

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

Edexcel Higher - 2026

Higher Tier

Variant 4

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.

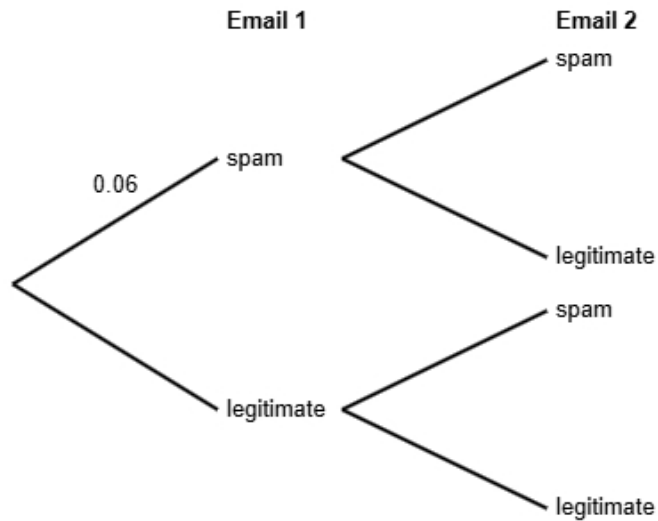
**Copyright**

- 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 A study shows that 6% of emails received by a certain email provider are spam.  
All other emails are legitimate.  
Maria receives two emails in her inbox.  
She does not know if each email is spam or legitimate.



- (a) Complete the probability tree diagram.

(2 marks)

The branches for each stage must add up to 1.  
Each test is independent so will have the same probabilities.

(b) Find the probability that both of Maria's emails are legitimate.

(2 marks)

You will need to find  $P(\text{legitimate})$  AND  $P(\text{legitimate})$ .

Remember, AND means  $\times$  in probability.

(c) Maria states that the probability that exactly one email is spam is less than 12%

Find out whether or not Maria is correct.

(3 marks)

Find the probability of exactly one email is spam (there are two outcomes on the tree diagram).

Select **one** box.

The probability that exactly one email is spam is less than 12%, so Maria is not correct.

The probability that exactly one email is spam is less than 12%, so Maria is correct.

The probability that exactly one email is spam is more than 12%, so Maria is correct.

The probability that exactly one email is spam is more than 12%, so Maria is not correct.

2 The table shows information about bicycles for sale in Birmingham.

gear types	number of bicycles
1	175
2	350
3	200
4	450
5 or more	425
Total	1600

A researcher wants to investigate the price of these bicycles and takes a stratified sample of 64 bicycles according to the gear types.

(a) The researcher says the mode of the gear types for these bicycles is 4.

Explain how the researcher knows this.

(1 mark)

Select **one** box.

4 gears has the highest frequency.

4 bicycles has the highest frequency.

4 is the difference between the largest and smallest number.

4 is the middle number.

(b) Work out the number of bicycles in the sample for each gear type.

gear types	number of bicycles in the sample
1	
2	
3	
4	
5 or more	

(3 marks)

Find the divisor for the stratified sample:  $\frac{\text{total}}{\text{sample size}}$

Divide each frequency by this number to find the required sample in each group

(c) Describe how the 64 bicycles in the sample should be selected.

(3 marks)

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

- Ensure that all 1600 bicycles are included in the sample.
- Select the first 64 bicycles.
- Generate random numbers, remove repeats or numbers out of range.
- Use a sampling frame for each strata.
- Complete two of the strata.
- Number each of the bicycles, and then use the random numbers to select the required amount of bicycles.

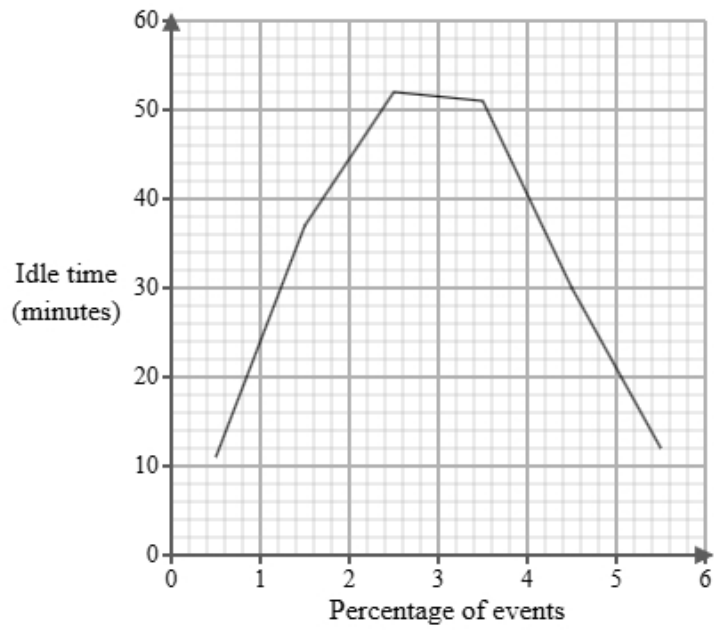
- 3 Tom works for a logistics company. He has been tasked with investigating delivery vehicle idle times. Below is a section of the spreadsheet he used to record his findings.

Idle time (Minutes)	Percentage of events
$0 < d \leq 1$	6
$1 < d \leq 2$	four
$2 < d \leq 3$	9
$3 < d \leq 4$	119
$4 < d \leq 5$	47
$5 < d \leq 6$	15
Total	100

Tom cleans the data to create the table below.

Idle time (Minutes)	Percentage of events
$0 < d \leq 1$	6
$1 < d \leq 2$	4
$2 < d \leq 3$	9
$3 < d \leq 4$	19
$4 < d \leq 5$	47
$5 < d \leq 6$	15
Total	100

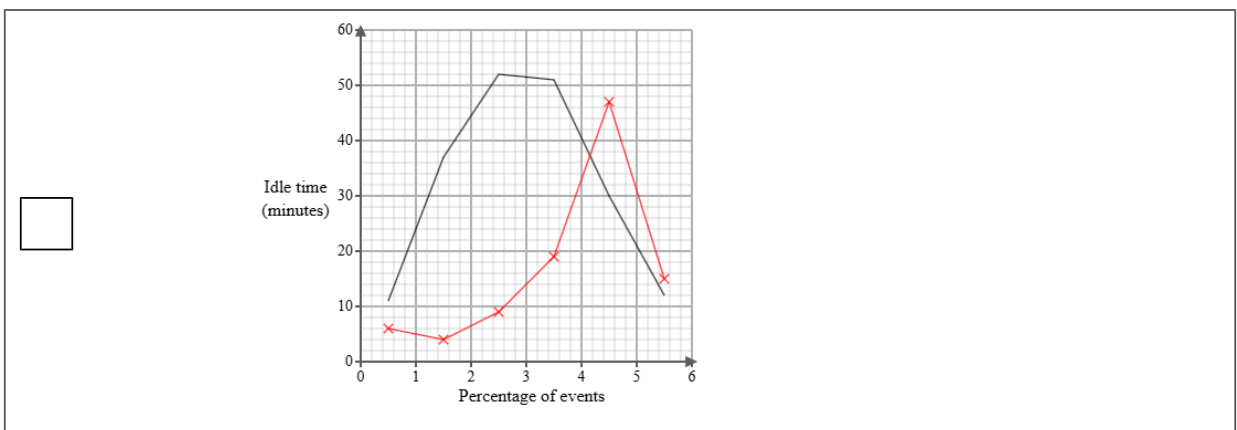
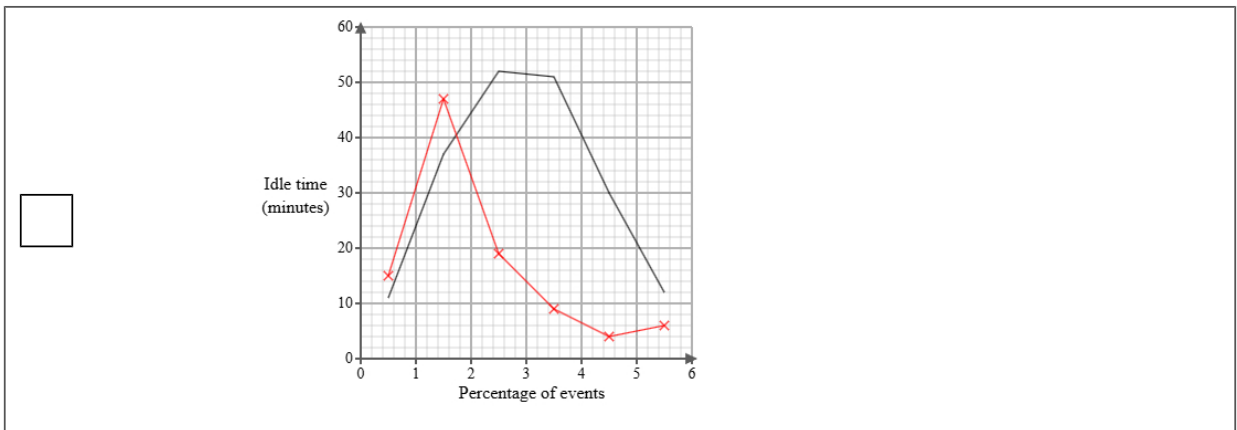
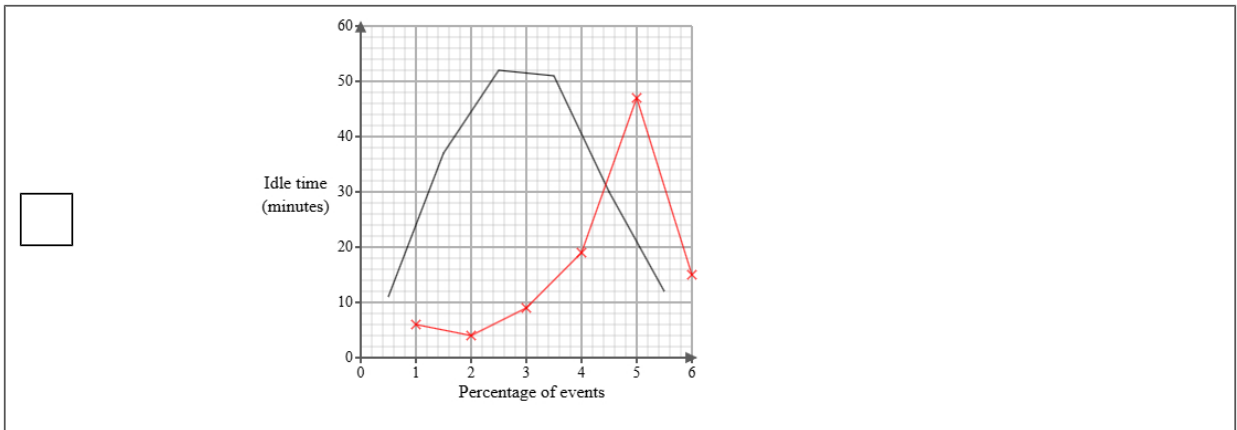
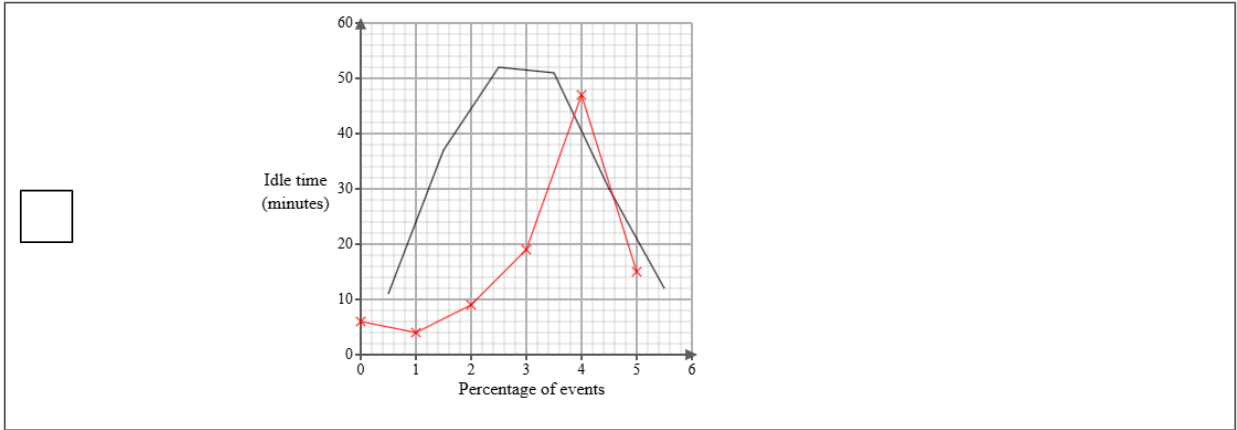
A frequency polygon has been drawn for delivery vehicle idle times at a rival firm.



- i) On the same graph, draw the frequency polygon for delivery vehicle idle times for Tom's company.
- ii) Using the two frequency polygons, compare the skew of the distributions and explain what your comparison means in context.

(4 marks)

Select the correct answer.



Select the **two** correct statements (**four** statements are incorrect).

- The distribution of delivery vehicle idle times at a rival firm is symmetrical whereas the distribution of delivery vehicle idle times for Tom's company is negatively skewed.
- The distribution of delivery vehicle idle times at a rival firm is positively skewed whereas the distribution of delivery vehicle idle times for Tom's company is symmetrical.
- This means that idle times for the rival firm were equally spread out on either side of the median and idle times for the for Tom's company were mainly at the upper end of the distribution.
- The distribution of delivery vehicle idle times at a rival firm is negatively skewed whereas the distribution of delivery vehicle idle times for Tom's company is symmetrical.
- This means that idle times for the rival firm were mainly at the lower end of the distribution and idle times for the for Tom's company were mainly at the upper end of the distribution.
- This means that idle times for the rival firm were mainly at the upper end of the distribution and idle times for the for Tom's company were equally spread out on either side of the median.

- 4 The table shows information about the retail price index (RPI) and cinema ticket price (£) in the United Kingdom for Jan 1990, Jan 2000 and Jan 2010.

	Jan 1990	Jan 2000	Jan 2010
retail price index	100	145	177
cinema ticket price (£)	2.81	3.99	5.97

Describe how the increase in cinema ticket price (£) compares with the RPI over the ten years to Jan 2000 and over the twenty years to Jan 2010.

(5 marks)

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

- Between Jan 1990 and Jan 2000 the change in price was less than the RPI.
- Between Jan 1990 and Jan 2000 the change in price was more than the RPI.
- $\frac{3.99}{145} \times 100 = 3$  (nearest integer)
- Between Jan 1990 and Jan 2010 the change in price was less than the RPI.
- $\frac{5.97}{2.81} \times 100 = 212$  (nearest integer)
- $\frac{5.97}{177} \times 100 = 3$  (nearest integer)
- Between Jan 1990 and Jan 2010 the change in price was more than the RPI.
- $\frac{3.99}{2.81} \times 100 = 142$  (nearest integer)

5 Daniel is investigating how the distance from city centre in km,  $x$ , affects the selling price (£),  $y$  for two types of houses, detached houses and semi-detached houses.

He found ten houses of each type and recorded their distance from city centre and selling price and plotted each on scatter diagrams.

He then drew a line of best fit on each diagram and found the gradient and y-intercept of each line.

Here are the results:

House type	Gradient of line of best fit	y-intercept of line of best fit
Detached houses	-5000	450000
Semi-detached houses	-4200	380000

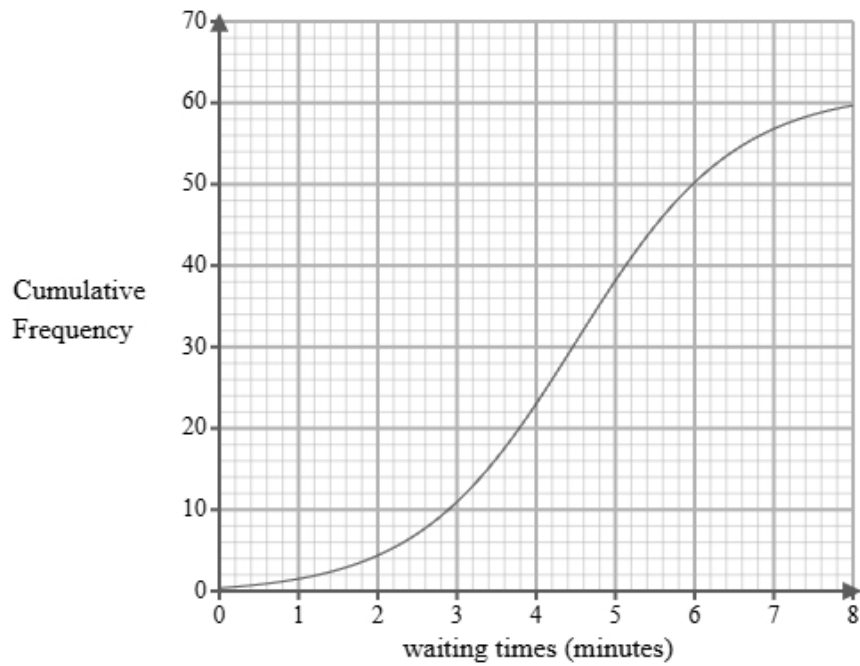
Interpret and compare these results in context.

(5 marks)

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

- House type Semi-detached houses reduces in selling price by £4200 per km.
- House type Semi-detached houses has a greater initial selling price.
- House type Detached houses reduces in selling price less than House type Semi-detached houses.
- House type Detached houses reduces in selling price more per km than House type Semi-detached houses.
- Both houses increase in selling price as the distance from city centre increase.
- House type Detached houses changes in selling price by £450000 per km.
- House type Detached houses has a greater initial selling price.
- House type Detached houses reduces in selling price by £5000 per km.
- Both houses decrease in selling price as the distance from city centre increase.
- House type Semi-detached houses changes in selling price by £380000 per km.

- 6 A researcher measures the waiting times, in minutes, of 60 customers at a café.  
A cumulative frequency diagram is drawn from the data.



Complete the table below from the cumulative frequency diagram.

Lower quartile	Median	Upper quartile

(2 marks)

Select the correct answer.

<input type="checkbox"/>	Lower quartile	Median	Upper quartile
	2.1	5	7.2

<input type="checkbox"/>	Lower quartile	Median	Upper quartile
	2.9	5	6.3

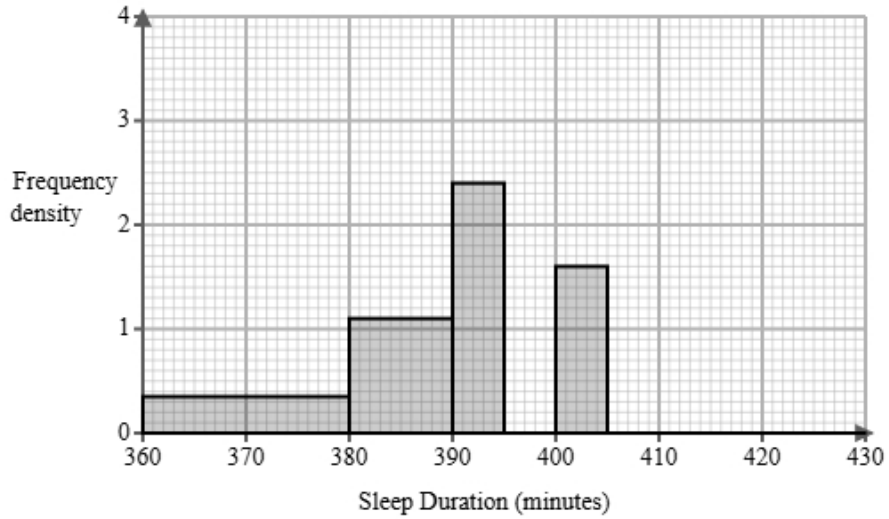
<input type="checkbox"/>	Lower quartile	Median	Upper quartile
	3.4	4.5	5.5

<input type="checkbox"/>	Lower quartile	Median	Upper quartile
	3.8	4.5	4.7

7 The amount of sleep is recorded in minutes.

A sleep researcher is analysing the duration of sleep that 60 adult women get following a 12-hour period of fasting.

The partially completed histogram and grouped frequency table provide details about the sleep durations recorded in the study.



Sleep Duration $s$ (minutes)	Frequency
$360 < s \leq 380$	7
$380 < s \leq 390$	11
$390 < s \leq 395$	
$395 < s \leq 400$	12
$400 < s \leq 405$	
$405 < s \leq 430$	10

(a) Complete the table using the information from the histogram.

(2 marks)

Use the frequency density formula

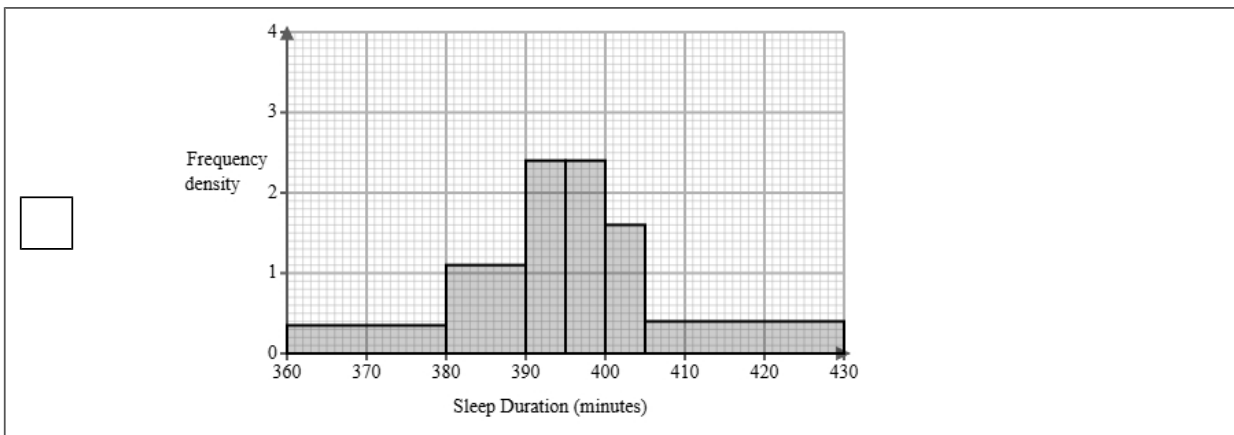
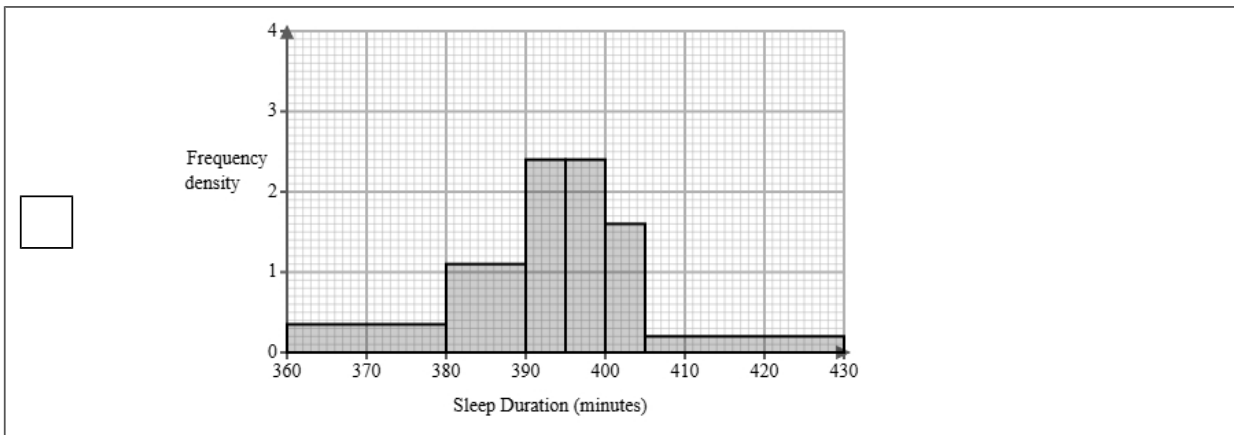
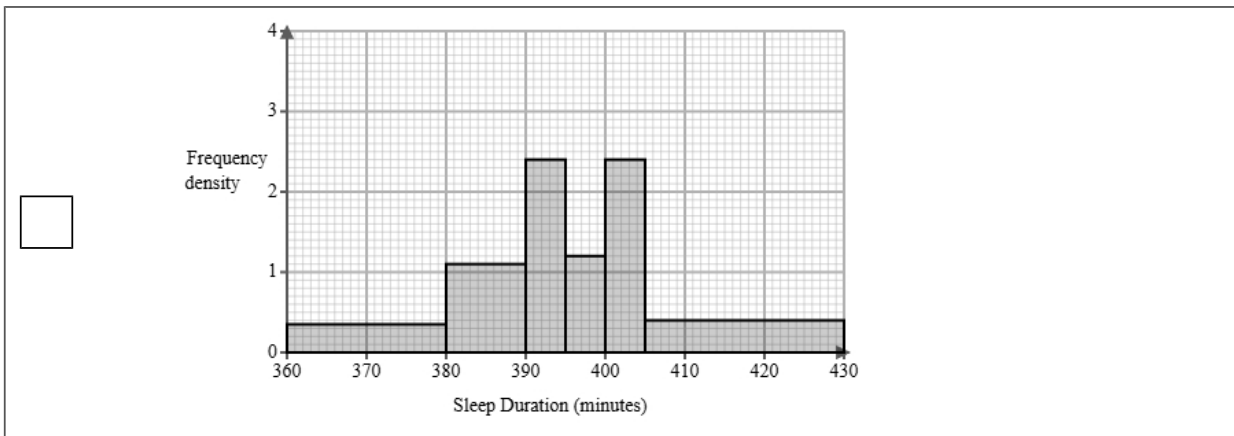
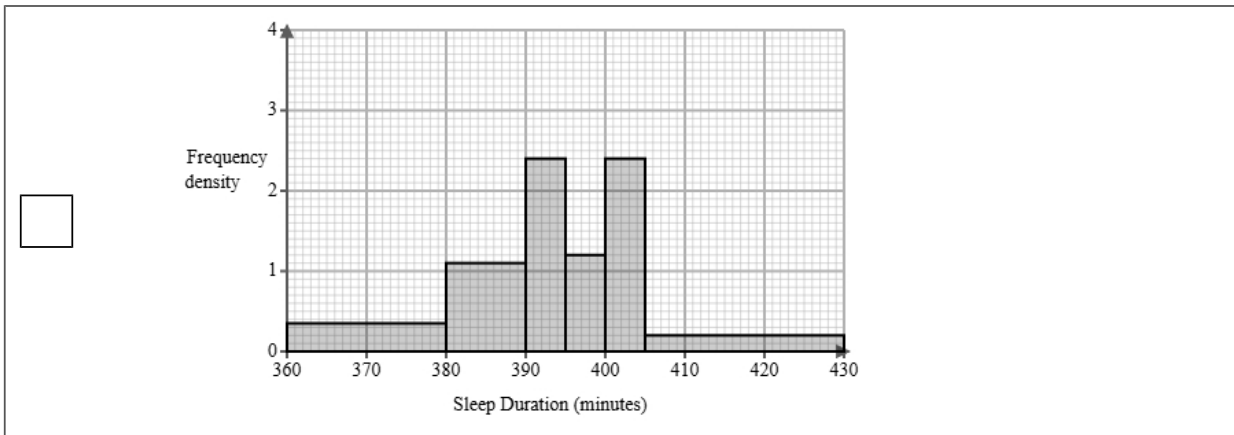
$$\text{frequency} = \text{frequency density} \times \text{class width}$$

Sleep Duration $s$ (minutes)	Frequency
$360 < s \leq 380$	7
$380 < s \leq 390$	11
$390 < s \leq 395$	_____
$395 < s \leq 400$	12
$400 < s \leq 405$	_____
$405 < s \leq 430$	10

(b) Complete the histogram using the information from the table.

(2 marks)

Select the correct answer.



(c) The sleep researcher finds the following summary statistics for the data.

$$\sum s = 23730$$

$$\sum s^2 = 9393238$$

$$n = 60$$

Explain whether or not there may be any outliers in the sleep researcher's data by calculating the limits for outliers using the mean and standard deviation.

You must round all values to 2 decimal places.

(5 marks)

mean = \_\_\_\_\_

standard deviation = \_\_\_\_\_

lower outlier limit = \_\_\_\_\_

upper outlier limit = \_\_\_\_\_

Select **one** box.

There are outliers because the limits are within the data.

There are no outliers because the limits are outside the ranges of the data.

It is possible that there is an outlier as the lower outlier limit is within the group  $360 < s \leq 380$ .

- (d) A different sleep researcher is analysing the duration of sleep adult men get following a 12-hour period of fasting.

They find the following summary statistics for the data.

mean = 347.54

median = 380

standard deviation = 11.4

Calculate and interpret the skew for the men.

You must round your answer to 2 decimal places.

(3 marks)

Skew = \_\_\_\_\_

Select **one** box.

- There is a negative skew showing the data is not normally distributed.
- The skew shows there is a negative correlation.
- The skew shows that the three averages are equal.

- (e) Find the class interval that contains the 30th percentile.

(1 mark)

Select **one** box.

- $390 < s \leq 395$
- $380 < s \leq 390$
- $395 < s \leq 400$
- $360 < s \leq 380$
- $400 < s \leq 405$

- 8 Ethan is researching the final league position of basketball teams in a local league and the mean heights of all the players in each team.

The table below shows the data collected.

Team	Mean Height (cm)	Height Rank	Final Position	d	d <sup>2</sup>
Aylesbury	168	1	1	0	0
Bedford	180	7	7	0	0
Chatham	181	8	8	0	0
Dorking	171	3	3	0	0
Eastleigh	179	6	5		
Farnborough	176	5	6		
Guildford	169	2	4		
Havant	173	4	2		

- (a) Ethan would like to see if there is an association between the final position and the mean value.

Suggest a diagram that Ethan could draw.

(1 mark)

Select **one** box.

Histogram

Cumulative frequency diagram

Scatter diagram

Venn diagram

(b) i) Calculate Spearman's rank correlation coefficient from the data in the table and leave your answer to 2 decimal places.

ii) Interpret your answer to **part i**, referring to the effects of any anomalous data.

(5 marks)

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

- As the mean height of players increases, the position of the team in the league is lower.
- Anomalous data would decrease the correlation.
- As the mean height of players increases, the position of the team in the league is higher.
- Anomalous data would increase the correlation.

(c) Ethan used Spearman's rank correlation coefficient to analyse the data.

Emily suggests that Ethan could have used Pearson's product moment correlation coefficient.

Discuss whether using Pearson's product moment correlation coefficient is appropriate for this data.

(3 marks)

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

- PMCC measures linear correlation.
- Spearman's rank correlation is influenced by outliers.
- Ethan used the correct method, Emily's suggestion is not appropriate.
- Emily's suggestion is more appropriate than Ethan's method.
- PMCC compares bivariate data
- Spearman's rank correlation is used for ranked data

- 9 Tom works for a logistics company. He has been tasked with investigating delivery vehicle idle times. Below is a section of the spreadsheet he used to record his findings.

Idle time (Minutes)	Percentage of events
$0 < d \leq 1$	6
$1 < d \leq 2$	four
$2 < d \leq 3$	9
$3 < d \leq 4$	119
$4 < d \leq 5$	47
$5 < d \leq 6$	15
Total	100

Tom cleans the data to create the table below.

Idle time (Minutes)	Percentage of events
$0 < d \leq 1$	6
$1 < d \leq 2$	4
$2 < d \leq 3$	9
$3 < d \leq 4$	19
$4 < d \leq 5$	47
$5 < d \leq 6$	15
Total	100

- (a) Give a reason Tom cleaned the data.

(1 mark)

Select **one** box.

- The data should be spread evenly between each group.
- One of the percentages says four not 4.
- It should show frequencies not percentages.

(b) Use linear interpolation to work out an estimate of the median idle time.

Round your answer to one decimal place.

(3 marks)

Find the group where the 50th value is in

You can use the interpolation formula to find the median

$$\text{estimated median} = L + \frac{\frac{n}{2} - F}{f} \times w$$

lower boundary = L

number of values = n

cumulative frequency before group = F

frequency of group = f

width of group = w

\_\_\_\_\_ minutes

- 10 Ethan collected the steps for remote and on-site workers in an hour within their day. Both groups recorded their steps over the same period. The box plot presents data on the steps for the remote workers.



The table gives information about the steps for the on-site workers.

Least tall	Lower quartile	Median	Upper quartile	Most tall
300	550	800	850	900

Compare the two distributions of steps.

Give three comparisons and interpret one of these comparisons.

(4 marks)

Select **one** box.

- The median is bigger.
- The median steps for remote workers is greater than on-site workers.
- The median steps for remote workers is lower than on-site workers.
- The median steps for remote and on-site workers are equal.

Select **one** box.

- The IQR is bigger.
- The IQR for the steps of the remote workers is greater than on-site workers.
- The IQR for the steps of the remote workers is lower than on-site workers.
- The IQR for the steps of the remote and on-site workers are equal.

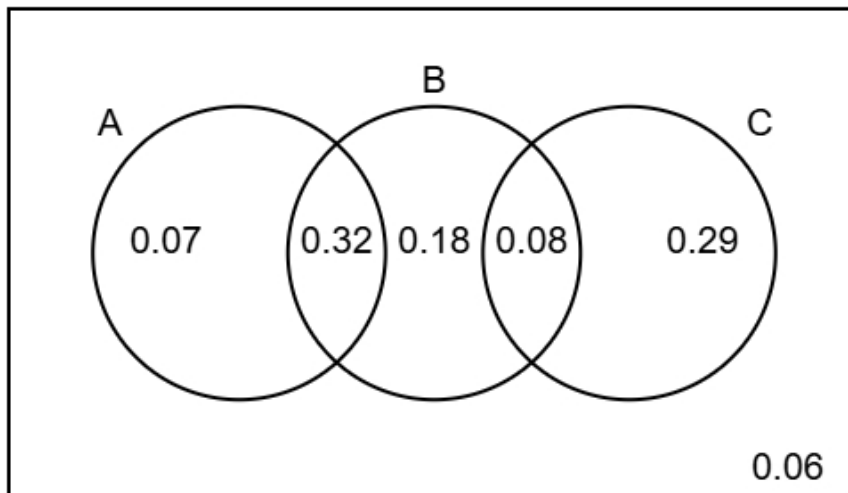
Select **one** box.

- The skews for the steps of the remote and on-site workers are both positive.
- The skew for the steps of the remote workers is symmetrical and the skew for the on-site workers is positive.
- The skew for the steps of the remote workers is symmetrical and the skew for the on-site workers is negative.
- The skews for the steps of the remote and on-site workers are both symmetrical.

Select **one** box.

- The steps for the remote workers are more spread out than the on-site workers.
- The remote workers on average walk less than the on-site workers.
- The remote workers on average walk more than the on-site workers.
- The remote workers are more skewed than on-site workers.

11 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.
- because they only overlap once.
- because they do not overlap.
- B and C are mutually exclusive
- A and C are mutually exclusive
- A and B are mutually exclusive

(b) Find  $P(B)$

(1 mark)

We are looking for the probabilities inside B.

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

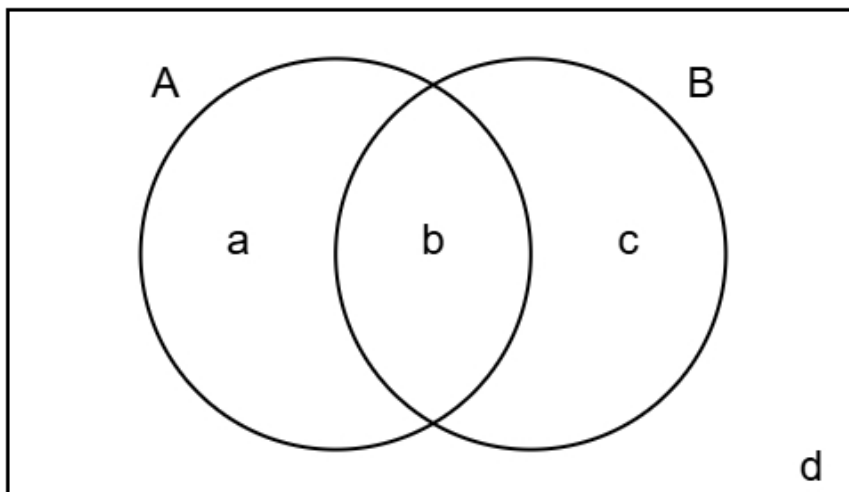
(2 marks)

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

12 The figures below show the amount, in millions, of tourists who visited USA between 2009 and 2015

72 75 78 74 80 83 85

The table gives a summary of the amount, in millions, of tourists who visited Italy between 2009 and 2015

Mean	Standard Deviation	Largest Amount
81	4	83

Compare the amount of tourists in USA and Italy between 2009 and 2015

You may use:

$$72^2 + 75^2 + 78^2 + 74^2 + 80^2 + 83^2 + 85^2 = 42883$$

(5 marks)

Find the mean for USA

$$\text{mean} = \frac{\text{sum}}{\text{amount}}$$

Find the standard deviation for USA

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

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

On average more tourists visited USA than Italy because the mean was greater for USA.

The amount of tourists varied less for USA than Italy because the standard deviation was smaller for USA.

On average less tourists visited USA than Italy because the mean was smaller for USA.

The amount of tourists varied more for USA than Italy because the standard deviation was greater for USA.

**13** The daily screen times of a group of teenagers have a mean of 3.7 hours and a standard deviation of 1.2 hours.

(a) David is teenager with a standardised score of 0.

Find David's daily screen time.

(1 mark)

A standardized score of 0 indicates that the value is equal to the mean of the distribution.

\_\_\_\_\_ hours

(b) Ben and Charlie are both teenagers in the group.

Ben's standardised score for daily screen time is 0.6 hours.

Charlie's standardised score for daily screen time is -1.1 hours.

Ben had a higher screen time than Charlie.

How much more screen time is spent by Ben?

(3 marks)

Rearrange the formula to make the value the subject

$$\text{Standardised score} = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

\_\_\_\_\_ hours

(c) Kieran takes a sample of 4 teenagers from the group.

He wants to calculate the standardised score for the sample mean of their ages.

(i) Discuss the appropriateness of using 3.7 hours as the mean in the calculation of the standardised score,

(ii) Discuss the appropriateness of using 1.2 hours as the standard deviation in the calculation of the standardised score.

(4 marks)

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

Using 3.7 hours as the mean is appropriate

because the sample mean will be close to the population mean.

because the sample mean will be smaller than the population mean.

Using 3.7 hours as the mean is not appropriate

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

Using 1.2 hours as the standard deviation is not appropriate

because the sample mean will be more closely distributed than the individual values.

because the sample standard deviation will be close to the population standard deviation.

Using 1.2 hours as the standard deviation is appropriate