3. Stav is studying the large data set for September 2015

He codes the variable Daily Mean Pressure, x, using the formula $y = x - 1010$

The data for all 30 days from Hurn are summarised by

$$\sum y = 214 \qquad \sum y^2 = 5912$$

(a) State the units of the variable x

(1)

(b) Find the mean Daily Mean Pressure for these 30 days.

(2)

(c) Find the standard deviation of Daily Mean Pressure for these 30 days.

(3)

Stav knows that, in the UK, winds circulate

- in a **clockwise** direction around a region of **high** pressure
- in an **anticlockwise** direction around a region of **low** pressure

The table gives the Daily Mean Pressure for 3 locations from the large data set on 26/09/2015

Location	Heathrow	Hurn	Leuchars
Daily Mean Pressure	1029	1028	1028
Cardinal Wind Direction			

The Cardinal Wind Directions for these 3 locations on 26/09/2015 were, in random order,

W NE E

You may assume that these 3 locations were under a single region of pressure.

(d) Using your knowledge of the large data set, place each of these Cardinal Wind Directions in the correct location in the table.
Give a reason for your answer.

(2)

1. (a) State one disadvantage of using quota sampling compared with simple random sampling.

(1)

In a university 8% of students are members of the university dance club.

A random sample of 36 students is taken from the university.

The random variable X represents the number of these students who are members of the dance club.

(b) Using a suitable model for X, find

(i) $P(X = 4)$

(ii) $P(X \geqslant 7)$

(3)

Only 40% of the university dance club members can dance the tango.

(c) Find the probability that a student is a member of the university dance club and can dance the tango.

(1)

A random sample of 50 students is taken from the university.

(d) Find the probability that fewer than 3 of these students are members of the university dance club and can dance the tango.

(2)

4. Charlie is studying the time it takes members of his company to travel to the office. He stands by the door to the office from 08 40 to 08 50 one morning and asks workers, as they arrive, how long their journey was.

(a) State the sampling method Charlie used.

(1)

(b) State and briefly describe an alternative method of non-random sampling Charlie could have used to obtain a sample of 40 workers.

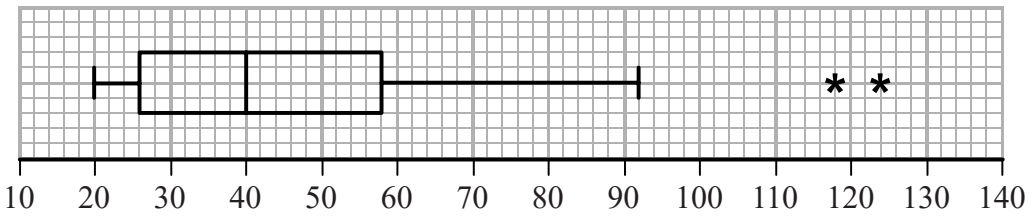
(2)

Taruni decided to ask every member of the company the time, x minutes, it takes them to travel to the office.

(c) State the data selection process Taruni used.

(1)

Taruni's results are summarised by the box plot and summary statistics below.



Journey time (minutes)

$n = 95 \quad \sum x = 4133 \quad \sum x^2 = 202\,294$

(d) Write down the interquartile range for these data.

(1)

(e) Calculate the mean and the standard deviation for these data.

(3)

(f) State, giving a reason, whether you would recommend using the mean and standard deviation or the median and interquartile range to describe these data.

(2)

Rana and David both work for the company and have both moved house since Taruni collected her data.

Rana's journey to work has changed from 75 minutes to 35 minutes and David's journey to work has changed from 60 minutes to 33 minutes.

Taruni drew her box plot again and only had to change two values.

(g) Explain which two values Taruni must have changed and whether each of these values has increased or decreased.

(3)

4. Magali is studying the mean total cloud cover, in oktas, for Leuchars in 1987 using data from the large data set. The daily mean total cloud cover for all 184 days from the large data set is summarised in the table below.

Daily mean total cloud cover (oktas)	0	1	2	3	4	5	6	7	8
Frequency (number of days)	0	1	4	7	10	30	52	52	28

One of the 184 days is selected at random.

(a) Find the probability that it has a daily mean total cloud cover of 6 or greater.

(1)

Magali is investigating whether the daily mean total cloud cover can be modelled using a binomial distribution.

She uses the random variable X to denote the daily mean total cloud cover and believes that $X \sim \text{B}(8, 0.76)$

Using Magali's model,

(b) (i) find $\text{P}(X \geqslant 6)$

(2)

(ii) find, to 1 decimal place, the expected number of days in a sample of 184 days with a daily mean total cloud cover of 7

(2)

(c) Explain whether or not your answers to part (b) support the use of Magali's model.

(1)

There were 28 days that had a daily mean total cloud cover of 8
For these 28 days the daily mean total cloud cover for the **following** day is shown in the table below.

Daily mean total cloud cover (oktas)	0	1	2	3	4	5	6	7	8
Frequency (number of days)	0	0	1	1	2	1	5	9	9

(d) Find the proportion of these days when the daily mean total cloud cover was 6 or greater.

(1)

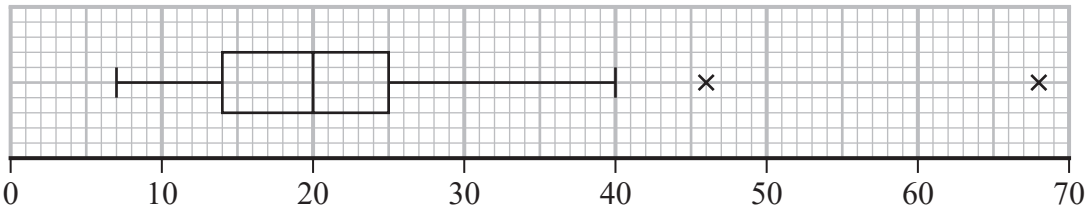
(e) Comment on Magali's model in light of your answer to part (d).

(2)

3. Each member of a group of 27 people was timed when completing a puzzle.

The time taken, x minutes, for each member of the group was recorded.

These times are summarised in the following box and whisker plot.



Time (x minutes)

(a) Find the range of the times.

(1)

(b) Find the interquartile range of the times.

(1)

For these 27 people $\sum x = 607.5$ and $\sum x^2 = 17\,623.25$

(c) calculate the mean time taken to complete the puzzle,

(1)

(d) calculate the standard deviation of the times taken to complete the puzzle.

(2)

Taruni defines an outlier as a value more than 3 standard deviations above the mean.

(e) State how many outliers Taruni would say there are in these data, giving a reason for your answer.

(1)

Adam and Beth also completed the puzzle in a minutes and b minutes respectively, where $a > b$.
When their times are included with the data of the other 27 people

- the median time increases
- the mean time does not change

(f) Suggest a possible value for a and a possible value for b, explaining how your values satisfy the above conditions.

(3)

(g) Without carrying out any further calculations, explain why the standard deviation of all 29 times will be lower than your answer to part (d).

(1)

SECTION A: STATISTICS

Answer ALL questions. Write your answers in the spaces provided.

1. The number of hours of sunshine each day, y, for the month of July at Heathrow are summarised in the table below.

Hours	$0 \leqslant y < 5$	$5 \leqslant y < 8$	$8 \leqslant y < 11$	$11 \leqslant y < 12$	$12 \leqslant y < 14$
Frequency	12	6	8	3	2

A histogram was drawn to represent these data. The $8 \leqslant y < 11$ group was represented by a bar of width 1.5 cm and height 8 cm.

(a) Find the width and the height of the $0 \leqslant y < 5$ group.

(3)

(b) Use your calculator to estimate the mean and the standard deviation of the number of hours of sunshine each day, for the month of July at Heathrow.
Give your answers to 3 significant figures.

(3)

The mean and standard deviation for the number of hours of daily sunshine for the same month in Hurn are 5.98 hours and 4.12 hours respectably.
Thomas believes that the further south you are the more consistent should be the number of hours of daily sunshine.

(c) State, giving a reason, whether or not the calculations in part (b) support Thomas' belief.

(2)

(d) Estimate the number of days in July at Heathrow where the number of hours of sunshine is more than 1 standard deviation above the mean.

(2)

Helen models the number of hours of sunshine each day, for the month of July at Heathrow by $N(6.6, 3.7^2)$.

(e) Use Helen's model to predict the number of days in July at Heathrow when the number of hours of sunshine is more than 1 standard deviation above the mean.

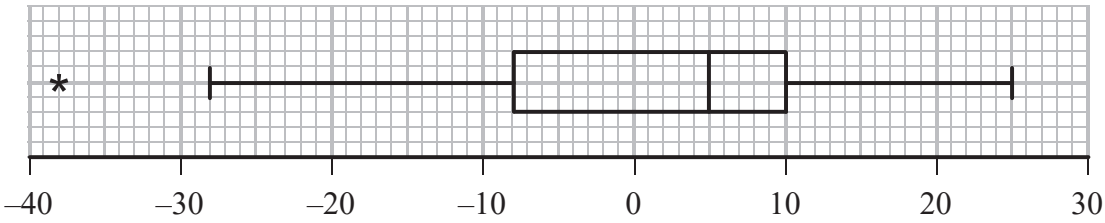
(2)

(f) Use your answers to part (d) and part (e) to comment on the suitability of Helen's model.

(1)

Leave blank

1. At the start of a course, an instructor asked a group of 80 apprentices to estimate the length of a piece of pipe. The error (true length – estimated length) was recorded in centimetres. The results are summarised in the box plot below.



(a) Find the range for these data.

(1)

(b) Find the interquartile range for these data.

(1)

One month later, the instructor asked the 80 apprentices to estimate the length of a different piece of pipe and recorded their errors. The results are summarised in the table below.

Error (e cm)	Number of apprentices
$-40 < e \leqslant -16$	2
$-16 < e \leqslant -8$	18
$-8 < e \leqslant 0$	33
$0 < e \leqslant 8$	14
$8 < e \leqslant 16$	10
$16 < e \leqslant 40$	3

(c) Use linear interpolation to estimate the median error for these data.

(2)

(d) Show that the upper quartile for these data, to the nearest centimetre, is 4.

(1)

For these data, the lower quartile is –8 and the five worst errors were –25, –21, 18, 23, 28

An outlier is a value that falls either
more than 1.5 × (interquartile range) above the upper quartile or
more than 1.5 × (interquartile range) below the lower quartile.

(e) (i) Show that there are only 2 outliers for these data.

(ii) Draw a box plot for these data on the grid on page 3.

(6)

(f) State, giving reasons, whether or not the apprentices' ability to estimate the length of a piece of pipe has improved over the first month of the course.

(3)

Leave blank

1. Nina weighed a random sample of 50 carrots from her shop and recorded the weight, in grams to the nearest gram, for each carrot. The results are summarised below.

Weight of carrot	Frequency (f)	Weight midpoint (x grams)
45–54	5	49.5
55–59	10	57
60–64	22	62
65–74	13	69.5

(You may use $\sum fx^2 = 192\,102.5$)

(a) Use linear interpolation to estimate the median weight of these carrots.

(2)

(b) Find an estimate for the mean weight of these carrots.

(2)

(c) Find an estimate for the standard deviation of the weights of these carrots.

(2)

A carrot is selected at random from Nina's shop.

(d) Estimate the probability that the weight of this carrot is more than 70 grams.

(2)

Leave blank

2. The time taken to complete a puzzle, in minutes, is recorded for each person in a club. The times are summarised in a grouped frequency distribution and represented by a histogram.

One of the class intervals has a frequency of 20 and is shown by a bar of width 1.5 cm and height 12 cm on the histogram. The total area under the histogram is 94.5 cm^2

Find the number of people in the club.

(3)

Leave blank

8.

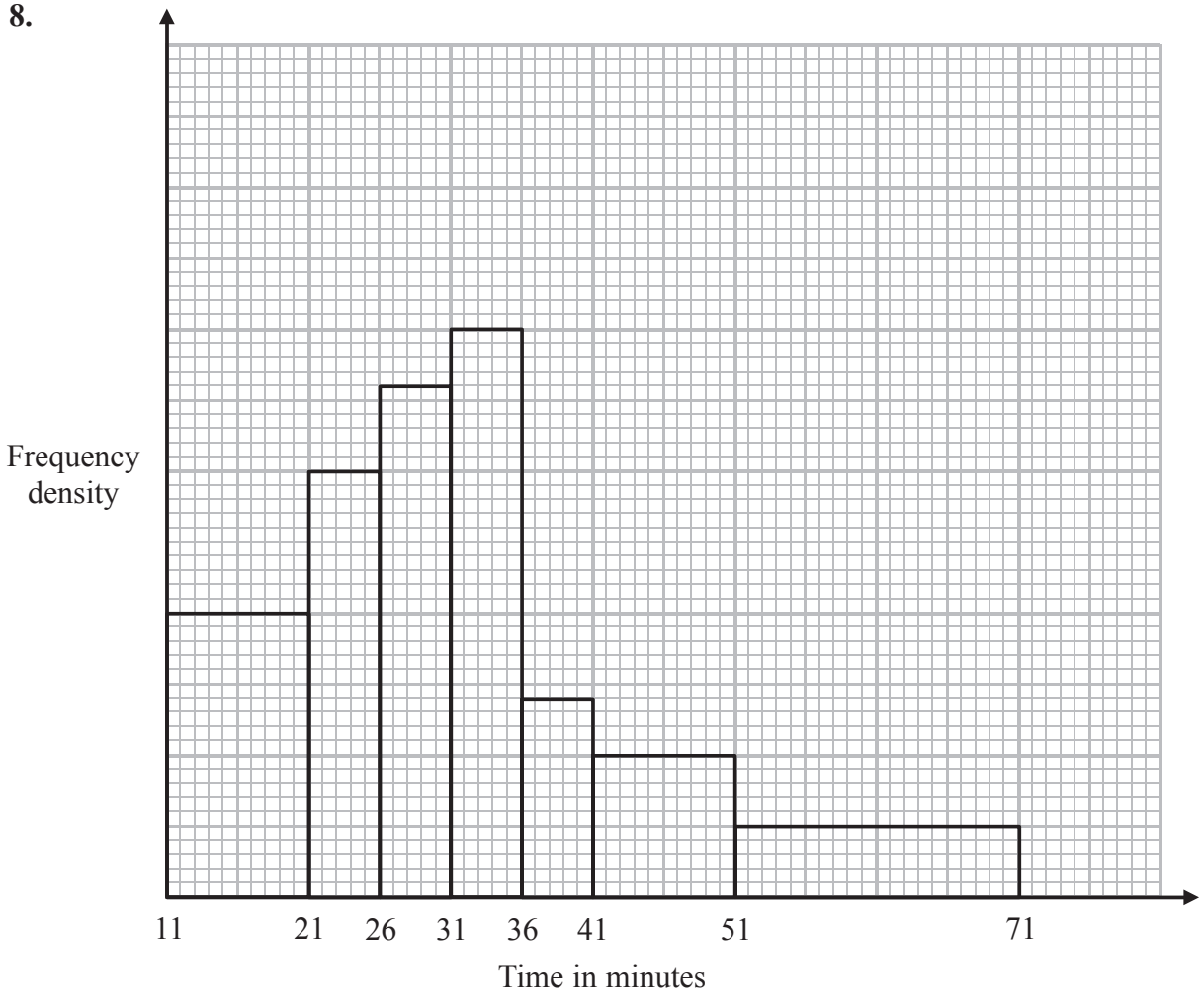


Figure 1

The histogram in Figure 1 summarises the times, in minutes, that 200 people spent shopping in a supermarket.

(a) Give a reason to justify the use of a histogram to represent these data.

(1)

Given that 40 people spent between 11 and 21 minutes shopping in the supermarket, estimate

(b) the number of people that spent between 18 and 25 minutes shopping in the supermarket,

(3)

(c) the median time spent shopping in the supermarket by these 200 people.

(2)

The mid-point of each bar is represented by x and the corresponding frequency by f.

(d) Show that $\sum fx = 6390$

(2)

Leave blank

2. The table below shows the distances (to the nearest km) travelled to work by the 50 employees in an office.

Distance (km)	Frequency (f)	Distance midpoint (x)
0 – 2	16	1.25
3 – 5	12	4
6 – 10	10	8
11 – 20	8	15.5
21 – 40	4	30.5

[You may use $\sum fx = 394$, $\sum fx^2 = 6500$]

A histogram has been drawn to represent these data.

The bar representing the distance of 3 – 5 has a width of 1.5 cm and a height of 6 cm.

(a) Calculate the width and height of the bar representing the distance of 6 – 10

(3)

(b) Use linear interpolation to estimate the median distance travelled to work.

(2)

(c) (i) Show that an estimate of the mean distance travelled to work is 7.88 km.

(ii) Estimate the standard deviation of the distances travelled to work.

(4)

(d) Describe, giving a reason, the skewness of these data.

(2)

Peng starts to work in this office as the 51st employee.

She travels a distance of 7.88 km to work.

(e) Without carrying out any further calculations, state, giving a reason, what effect Peng's addition to the workforce would have on your estimates of the

(i) mean,

(ii) median,

(iii) standard deviation

of the distances travelled to work.

(3)

Leave blank

1. Two classes of students, class A and class B, sat a test.

Class A has 10 students. Class B has 15 students.

Each student achieved a score, x, on the test and their scores are summarised in the table below.

	n	$\sum x$	$\sum x^2$
Class A	10	770	59610
Class B	15	t	58035

The mean score for Class A is 77 and the mean score for Class B is 61

(a) Find the value of t

(1)

(b) Calculate the variance of the test scores for each class.

(3)

The highest score on the test was 95 and the lowest score was 45

These were each scored by students from the same class.

(c) State, with a reason, which class you believe they were from.

(1)

The two classes are combined into one group of 25 students.

(d) (i) Find the mean test score for all 25 students.

(ii) Find the variance of the test scores for all 25 students.

(4)

The teacher of class A later realises that he added up the test scores for his class incorrectly. Each student's test score in class A should be increased by 3

(e) Without further calculations, state, with a reason, the effect this will have on

(i) the variance of the test scores for class A

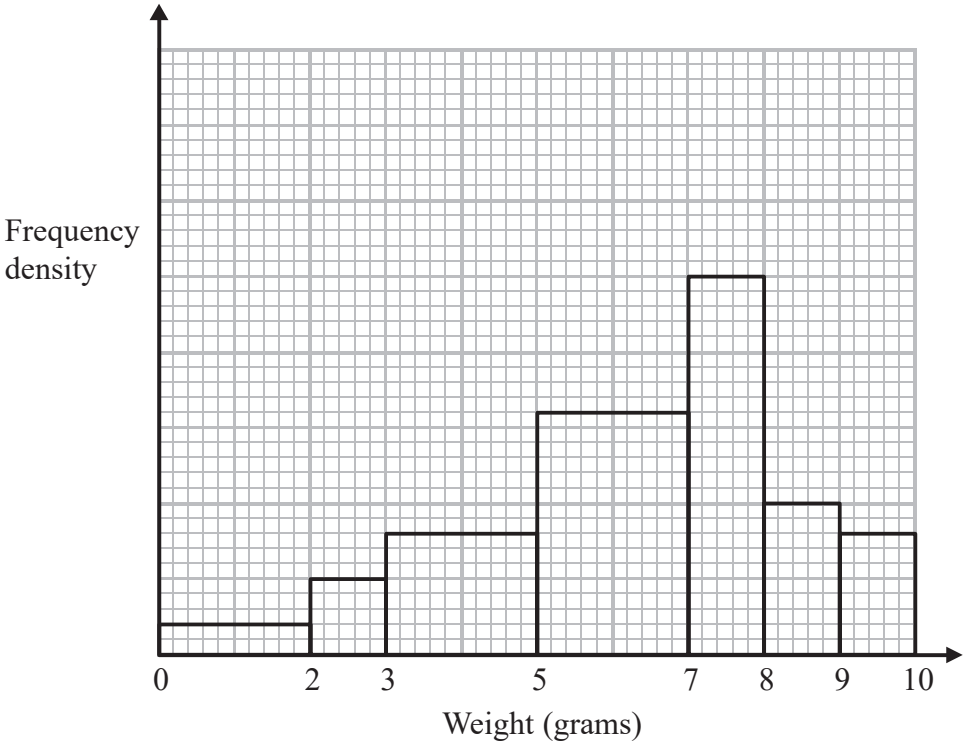
(ii) the mean test score for all 25 students

(iii) the variance of the test scores for all 25 students.

(3)

Leave blank

1. Ralph records the weights, in grams, of 100 tomatoes. This information is displayed in the histogram below.



Given that 5 of the tomatoes have a weight between 2 and 3 grams,

(a) find the number of tomatoes with a weight between 0 and 2 grams.

(2)

One of the tomatoes is selected at random.

(b) Find the probability that it weighs more than 3 grams.

(2)

(c) Estimate the proportion of the tomatoes with a weight greater than 6.25 grams.

(2)

(d) Using your answer to part (c), explain whether or not the median is greater than 6.25 grams.

(1)

Given that the mean weight of these tomatoes is 6.25 grams and using your answer to part (d),

(e) describe the skewness of the distribution of the weights of these tomatoes. Give a reason for your answer.

(1)

Two of these 100 tomatoes are selected at random.

(f) Estimate the probability that both tomatoes weigh within 0.75 grams of the mean.

(4)

Leave blank

2.

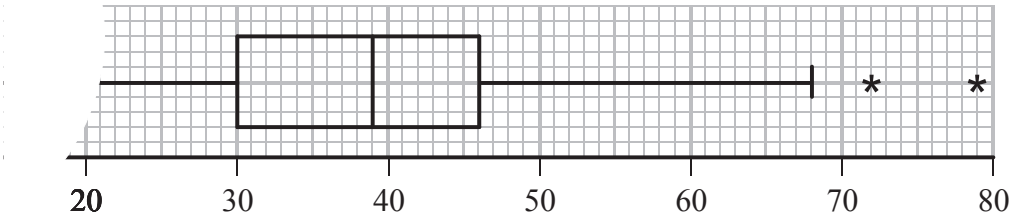


Figure 1

Figure 1 shows part of a box and whisker plot for the marks in an examination with a large number of candidates. Part of the lower whisker has been torn off.

(a) Given that 75% of the candidates passed the examination, state the lowest mark for the award of a pass.

(1)

(b) Given that the top 25% of the candidates achieved a merit grade, state the lowest mark for the award of a merit grade.

(1)

An outlier is defined as any value greater than c or any value less than d where

$$c = Q_3 + 1.5(Q_3 - Q_1)$$
$$d = Q_1 - 1.5(Q_3 - Q_1)$$

(c) Find the value of c and the value of d.

(2)

(d) Write down the 3 highest marks scored in the examination.

(2)

The 3 lowest marks in the examination were 5, 10 and 15

(e) On the diagram on page 7, complete the box and whisker plot.

(3)

Three candidates are selected at random from those who took this examination.

(f) Find the probability that all 3 of these candidates passed the examination but only 2 achieved a merit grade.

(3)

Leave blank

8. A manager records the number of hours of overtime claimed by 40 staff in a month.

The histogram in Figure 1 represents the results.

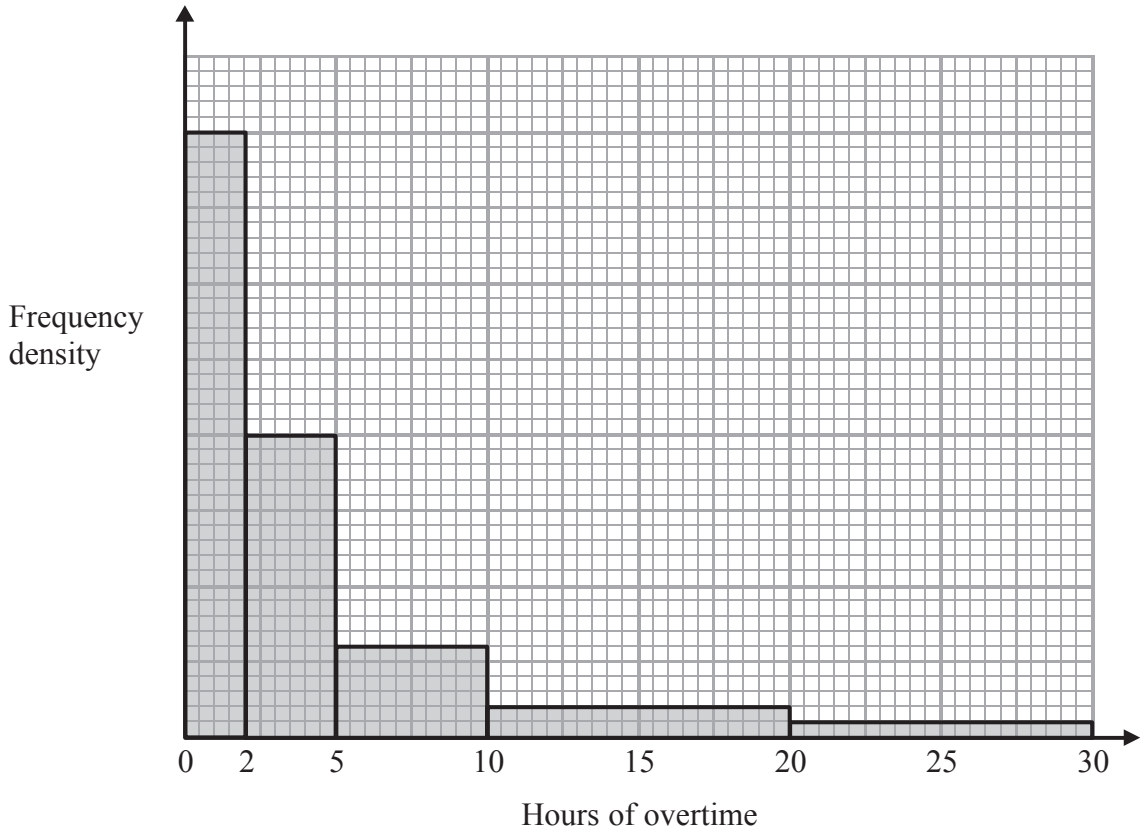


Figure 1

(a) Calculate the number of staff who have claimed less than 10 hours of overtime in the month.

(4)

(b) Estimate the median number of hours of overtime claimed by these 40 staff in the month.

(2)

(c) Estimate the mean number of hours of overtime claimed by these 40 staff in the month.

(2)

The manager wants to compare these data with overtime data he collected earlier to find out if the overtime claimed by staff has decreased.

(d) State, giving a reason, whether the manager should use the median or the mean to compare the overtime claimed by staff.

(2)