

**Statistics GCSE****Paper 2**

2025

Edexcel Higher

Variant 2

1ST0/2H

**Instructions**

- Write all answers in the spaces provided.
- Answer all questions.
- You must show all your working.
- There may not be enough space to show all your working out.

**Information**

- This is a practise paper to aid your revision for your exams.
- This site, and all that work on it, have no affiliation or relationship with any exam board.
- This site is not endorsed by any company or charity, unless we state otherwise.

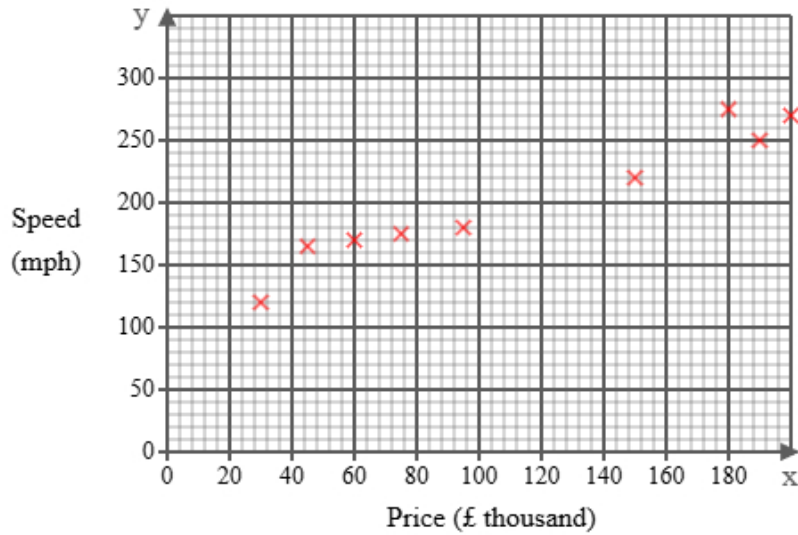
**Copyright**

- Sharing this PDF is strictly forbidden unless you have the author's permission.
- This paper is only authorised to be used by the person who has bought it.
- You may not store this online or any shared area such as an intranet.
- Please contact us if you have any queries.

**Advice**

- You can get support for all these questions at our website: [www.statsgcse.com](http://www.statsgcse.com)
- This paper and more are available on our site with questions that change subtly after each attempt.
- Good luck!

1 Emma recorded data on 11 cars, recording their top speed (in miles per hour) and their price (in thousands of pounds). She represented her findings in the scatter diagram below.



(a) One of the 11 cars has a top speed of 225 mph.  
For this car, write down its price.

(1 mark)

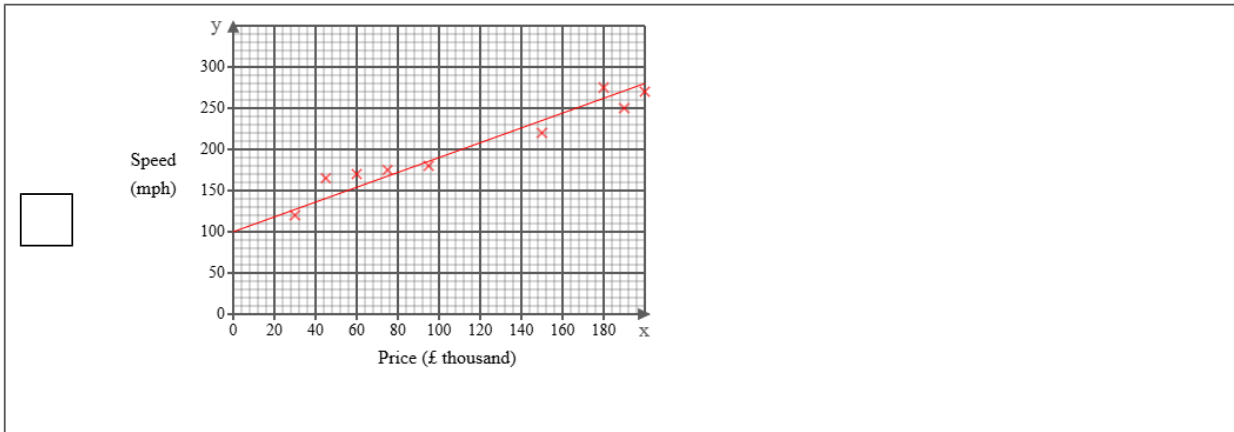
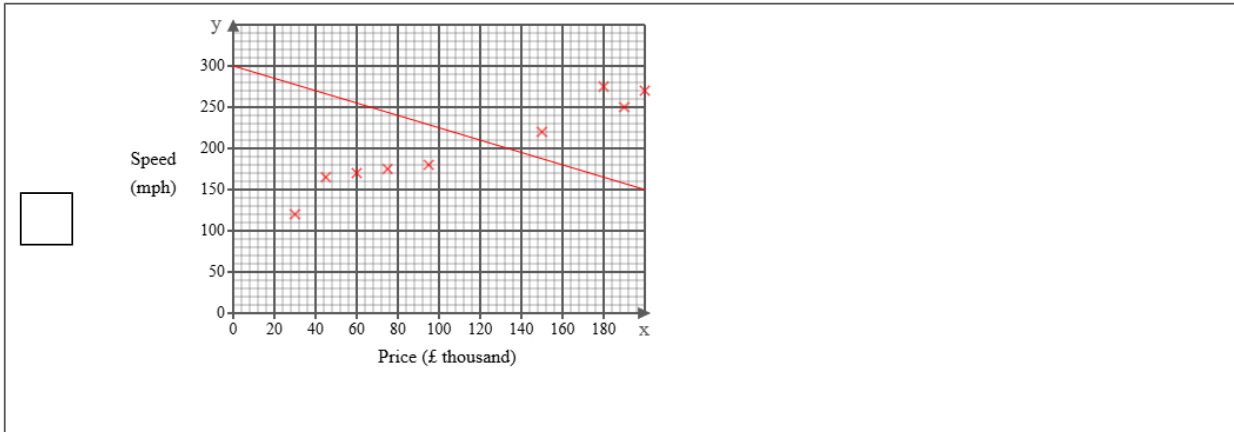
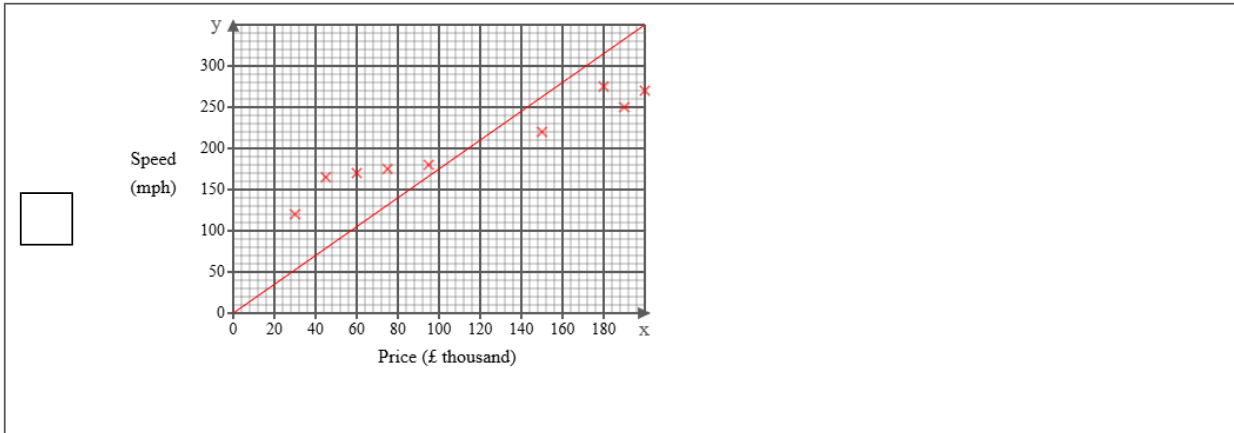
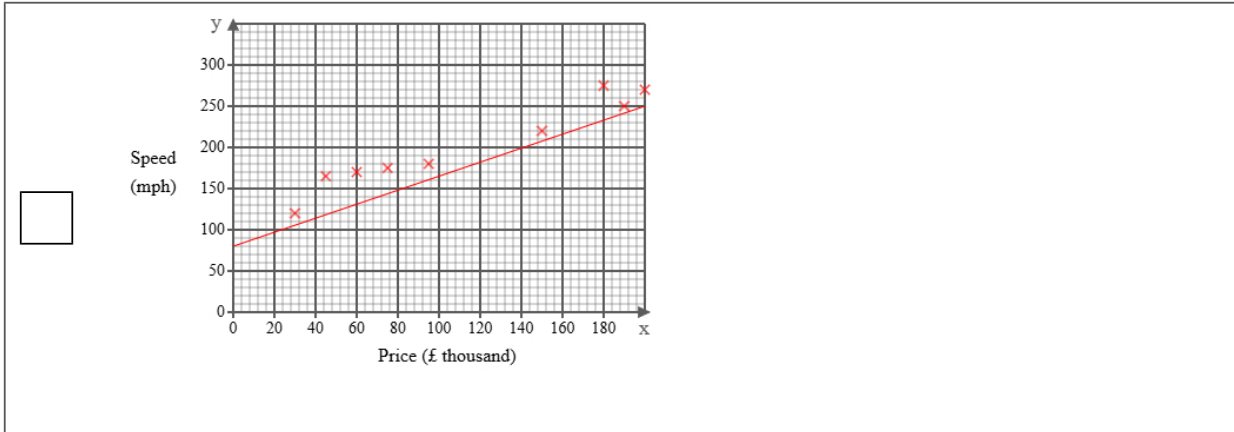
Find the cross on the scatter graph that is at 225 mph on the x-axis (the bottom axis), then read off the value from the y-axis (the side axis).

\_\_\_\_\_ mph

**(b)** Draw a line of best fit on the scatter diagram.

(1 mark)

Select the correct answer.



(c) Describe and interpret the type of correlation shown by the scatter diagram.

(3 marks)

Number the **two** correct statements in the correct order (**three** statements are incorrect).

- The correlation is positive and
- There is no correlation but it is
- weak
- strong
- The correlation is negative and

Select **one** box.

- As the price increases the top speed decreases.
- A high price car will have a low top speed.
- As the price increases the top speed increases.
- A high price car will have a high top speed.

(d) A new car will be releasing soon with a price of £250,000.

Emma is planning on using the line of best fit on the scatter diagram to predict the top speed of the new car.

Explain whether or not it is appropriate to use the line of best fit for this prediction.

(2 marks)

Number the **two** correct statements in the correct order (**two** statements are incorrect).

- because the point is inside the range of the data.
- because the point is outside the range of the data.
- This is appropriate
- This is not appropriate

A city council is considering adding more public transport routes.

Ethan wants to conduct a survey to learn what all the residents in the city think about the plan.

Ethan thinks that he should take a sample rather than a census.

- 2 Ethan has decided to use the electoral register as a sampling frame.

State one problem Ethan may have using the electoral register as a sampling frame.

(1 mark)

Select **one** box.

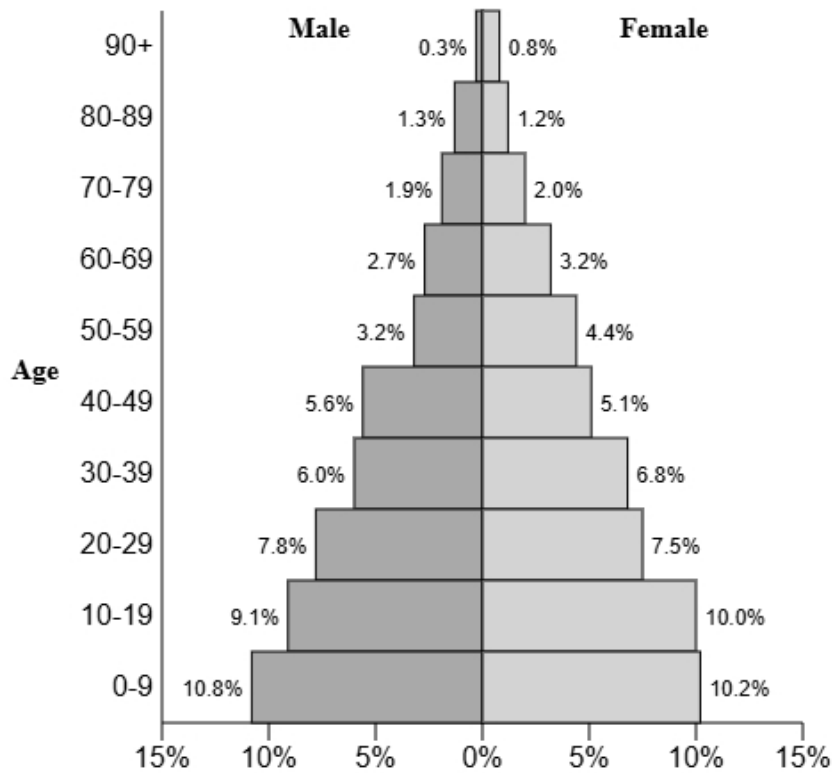
There will be too many names.

The electoral register would also include people's addresses.

Only those registered to vote would be included.

Unreliable.

3 The population pyramid below shows the percentage of males and females in each age group for the town Elderleigh.



(a) Find the age group that has 19.1% of the population.

(1 mark)

Select *one* box.

- 30-39
- 10-19
- 20-29
- 40-49

(b) Compare the percentage of the population aged 20-49 between males and females.

(1 mark)

Select **one** box.

- There are more males.
- There are more females.
- They are both the same.

(c) Give a reason why the sum of all the percentages is 99.9% and not 100%.

(1 mark)

Select **one** box.

- It is incomplete.
- There are gaps in the ages.
- The figures are wrong.
- They have been rounded.

A scientist is conducting an experiment to investigate how caffeine affects concentration.

She plans to use a matched pairs design.

Layla is one of the participants in the study.

As part of the experiment, she takes four concentration tests.

Each test has a different weighting.

The table below shows the weightings and Layla's scores for each test.

Test	Weighting	Score
A	1	15
B	2	20
C	3	18
D	4	25

- 4 Calculate the weighted mean score for Layla's four tests.

(3 marks)

Use the formula

$$\text{weighted mean} = \frac{\sum w \times s}{\sum w}$$

5 Luca works for a public health institute. He has been tasked with investigating noise levels in urban parks.

Below is a section of the spreadsheet he used to record his findings.

Noise dB(A)	Percentage of parks
$30 < n \leq 35$	8
$35 < n \leq 40$	six
$40 < n \leq 45$	8
$45 < n \leq 50$	111
$50 < n \leq 55$	57
$55 < n \leq 60$	10
<b>Total</b>	100

Luca cleans the data to create the table below.

Noise dB(A)	Percentage of parks
$30 < n \leq 35$	8
$35 < n \leq 40$	6
$40 < n \leq 45$	8
$45 < n \leq 50$	11
$50 < n \leq 55$	57
$55 < n \leq 60$	10
<b>Total</b>	100

(a) Give a reason Luca cleaned the data.

(1 mark)

Select **one** box.

- Luca must have more data.
- Luca needs the raw data.
- Data must be in the same format before it can be used

(b) Luca realised that the value of 111 in the original table was incorrect.  
Explain how Luca knew this.

(1 mark)

Select **one** box.

- The total is 100, so no value can be more than this.
- Luca removed it because it was an outlier.
- Luca wanted the data to be closer together.

(c) Use linear interpolation to work out an estimate of the median noise level.  
Round your answer to one decimal place.

(3 marks)

\_\_\_\_\_ dB(A)

6 The table shows the average monthly oil price, in dollars, over a six-month period in 2022.

It also provides selected chain base index numbers, rounded to one decimal place, for the same data.

	Jan	Feb	Mar	Apr	May	Jun
oil price	70	72	75	73	76	78
Chain base index number		102.9	104.2	97.3	104.1	

(a) Calculate the chain base index for June.

Round your answer to one decimal place.

(2 marks)

Use the formula:

$$\text{current index} = \frac{\text{current value}}{\text{previous value}} \times \text{previous index}$$

(b) i) Calculate the geometric mean of the five chain base index numbers, showing all your working and rounding your answer to one decimal place.

ii) Interpret your answer in context.

(4 marks)

For part i, use the formula:

$$\sqrt[5]{102.9 \times 104.2 \times 97.3 \times 104.1 \times 102.6}$$

Number the **two** correct statements in the correct order (**two** statements are incorrect).

- by about 2.2% per month.
- by about 2.6% per month.
- On average, oil prices decreased
- On average, oil prices increased

- 7 In a library, 60% of members are adults and 40% are children.  
Priya and Thomas plan to conduct a feedback survey.

Priya decides to use simple random sampling to select 100 participants.  
She uses the library membership database as a sampling frame, assigning a number to each participant.  
She then generates 100 random numbers and selects her sample accordingly.

Thomas decides to use quota sampling to collect a sample of 100 participants.  
He plans to sit at the library's exit until 60 adults and 40 children have been interviewed.

- (a) Give two reasons why Priya's method may **not** produce a sample of 100 participants.

(2 marks)

Select **two** boxes.

- The random numbers may not be whole numbers.
- The numbers may not have been mixed properly.
- Some people may have left the library.
- Some of the random numbers may be repeated.
- The method may produce more people than is required.

- (b) Give **two** advantages of quota sampling.

(2 marks)

Number the **two** correct statements in the correct order (**two** statements are incorrect).

- Quota sampling is cheap
- Quota sampling is the standard method
- and will allow results to be published.
- and allows for comparison between adults and children.

(c) Explain why the quota sample used by Thomas is not a random sample.

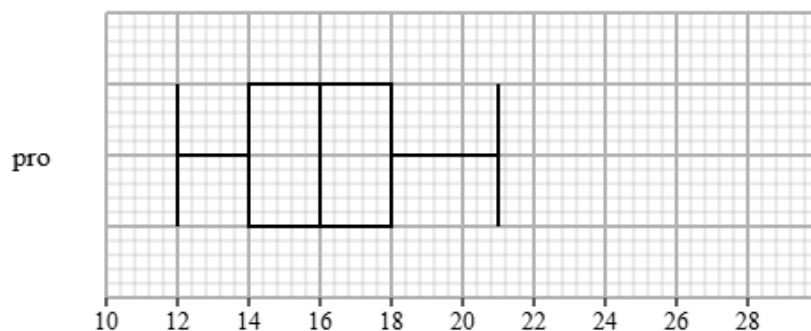
(1 mark)

Select **one** box.

- The sample was selected from a large population.
- The participants volunteered to take part.
- Only the people at the library that day are considered.
- He made sure the sample was balanced.

Liam recorded the completion times for pro and beginner runners in a 5K race.  
Both groups ran the same course.

The box plot presents data on the completion times for the pro runners.



The table gives information about the completion times for the beginner runners.

Least tall	Lower quartile	Median	Upper quartile	Most tall
17	19	20	24	30

8 Compare the two distributions of completion times.

Give three comparisons and interpret one of these comparisons.

(4 marks)

Select **one** box.

- The median is bigger.
- The median completion times for pro runners is greater than beginner runners.
- The median completion times for pro runners is lower than beginner runners.
- The median completion times for pro and beginner runners are equal.

Select **one** box.

- The IQR is bigger.
- The IQR for the completion times of the pro runners is greater than beginner runners.
- The IQR for the completion times of the pro and beginner runners are equal.
- The IQR for the completion times of the pro runners is lower than beginner runners.

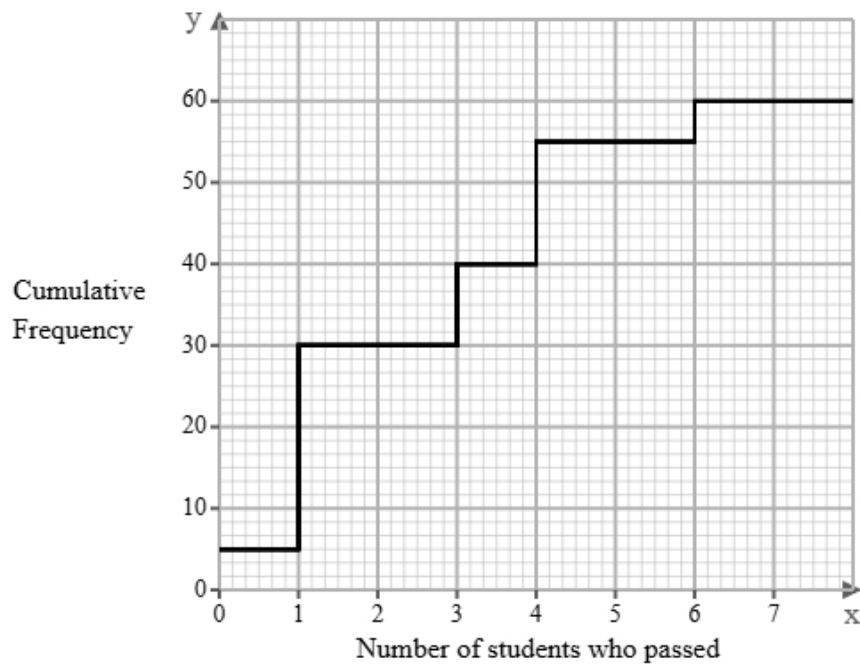
Select **one** box.

- The skews for the completion times of the pro and beginner runners are both positive.
- The skew for the completion times of the pro runners is symmetrical and the skew for the beginner runners is positive.
- The skews for the completion times of the pro and beginner runners are both symmetrical.
- The skew for the completion times of the pro runners is symmetrical and the skew for the beginner runners is negative.

Select **one** box.

- The times for the pro runners are more spread out than the beginner runners.
- The pro runners are on average slower than the beginner runners.
- The pro runners are on average faster than the beginner runners.
- The pro runners are more skewed than beginner runners.

- 9 The cumulative frequency step polygon shows information about the number of students who passed a daily maths quiz over 60 days.



- (a) Find the mode of the number of students who passed a daily maths quiz.

(1 mark)

The mode is the number that came up the most (the highest frequency).  
Look at the cumulative frequency step polygon and see where it 'jumps up' the most.

(b) Find the number of days where there were:

- i) exactly 5 students who passed.
- ii) more than 5 students who passed.

(3 marks)

The frequency is shown by how much the graph 'goes up' at each point.  
Remember, the overall frequency is 60.

i) Exactly 5 students who passed: \_\_\_\_\_

ii) More than 5 students who passed: \_\_\_\_\_

(c) In 40 days fewer than  $x$  students passed.

Find the value of  $x$

(1 mark)

Draw a line across from 40 on the graph and see where all the 'jumps up' to this line are under.

\_\_\_\_\_

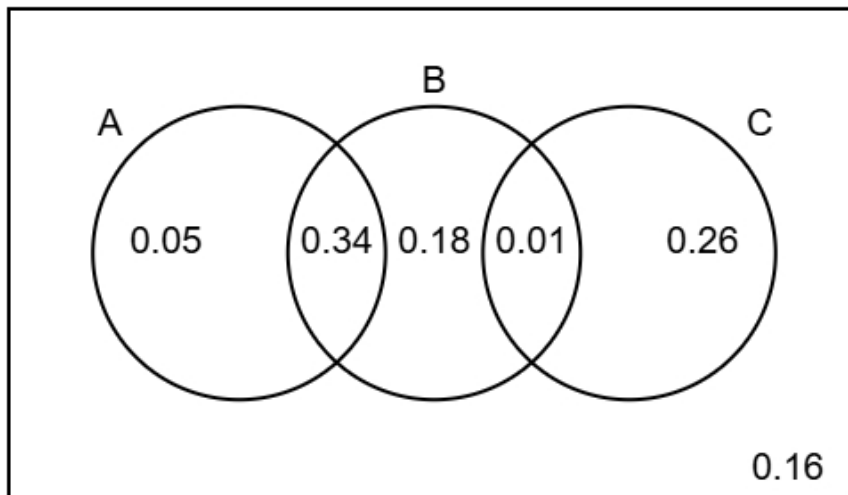
- (d) Peter believes the interquartile range of the number of students who passed is 8.  
Explain why the interquartile range for this data cannot be 8.

(1 mark)

Select **one** box.

- The range is 6, so the IQR must be less than 6.
- The range is 7, so the IQR must be less than 7.
- The range is 6, so the IQR must be more than 6.
- The range is 7, so the IQR must be more than 7.

10 The Venn diagram illustrates the probabilities associated with events A, B, and C.



(a) Identify the **two** events that are mutually exclusive, giving a reason for your answer. (2 marks)

Number the **two** correct statements in the correct order (**four** statements are incorrect).

- because they have the lowest total probability.
- B and C are mutually exclusive
- A and C are mutually exclusive
- because they do not overlap.
- A and B are mutually exclusive
- because they only overlap once.

(b) Find  $P(B)$

(1 mark)

We are looking for the probabilities inside B.

(c) Find  $P(A \text{ or } C)$

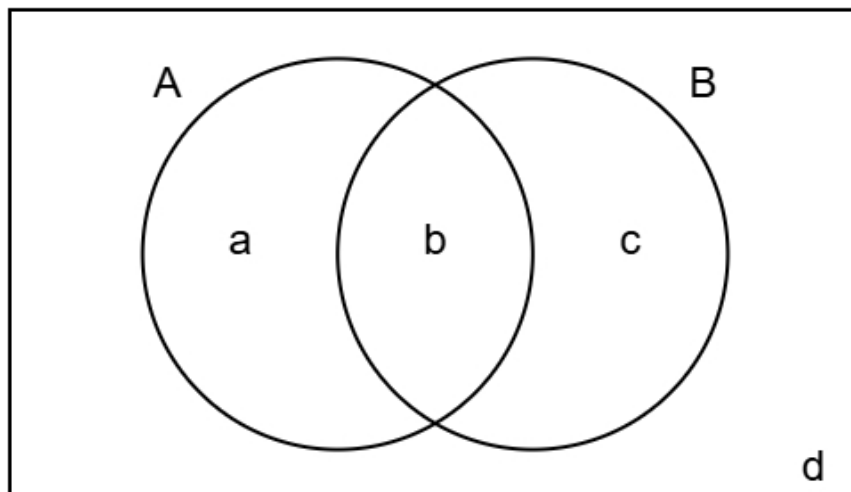
(2 marks)

We are looking for the probabilities inside A or C.

(d) Complete the Venn diagram to show **only** the probabilities for events A and B.

(2 marks)

Combine the probabilities from C into either B or the outside area.



**a** = \_\_\_\_\_                      **b** = \_\_\_\_\_

**c** = \_\_\_\_\_                      **d** = \_\_\_\_\_

Lucas is researching customer spend and average number of daily customers for 10 shops.

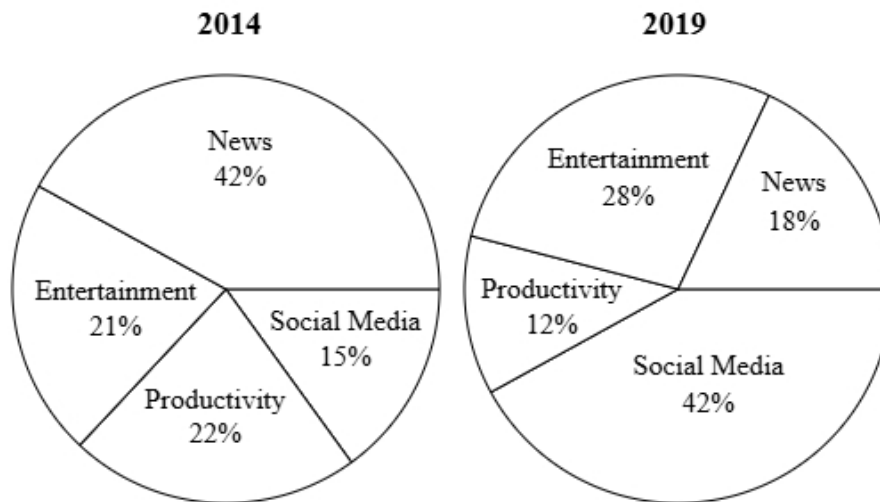
- 11** Discuss whether Lucas should have used Pearson's product moment correlation coefficient instead of Spearman's rank correlation coefficient to measure the correlation.

(3 marks)

Number the **three** correct statements in the correct order (**three** statements are incorrect).

- therefore either would be suitable.
- the PMCC only measures linear correlation
- the PMCC also measures correlation
- and Spearman's rank correlation is used for ranked data.
- Lucas should not use PMCC because
- Lucas should use PMCC because

12 The pie charts show a country's mobile phone app downloads in 2014 and 2019.



- (a) The number of news app downloads in 2014 was 1922130.  
Find the number of social media app downloads in 2014.

(2 marks)

The number of news app downloads in 2014 was 1922130 which is shown by 42% on the pie chart.

$$42\% = 1922130$$

Find 1%

Then find the number of social media app downloads (15%).

- (b) In 2014 the total number of app downloads was 4577000 (nearest thousand).  
In 2019 the total number of app downloads was 55345000 (nearest thousand).

Santi wants to use the totals to draw pie charts.

Explain, giving reasons, how Santi can use the totals to draw these pie charts.

(5 marks)

Select **one** box.

- Santi can use dual pie charts.
- Santi can use comparative pie charts.
- Santi can use a stem and leaf diagram.
- Santi can use merged pie charts.

Select **one** box.

- $\frac{55345000^2}{4577000^2}$
- $\sqrt{\frac{55345000}{4577000}}$
- $\left(\frac{55345000}{4577000}\right)^2$
- $\frac{55345000}{4577000}$

The radius of the 2014 pie chart will be \_\_\_\_\_ (2 d.p.) larger than the 2019 pie chart.

Select **one** box.

- This approach is faster than traditional pie chart methods.
- The areas will now decrease as the frequency increases, making the charts more intuitive.
- Each sector's proportion will vary depending on the total value of its respective chart.
- These pie charts have different totals, so their areas will reflect those totals while maintaining the same proportions within each chart.

**13** Changes in the cost of living in the United Kingdom are measured by the Consumer Price Index (CPI).

The table below shows the index numbers for 2011, using 2005 as the base year, for the items that contribute to the CPI.

It also shows the weightings representing how the Williams family's spending is distributed among these items.

Item	Index Number (2011)	Williams Family Weighting (%)
Food, alcoholic beverages & tobacco	110.3	14
Clothing and footwear	107.6	8
Housing and household services	108.8	29
Transport	108.7	12
Recreation, culture and communication	108.2	26
Health, education and other	108.0	11

Using 2005 as base year, the national CPI for 2011 was 110.2.

Compare the total variation in the Jones family's cost of living from 2015 to 2011 with the change in the Consumer Price Index (CPI) over the same period and explain whether your calculations allow you to determine if the Jones family is financially better off or worse off in 2011 compared to 2015.

(5 marks)

Williams family weighted mean = a (1 decimal place)

$a < 110.2$

**a** = \_\_\_\_\_

*Select **one** box.*

- Cost of living has risen less for the Williams family.
- Cost of living has risen more for the Williams family.
- Cost of living has risen in line with the national CPI for the Williams family.

*Select **one** box.*

- We do not know about the Williams family's income so cannot determine if they are better off.
- The Williams family are better off because they have a better cost of living.
- The Williams family are worse off because they have a better cost of living.

**14** A company produces chocolate chip cookies.

The cookies have a target mass of 60 g.

The company uses quality assurance to monitor the mass of each cookie.

Samples of the cookies are taken from the production line at regular intervals and the mean mass of the cookies in each sample is found.

The sample means should be normally distributed with a mean of 60 g and a standard deviation of 1.2 g.

**(a)** Find the upper action limit for the sample means for the cookies.

(2 marks)

Use the formula for the upper action limit

$$\text{Upper action limit} = \mu + 3\sigma$$

\_\_\_\_\_ g

**(b)** The upper action limit will be set closer to the target mass of 60 g.

Describe the effect this will have on the frequency of production process stoppages.

(1 mark)

Select *one* box.

The production process will work faster.

The production process would reset more frequently.

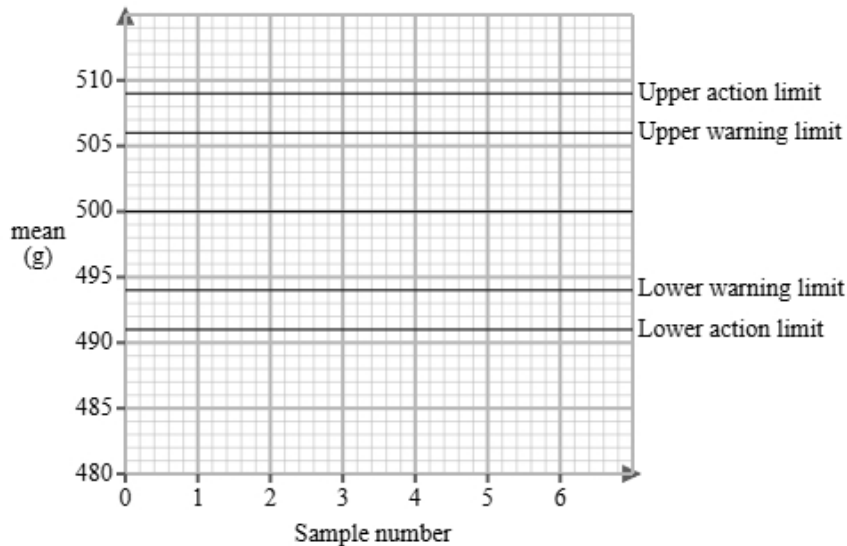
The production process would reset less frequently.

This is unrelated to the production process so will have no effect.

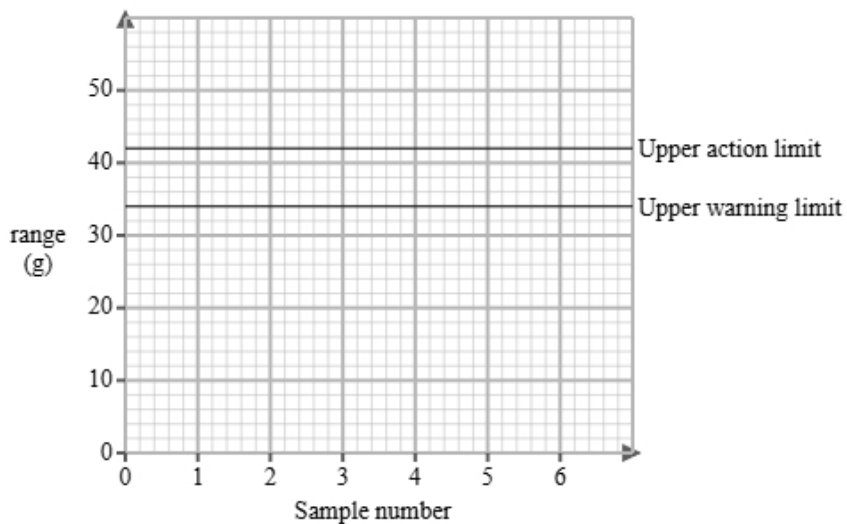
(c) The company also produces sponge cake and uses quality assurance to monitor the mass of each cake.

Here are the control charts for the sample means and for the sample ranges of the masses of the cakes.

**Control chart for means**



**Control chart for ranges**



A sample is taken and is found to have a mean of 496 g and a range of 31 g.

Use the sample mean and range to determine what action, if any, needs to be taken.

(2 marks)

Number the **two** correct statements in the correct order (**two** statements are incorrect).

- The production process should continue because the range is within the warning limit
- The production process should be stopped as the range is outside the upper action limit
- although the mean is outside the upper warning limit but not the action limit.
- and the mean is within the upper warning limit.

- 15** A study took place in Japan to find if there was a relationship between screen time and sleep duration of teenagers.

The researchers found the equations of the regression lines for the relationship between screen time ( $x$  hours) and sleep duration ( $y$  hours) for male teenagers and female teenagers on school nights and weekend nights.

The table below gives the equations of the regression lines.

	school nights	weekend nights
male teenagers	$y = -0.4x + 7.5$	$y = -0.2x + 8.2$
female teenagers	$y = -0.3x + 7.8$	$y = -0.15x + 8.5$

- (a) Compare the relationships between screen time and sleep duration in male and female teenagers. Include in your comparisons reference to whether it is a school night or weekend night.

(3 marks)

Select the **three** correct statements (**three** statements are incorrect).

- As screen time increased, the sleep duration decreased more rapidly on weekend nights compared to school nights.
- For all teenagers, an increase in screen time led to a decrease in sleep duration.
- For all teenagers, an increase in screen time led to a increase in sleep duration.
- The sleep duration of female teenagers decreased more per hours of screen time than male teenagers.
- As screen time increased, the sleep duration decreased more rapidly on school nights compared to weekend nights.
- The sleep duration of male teenagers decreased more per hours of screen time than female teenagers.

(b) The researchers would like to use a normal distribution as a model for the sleep duration of male teenagers on school nights.

i) Explain how they could check whether a normal distribution is a suitable model by drawing a histogram.

ii) Explain how they could check whether a normal distribution is a suitable model by calculating the averages and the standard deviation.

(3 marks)

Select **one** box.

If the histogram shows different heights, a normal distribution could be a suitable model.

If the histogram is a linear shape, a normal distribution could be a suitable model.

If the histogram is a bell shape, a normal distribution could be a suitable model.

If the histogram shows equal heights, a normal distribution could be a suitable model.

Number the **two** correct statements in the correct order (**two** statements are incorrect).

If the mean, median and mode are equal

and 95% of data is within 1 standard deviation from the mean, a normal distribution could be a suitable model.

If the mean, median and mode are larger than the standard deviation

and 68% of data is within 1 standard deviation from the mean, a normal distribution could be a suitable model.

**16** A fair coin is tossed 4 times.

The number of heads obtained is recorded.

**(a)** Identify two conditions needed so that a binomial distribution is a suitable model for the number of heads recorded.

(2 marks)

Select *two* boxes.

- The coin is biased.
- Each of the emails are independent.
- The chance of getting a head remains constant.
- A different coin is used for each trial.

**(b)** Calculate the probability, as a fraction, that all 4 of the coins land on heads.

(2 marks)

Work out  $p^n$  where  $p$  is the given probability and  $n$  is the amount.

---

---

---

---

**(c)** Calculate the probability, as a fraction, that at least 2 of the coins land on heads.

(3 marks)

You can use Pascal's triangle, or your calculator to find  $P(X < 2)$  then take this away from 1.

---

---

---

---