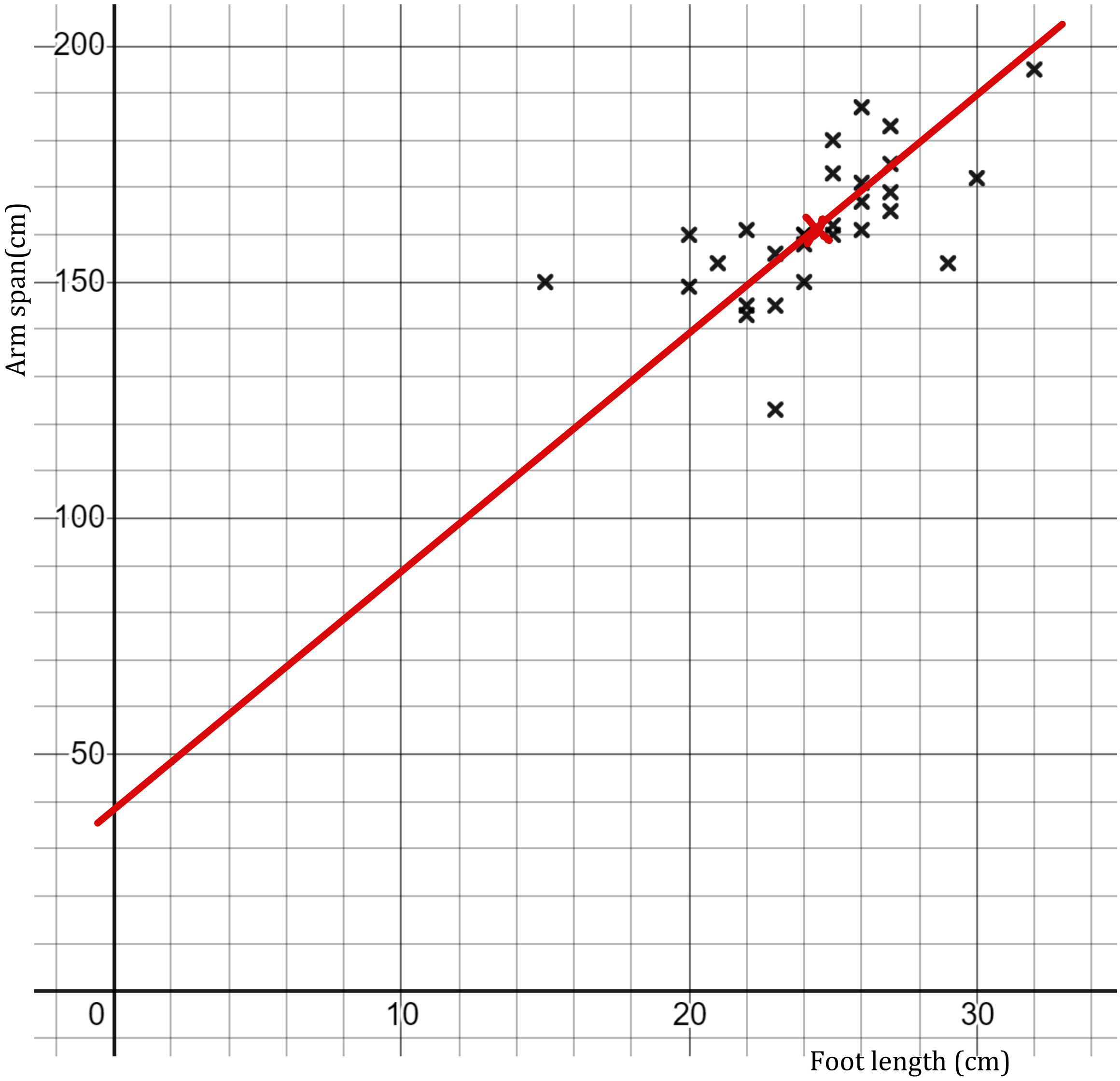
There are two outliers.

(b) The Pearson's correlation coefficient for this data once it has been cleaned is -0.272

Interpret this value in context.

There is a slight negative corelation between the variables.
Students with a faster reaction time are generally better at the memory test.

6. The following scatter graph shows foot length against arm span.



The mean point is (24.5, 161.7)

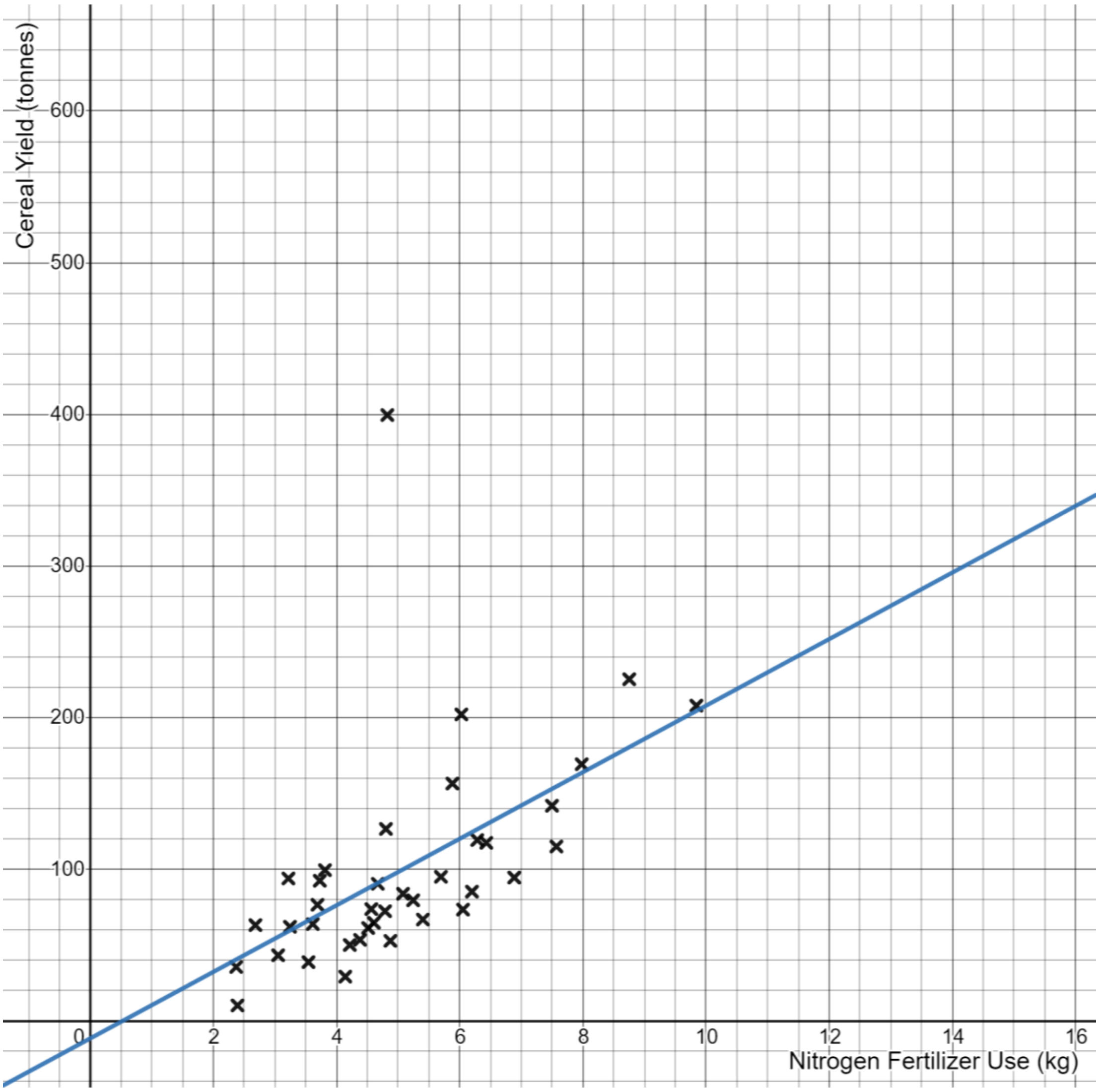
Draw a line of best fit on the scatter diagram and work out the equation of this line

Approx line shown – a lot of variation avaible in this line.

$y = 5x + 45$

7. The following scatter graph shows crop yield versus fertilizer use for the European countries in 2015.

Source: UN Food and Agriculture Organization



The equation of the line of best fit shown on the graph is

$$y = 21.9x - 11.5$$

(a) Interpret the coefficient 21.9

The yield increases 21.9 tonnes for every kg of nitrogen fertilizer.

(b) Interpret the constant -11.5

If you use no fertilizer your crop would be -11.5.

This is extrapolated data.

8.

(a) What is the base year for RPI?

1987

(b) What is the base year for CPI?

2015

9. The table gives the monthly average price of unleaded petrol so far in 2024.

Month	Jan	Feb	Mar	Apr	May
Prince per litre (pence)	139.36	141.47	144.66	148.85	149.31
Chain base index		101.5	102.25	102.9	100.3

(a) Calculate the chain base index numbers for April and May.
Give each value correct to 1 decimal place.

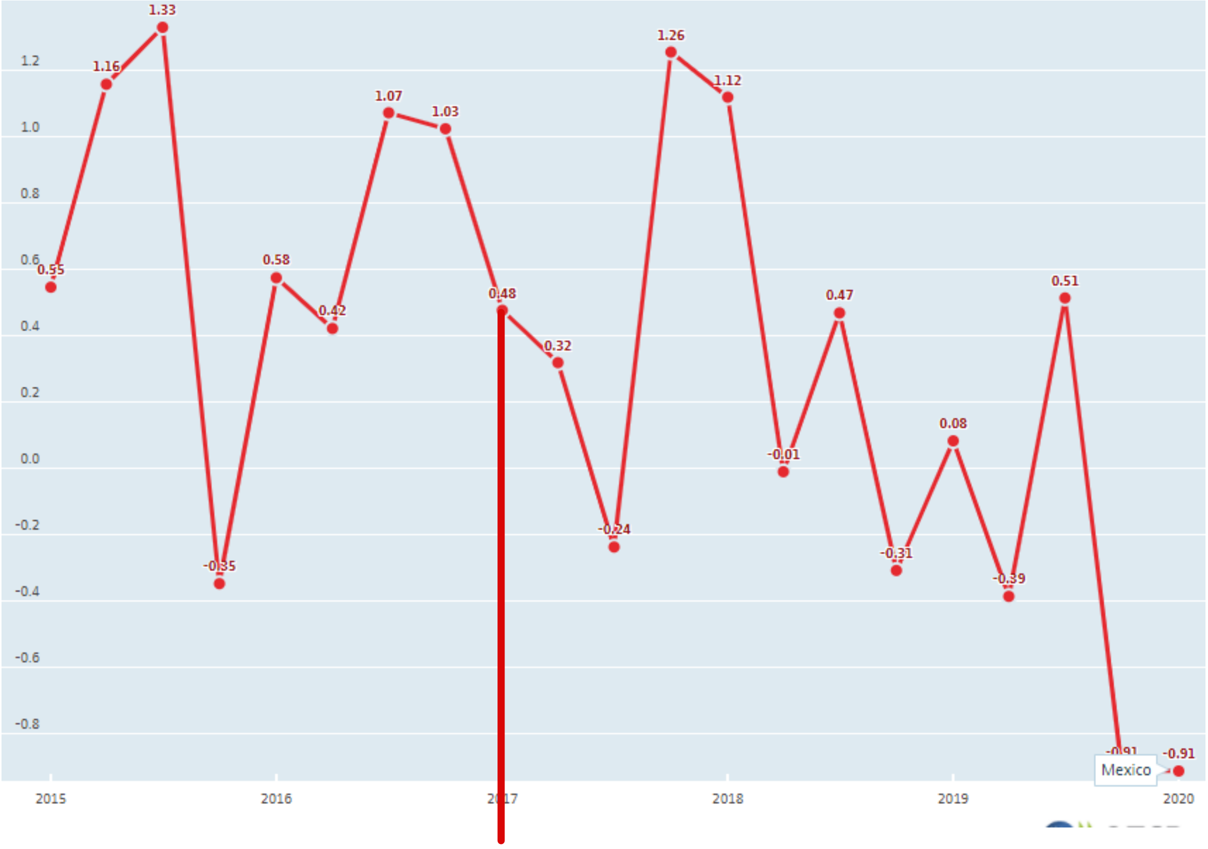
(b) Calculate the average monthly increase in petrol price per litre for 2024 so far.

$$\sqrt[4]{101.5 \times 102.25 \times 102.9 \times 100.3}$$

$$= 101.73$$

1.73% increase

10. State when this country Mexico first went into recession.
https://data.oecd.org/chart/6EBq



Q1 2017

Two successive quaters of fall in GDP.

11. Below are the number of deaths for two countries.

	Country A		Country B	
Age group	Number of Deaths	Population	Number of Deaths	Population
0-29	7 000	6 000 000	6 300	1 500 000
30-59	20 000	5 500 000	3 000	550 000
60+	120 000	2 500 000	12 000	120 000

(a) Calculate the crude death rate for the 60+ age group for both countries.

$$\text{Death rate} = \frac{\text{number of deaths} \times 1000}{\text{total population}}$$

A

$$\text{Crude death rate} = \frac{120000 \times 1000}{2500000}$$

$$= 48 \text{ deaths per thousand}$$

B

$$\text{Crude death rate} = \frac{12000 \times 1000}{120\,000}$$

$$= 100 \text{ deaths per thousand}$$

(b) Given the standard population for the two countries is:

Age group	Country A	Country B
0-29	429	691
30-59	392	253
60+	179	56

Use this to calculate the standardised death rate for the 60+age group for both countries

$$\text{Standardised rate} = \frac{\text{crude rate}}{1000} \times \text{standard population}$$

(c) Compare the two standardised rates

A

$\frac{48}{1000} \times 179 = \underline{\underline{8.592}}$

B

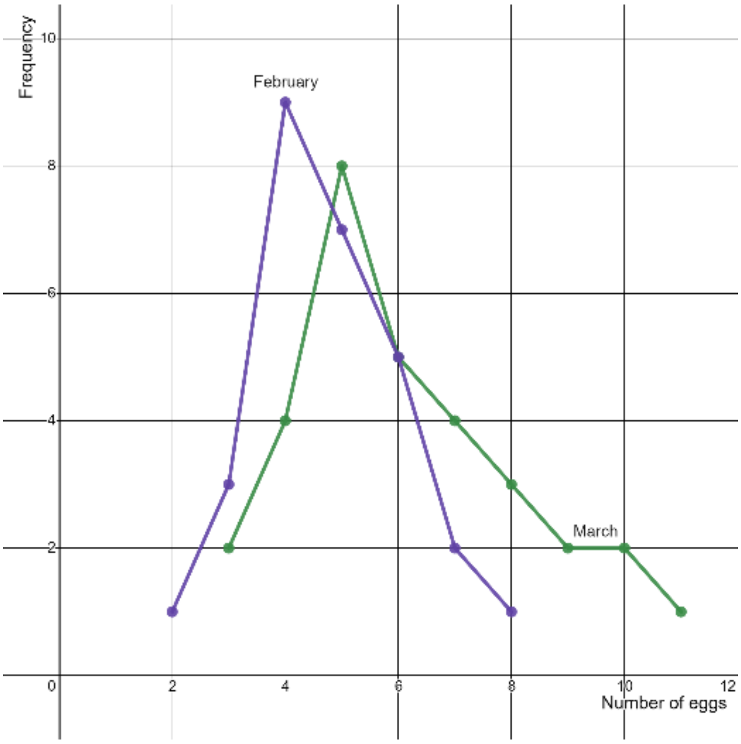
$\frac{100}{1000} \times 56 = \underline{\underline{5.6}}$

The standardised death rate for Country A is higher than the standardised death rate for Country B, meaning that a greater proportion of the 60+ age group die in country A.

12.

The frequency polygons below show the numbers of eggs collected by Conor from his hens during February and March 2021.

Compare the distributions



The mode of the eggs collected in February is 4 which is lower than the mode of the eggs collected in March, 5.

The range of the number of eggs collected in February is 4, which is lower than the range of the number of eggs collected in March, 8.

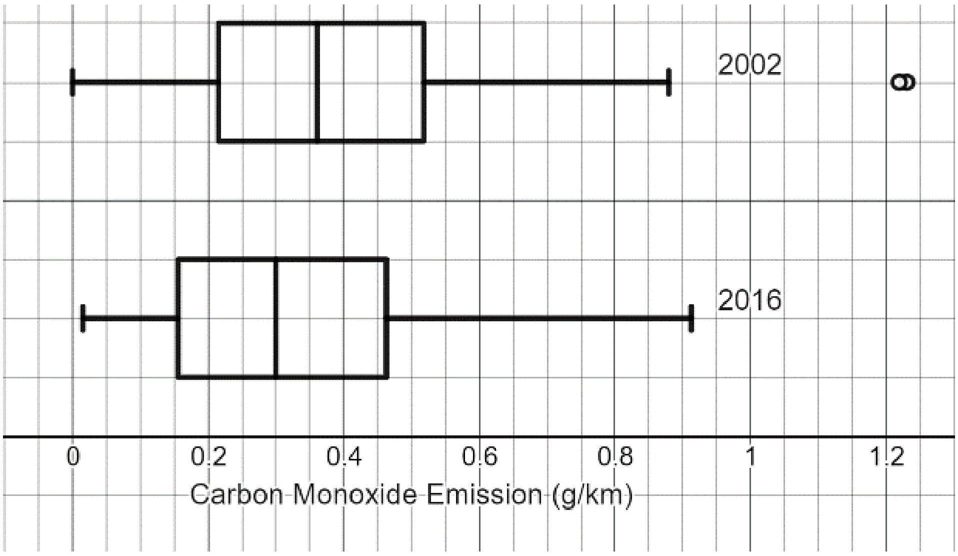
February has slight positive skew, where as March has more positive skew.

Overall there is more variation in the number of eggs collected in March.

13.

The box and whisker plots below show the carbon Monoxide emission for a sample of 100 cars manufactured in 2002 and a sample of 100 cars manufactured in 2016.

Compare the distributions



The CO emission of the cars manufactured in 2002 has a median of 0.35 which is higher than the median of 0.3 for cars manufactured in 2016

The CO emission of the cars manufactured in 2002 has an IQR of 0.3 which is the same as cars manufactured in 2016.

The CO emission of cars manufactured in 2002 has two outliers, where as 2016 has none.

The CO emission of cars manufactured in 2002 is almost symmetric, where as 2016 has positive skew.

Overall cars manufactured in 2002 have more emissions.

14.

(a) Calculate and estimate for the mean and standard deviation of this data set, you must show all your working:

Height	Frequency	m	m^2	mf	m^2f	
$100 < x \leq 110$	1	105	11025	105	11025	
$110 < x \leq 120$	0	115	13225	0	0	
$120 < x \leq 130$	1	125	15625	125	15625	
$130 < x \leq 140$	1	135	18225	135	18225	
$140 < x \leq 150$	3	145	21025	435	63075	
$150 < x \leq 160$	9 15	155	24025	1395	216225	
$160 < x \leq 170$	9	165	27225	1485	245025	
$170 < x \leq 180$	5	175	30625	875	153125	
$180 < x \leq 190$	2	185	34225	370	68450	
	31			4925		

$$\bar{x} = \frac{4925}{31} = 158.8$$

$$\sum m^2 f = 790775$$

$$\sigma^2 = \frac{\sum m^2 f}{31} - \bar{x}^2$$

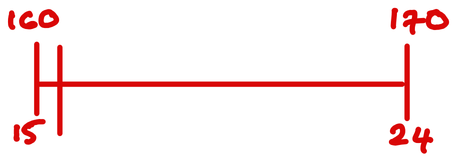
$$= 268.9$$

$$= \underline{\underline{16.4}}$$

(b) Comment on the skew of the data, showing any calculations you make.

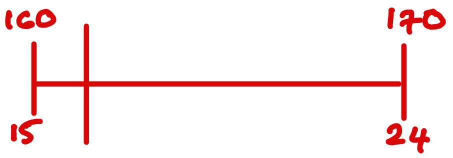
Median by interpolation: 15.5th or 16th.

15.5th.



$\frac{0.5}{9} \times 10 + 160 = 160.6$

16th



$\frac{1}{9} \times 10 + 160 = 161.1$

$\text{Skew} = \frac{3(\text{Mean} - \text{Median})}{\text{s.d}}$

Either: Skew $= -0.316$

or

Skew $= -0.408$

Slight negative skew.

(c) It turns out one of the data values in the interval $180 < x \leq 190$ had been recorded incorrectly and should have been in the interval $170 < x \leq 180$.

Without any further calculations, explain the effect this would have on the mean, median and standard deviation :

Mean

It will reduce

Median

Stay the same

Standard Deviation

Reduce, new value is closer to the mean

15. The percentage voter turn out for the last six general elections is given in the table below

Calculate the average voter turn out over the last six general elections

Year	Percentage turn out
2001	59.4%
2005	61.4%
2010	65.1%
2015	66.2%
2017	68.8%
2019	67.3%

$$\sqrt[6]{59{\cdot}4 \times 61{\cdot}4 \times 65{\cdot}1 \times 66{\cdot}2 \times 68{\cdot}8 \times 67{\cdot}3}$$

$$= 64{\cdot}6\%$$

16.

Tabitha collected data from her class about the number of siblings they had and their ages.

Below is an extract of her spreadsheet.

Give three reasons why she would need to clean this data:

Number of siblings	Your age	Age of sibling	Age of sibling	Age of sibling
2	13	8		
1	12	15		
0	12			
Three	12	9	14	16
1 brother, 1 sister	13	15	17	
2	1.2	5	8	
1	12			
1	13	16		

The number of siblings column contains the word "three" which needs changing to 3.

The number of siblings column contains the phrase "1 brother, 1 sister" which needs changing to 2.

Row 1 and 7 are missing the age of the sibling.

17.

Rhys has a database of information about weather in the UK at Heathrow.

Here is the information that the database contains about each day of the year in 2021.

Average temp	Max Temp	Min Temp	Max wind speed (gust)	Average wind speed	Dew point	Wind direction	Rainfall

Explain how Rhys can use technology to clean this data:

SPREADSHEET

- He can sort the data to identify missing data or data entry error

- He can use find and replace to deal with remaning units

- He can calculate statistics such as the mean and standard deviation quickly

GRAPHING SOFTWARE OR SPREADSHEET

- He can use graphs produced by software which will be more accurate, look nicer.

18.

Jayden wants to ask the following question in a survey:

Have you ever shoplifted? □ " Yes " □ "No"

Explain how they would use the random response technique to collect answers to the question:

Ask respondees to flip a coin.
If they get a head answer 'YES'
If they get a tail answer the question honestly.

Jayden thinks about 5% of people will have shoplifted at some point in their life.

Jayden gets the following data

Yes 132

No 98

Does this support his hypothesis?

$132 + 98 = 230$

$\frac{230}{5} = 115$

$132 - 115 = 17$

$\frac{17}{115} \times 100 = 14.8\%$

No, this does not support his hypothesis.

19.

A teacher believes that students who use flash cards learn definitions better.
She randomly splits her class into two groups A and B.
She asks group A to use flash cards to learn their definitions.
She asks group B to read the definitions from their book to learn the definitions.

She then gives the students a 50 question definition test.

(a) Identify the control group in this experiment

Group B

(b) Identify whether this is a field, natural or laboratory experiment.

Natural

(c) State one disadvantage of this experiment

No control over extraneous variable

(d) State one advantage of this experiment

Easy to set up

(e) Explain how the teacher could adjust this experiment to use matched pairs.

Pair up students with similar grades before the experiment and assign one of each pair to A and one of each pair to B.

20.

A butterfly farm is conducting a study to estimate the population of a particular butterfly species.

They initially captured and tagged 100 butterflies, then released them back into the farm.

After one day, they conducted a second capture and found that among the 80 butterflies captured, 20 were tagged.

Estimate the total population of butterflies in the butterfly farm.

Discuss the validity of your estimate.

$$\frac{100}{N} = \frac{20}{80}$$

$$N = 400$$

VALID:

Population closed, no butterflies can escape.

The tagging doesn't affect survival rate

One day is enough time for butterflies to mix.

21.

Below are the test results of Class A, which consisted of 31 pupils, displayed as a stem and leaf diagram.

4	5
5	2
6	3
7	0 2 2 2 3 4 5 6 6 8 9
8	0 1 2 3 4 5 7 7 8 9
9	0 1 5 6 7 8 8

Key 6|3 mean 63%

(a) Calculate the value of the median, lower quartile and upper quartile

$\frac{31+1}{2}$th value.   81

LQ 8th   73

UQ 24th   89

(b) Identify if there are any outliers by calculation.

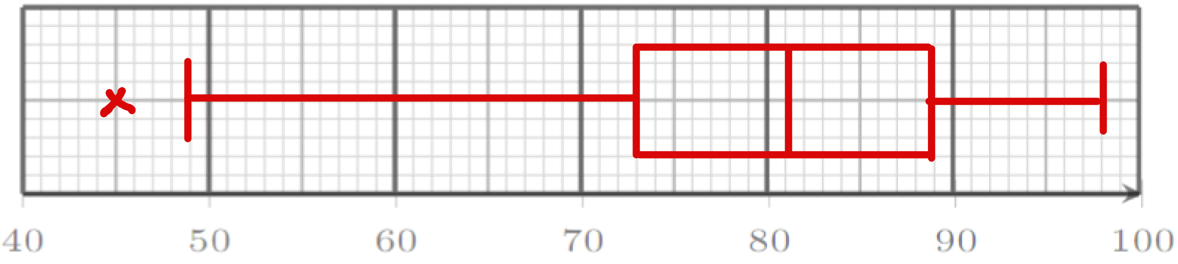
Outliers   1.5(IQR) = 1.5(16) = 24

Lower fence   73-24 = 49

Upper fence   89+24 = 113

45 < 49   45 is an outlier.

(c) Draw a box and whisker plot, showing outliers if they exist.



22.

(a) The following table gives Emily's percentage marks and their respective weights in her music exam:

	Performance	Theory	Composition
% Mark	75	80	85
Weight	40	35	25

An overall mark of 80% would give Emily a distinction for the exam. What percentage did Emily get? Did she gain a distinction?

$$\frac{75 \times 40 + 80 \times 35 + 85 \times 25}{100}$$

$$= \frac{7925}{100} = \underline{\underline{79.25\%}}$$

No.

(b) Fabio still has to complete his composition exam. What is the minimum mark he could achieve and gain a distinction?

	Performance	Theory	Composition
% Mark	80	75	
Weight	40	35	25

$$\frac{80 \times 40 + 75 \times 35 + x \times 25}{100} \geq 80$$

$$3200 + 2625 + 25x \geq 8000$$

$$25x \geq 2175$$

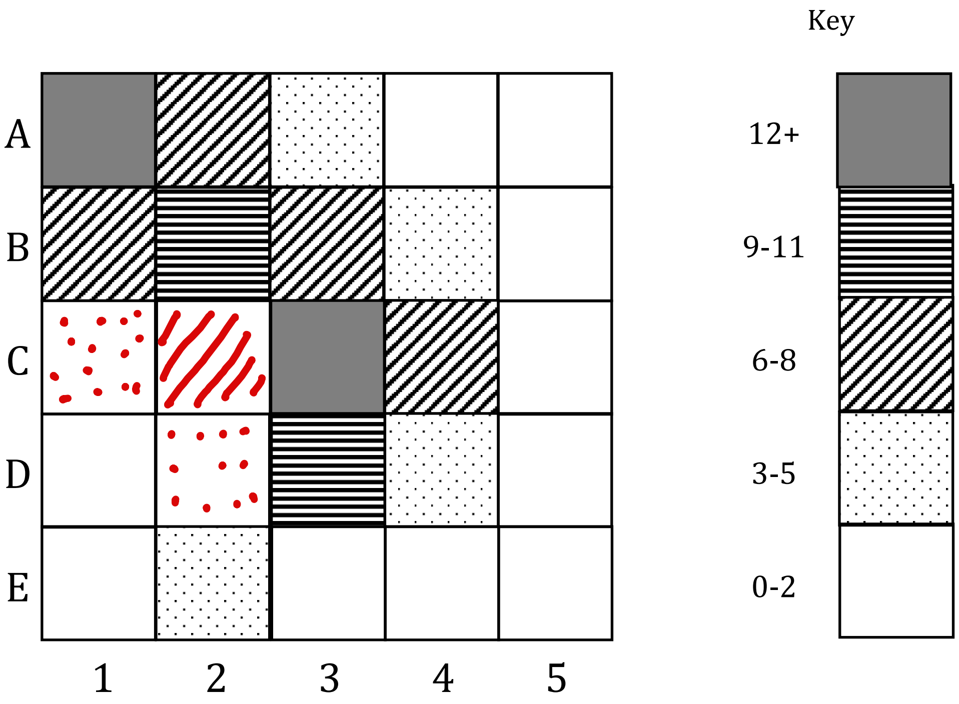
$$x \geq 87$$

Minimum = 87%

23. The grid below shows the amount of pieces of litter collected from the ground after lunch on a section of a school field.

A	12	8	4	0	1
B	7	9	8	3	0
C	3	6	14	6	1
D	1	5	10	4	0
E	2	3	2	0	0
	1	2	3	4	5

(a) Complete the choropleth map by filling in squares C1, C2, D1, D2



(b) The headteacher has decided to put one more bin on this section of the field.
Explain with reasons where he should put the bin

Allow any of A1, B2, C3 with reason, most litter is clustered around this area.

24. Jordan has surveyed his school on their use of Vinted.

He asks people if they receive the item "as described" or "not as described" and what the average five star rating of the seller is.

He displays his results in a table.

	As described	**Not as described**
4 stars or more	80	3
Less than 4 stars	45	10

Jordan says the relative risk of receiving an item "not as described" when the seller has an average rating of less than 4 stars, compared to when the seller has an average rating of 4 stars or more, is approximately 5.

(a) Show that Jordan is correct

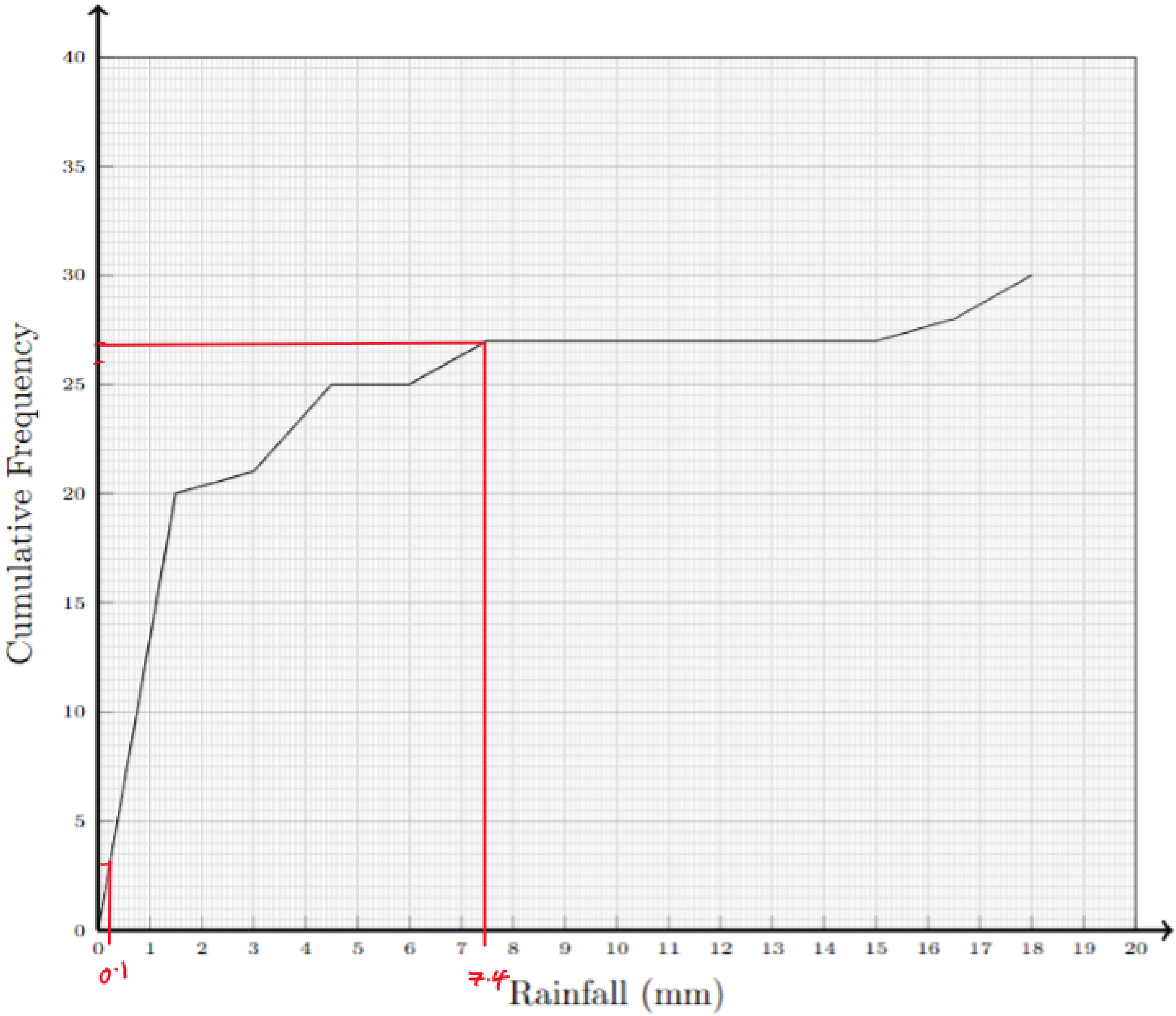
$$\text{R. risk} = \frac{\text{Abs risk in group}}{\text{Abs risk not in group}} = \frac{10/55}{3/83}$$

$$= 5.03$$

(b) Interpret a relative risk of 5 in context.

If you buy something from a seller with an average rating of less than 4 stars then you are 5 times more likely to receive an item "not as described"

25. The cumulative frequency graph shows the daily rainfall in Reigate during April 2024



(a) Find the 10th to 90th interpercentile range for rainfall in Reigate in April 2024

10% of 30 = 3 90% of 30 = 27

$7{\cdot}4 - 0{\cdot}1 = \underline{\underline{7{\cdot}3}}$

26. On a fruit stall at the market, the probability of an apple being bruised is 0.08.

Bradley buys 9 apples.

(a) Find the probability that 6 apples are bruised.

$X \sim Bin(9, 0.08)$

$$P(X=6) = {}^{9}C_{6}\ 0.08^{6}\ 0.92^{3}$$

$$= 0.0000171$$

(b) Discuss the validity of the assumptions you have made.

- Two outcomes, bruised/not bruised – VALID
- Trials are independent – this is unlikely to be valid because bruising is likely to happen to apples in the vacinity of one another.