Edexcel past paper questions

Statistics 1

Chapters 2-4 (Continuous)

Grouped frequency tables, continuous and discrete data

To find medians and quartiles

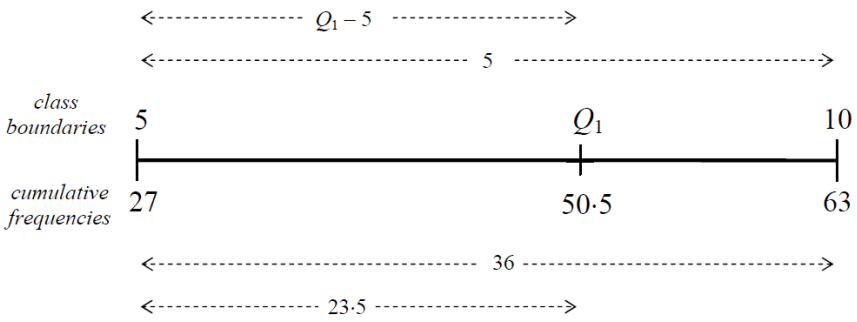
1. Find $k = \frac{n}{2}$ (for Q_2), $\frac{n}{4}$ (for Q_1), $\frac{3n}{4}$ (for Q_3).
2. **Do not round *k* up or change it in any way.**
3. Use linear interpolation to find median and quartiles – **note** that you must use the correct intervals for discrete data (start at the $\frac{1}{2}$ s).

Grouped frequency tables, continuous data

class boundaries	frequency	cumulative frequency
$0 \le x < 5$	27	27
5 to 10	36	63
10 to 20	54	117
20 to 30	49	166
30 to 60	24	190
60 to 100	12	202

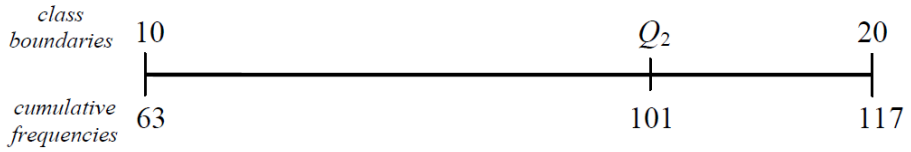
With *continuous* data, the end of one interval is the same as the start of the next – no gaps.

To find Q_1, $n = 202$ $\Rightarrow$ $\frac{n}{4} = 50\frac{1}{2}$ **do not change it**



From the diagram $\frac{Q_1 - 5}{5} = \frac{23{\cdot}5}{36} \Rightarrow Q_1 = 5 + 5 \times \frac{23{\cdot}5}{36} = 8{\cdot}263888889 = 8{\cdot}26$ to 3 S.F.

To find Q_2, $n = 202$ $\Rightarrow$ $\frac{n}{2} = 101$ **do not change it**



From the diagram $\frac{Q_2-10}{20-10} = \frac{101-63}{117-63} \Rightarrow Q_2 = 10 + 10 \times \frac{38}{54} = 17{\cdot}037\ldots = 17{\cdot}0$ to 3 s

Similarly for Q_3, $\frac{3n}{4} = 151{\cdot}5$, so Q_3 lies in the interval (20, 30)

$\Rightarrow$ $\frac{Q_3-20}{30-20} = \frac{151.5-117}{166-117} \Rightarrow Q_3 = 20 + 10 \times \frac{34.5}{49} = 27{\cdot}0408\ldots = 27{\cdot}0$ to 3 S.F.

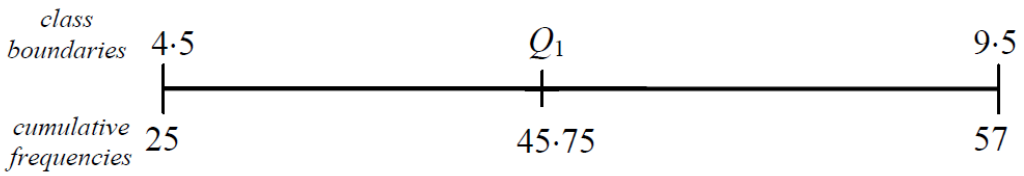
Grouped frequency tables, discrete data

The *discrete* data in grouped frequency tables is treated as *continuous*.

1. Change the class boundaries to the $4\frac{1}{2}$, $9\frac{1}{2}$ etc.
2. Proceed as for grouped frequency tables for continuous data.

class interval	class boundaries	frequency	cumulative frequency
0 – 4	0 to 4 ½	25	25
5 – 9	4 ½ to 9 ½	32	57
10 – 19	9 ½ to 19 ½	51	108
20 – 29	19 ½ to 29 ½	47	155
30 – 59	29 ½ to 59 ½	20	175
60 –99	59 ½ to 99 ½	8	183

To find Q_1, $n = 183$ $\Rightarrow$ $\frac{n}{4} = 45{\cdot}75$



From the diagram $\frac{Q_1-4{\cdot}5}{9{\cdot}5-4{\cdot}5} = \frac{45{\cdot}75-25}{57-25}$

$$\Rightarrow \quad Q_1 = 4{\cdot}5 + 5 \times \frac{20{\cdot}75}{32} = 7{\cdot}7421875\ldots = 7{\cdot}74 \text{ to 3 S.F.}$$

Q_2 and Q_3 can be found in a similar way.

Percentiles

Percentiles are calculated in exactly the same way as quartiles.

Example: For the 90th percentile, find $\frac{90n}{100}$ and proceed as above.

Histograms

Histograms are used for representing data that is **continuous** and are summarized in a grouped frequency distribution.

- There are no gaps between the bars.
- The area of the bar is proportional to the frequency.

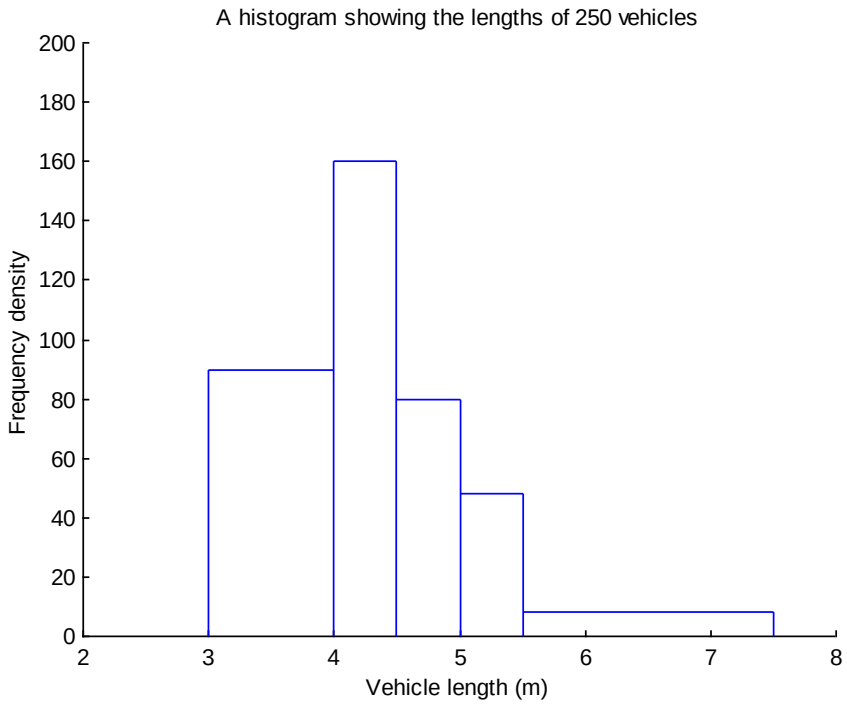
When you are asked to draw a histogram in a S1 examination, it is **essential** that you work out and plot the **FREQUENCY DENSITIES** on the y-axis, where

Frequency density = Frequency ÷ class width.

Example:

The lengths (in metres) of 250 vehicles aboard a cross-channel ferry are summarised in the following table:

Vehicle length (m)	*Class width*	*Frequency*	*Frequency density = Frequency ÷ class width*
3.0-4.0	1	90	90
4.0-4.5	0.5	80	160
4.5-5.0	0.5	40	80
5.0-5.5	0.5	24	48
5.5-7.5	2	16	8



N.B. It is important for you to label each axis and to give your graph a title.

Sometimes you have to think carefully about the width of each interval. You have to do this if the upper endpoint of one interval does not appear to match the lower endpoint of the next interval.

- **Example (rounded data)**

A class of 30 Year 5 children took part in a running race. The teacher recorded how long each child took to complete the race to the nearest second. Their times are shown in the table.

Time interval (seconds)	Frequency
40 – 49	5
50 – 54	8
55 – 59	6
60 – 69	7
70 –	4

The intervals in this table do not appear to meet because the data has been recorded to the nearest second. The first interval actually includes all times from 39.5 seconds up to (but not including) 49.5 seconds; the second interval all times from 49.5 up to 54.5 etc.

Also, the last interval does not have an upper end point. In such circumstances it is conventional to assume that the last interval has a width that is twice that of the previous interval.

We therefore have this new table:

Time interval (seconds)	Class width	Frequency	Frequency density
39.5 – 49.5	10	5	0.5
49.5 – 54.5	5	8	1.6
54.5 – 59.5	5	6	1.2
59.5 – 69.5	10	7	0.7
69.5 – 89.5	20	4	0.2

A histogram can then be drawn.

Example:

The height of twenty children (to the nearest cm) was recorded in the following frequency table. Draw a histogram to represent the data.

Height	Frequency f
120-124	1
124-129	5
130-134	7
135-139	4
140-149	3

There are two columns that we need to add: the class width and the frequency density.

Class width is the width of each group. Be careful when calculating to work out from the lower class boundary and the upper class boundary. For example, 120-125 is actually: 124.5-119.5 and so the class width is **5**.

$$\textbf{Frequency density} = \frac{\textbf{frequency}}{\textbf{class width}}$$

Height	Frequency f	Class Width	Frequency Density
120-124	1	5	0.2
125-129	5	5	1
130-134	7	5	1.4
135-139	4	5	0.8
140-149	3	10	0.3

When we have these values, we plot the **lower class** and **upper class boundaries** on the x axis and the **frequency density** on the y axis.

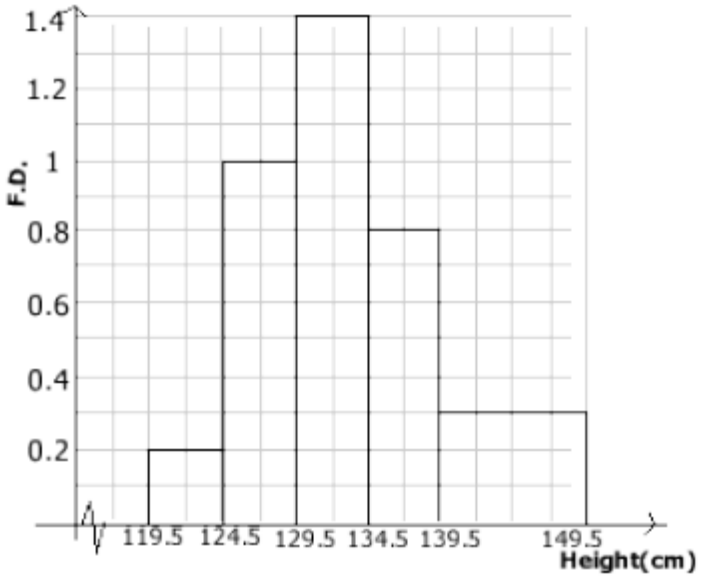


Diagram	When you can use it	Advantages	Disadvantages
Histogram	Continuous data	Allows you to see overall distribution of data – where the peak is, etc.	Actual data values are lost
Stem & Leaf	Discrete data	All the data values are retained. Makes it easy to find median and quartiles. You can still get an overall picture of distribution. You can use it to compare two distributions.	Only practicable foe a small amount of data.
Box plot	Discrete or continuous	Allows you to see the position of the '50% of data easily. Allows you to compare distributions.	Actual data values are lost. It does not give a detailed picture of the shape of the distribution.

1. The following grouped frequency distribution summarises the number of minutes, to the nearest minute, that a random sample of 200 motorists were delayed by roadworks on a stretch of motorway.

Delay (mins)	Number of motorists
4—6	15
7—8	28
9	49
10	53
11—12	30
13—15	15
16—20	10

(*a*) Using graph paper represent these data by a histogram. **(4 marks)**

(*b*) Give a reason to justify the use of a histogram to represent these data. **(1 mark)**

(*c*) Use interpolation to estimate the median of this distribution. **(2 marks)**

(*d*) Calculate an estimate of the mean and an estimate of the standard deviation of these data. **(6 marks)**

One coefficient of skewness is given by

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

(*e*) Evaluate this coefficient for the above data. **(2 marks)**

(*f*) Explain why the normal distribution may not be suitable to model the number of minutes that motorists are delayed by these roadworks. **(2 marks)**

Q5, Jan 2001

2. A meteorologist measured the number of hours of sunshine, to the nearest hour, each day for 100 days. The results are summarised in the table below.

Hours of sunshine	Days
1	16
2–4	32
5–6	28
7	12
8	9
9–11	2
12	1

(*a*) On graph paper, draw a histogram to represent these data.

(5)

(*b*) Calculate an estimate of the number of days that had between 6 and 9 hours of sunshine.

(2)

Q2, Jan 2002

3. The labelling on bags of garden compost indicates that the bags weigh 20 kg. The weights of a random sample of 50 bags are summarised in the table below.

Weight in kg	Frequency
14.6 – 14.8	1
14.8 – 18.0	0
18.0 – 18.5	5
18.5 – 20.0	6
20.0 – 20.2	22
20.2 – 20.4	15
20.4 – 21.0	1

(*a*) On graph paper, draw a histogram of these data. **(4)**

(*b*) Using the coding $y = 10(\text{weight in kg} - 14)$, find an estimate for the mean and standard deviation of the weight of a bag of compost. **(6)**

[Use $\Sigma fy^2 = 171\,503.75$]

(*c*) Using linear interpolation, estimate the median. **(2)**

The company that produces the bags of compost wants to improve the accuracy of the labelling. The company decides to put the average weight in kg on each bag.

(*d*) Write down which of these averages you would recommend the company to use. Give a reason for your answer.

(2)

Q6, May 2002

4. The total amount of time a secretary spent on the telephone in a working day was recorded to the nearest minute. The data collected over 40 days are summarised in the table below.

Time (mins)	90–139	140–149	150–159	160–169	170–179	180–229
No. of days	8	10	10	4	4	4

Draw a histogram to illustrate these data

(4)

Q1, Jan 2003

5. In a particular week, a dentist treats 100 patients. The length of time, to the nearest minute, for each patient's treatment is summarised in the table below.

Time (minutes)	4 – 7	8	9 – 10	11	12 – 16	17 – 20
Number of patients	12	20	18	22	15	13

Draw a histogram to illustrate these data. **(5)**

Q1, June 2003

6. The values of daily sales, to the nearest £, taken at a newsagents last year are summarised in the table below.

Sales	Number of days
1 – 200	166
201 – 400	100
401 – 700	59
701 – 1000	30
1001 – 1500	5

(*a*) Draw a histogram to represent these data. **(5)**
(*b*) Use interpolation to estimate the median and inter-quartile range of daily sales. **(5)**
(*c*) Estimate the mean and the standard deviation of these data. **(6)**
The newsagent wants to compare last year's sales with other years.

(*d*) State whether the newsagent should use the median and the inter-quartile range or the mean and the standard deviation to compare daily sales. Give a reason for your answer. **(2)**

Q5, Jan 2004

7. A college organised a 'fun run'. The times, to the nearest minute, of a random sample of 100 students who took part are summarised in the table below.

Time	Number of students
40–44	10
45–47	15
48	23
49–51	21
52–55	16
56–60	15

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

(1)

(*b*) Write down the upper class boundary and the lower class boundary of the class 40–44.

(1)

(*c*) On graph paper, draw a histogram to represent these data.

(4)

Q7, Nov 2004

8. The following table summarises the distances, to the nearest km, that 134 examiners travelled to attend a meeting in London.

Distance (km)	Number of examiners
41–45	4
46–50	19
51–60	53
61–70	37
71–90	15
91–150	6

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

(1)

(*b*) Calculate the frequency densities needed to draw a histogram for these data.

(DO NOT DRAW THE HISTOGRAM)

(2)

(*c*) Use interpolation to estimate the median Q_2, the lower quartile Q_1, and the upper quartile Q_3 of these data.

The mid-point of each class is represented by x and the corresponding frequency by f. Calculations then give the following values

$$\sum fx = 8379.5 \quad \text{and} \quad \sum fx^2 = 557489.75$$

(*d*) Calculate an estimate of the mean and an estimate of the standard deviation for these data.

(4)

One coefficient of skewness is given by

$$\frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}.$$

(*e*) Evaluate this coefficient and comment on the skewness of these data.

(4)

(*f*) Give another justification of your comment in part (*e*).

(1)
Q2, June 2005

9. Sunita and Shelley talk to each other once a week on the telephone. Over many weeks they recorded, to the nearest minute, the number of minutes spent in conversation on each occasion. The following table summarises their results.

Time (to the nearest minute)	Number of conversations
5–9	2
10–14	9
15–19	20
20–24	13
25–29	8
30–34	3

Two of the conversations were chosen at random.

(*a*) Find the probability that both of them were longer than 24.5 minutes.

(2)

The mid-point of each class was represented by x and its corresponding frequency by f, giving $\sum fx = 1060$.

(*b*) Calculate an estimate of the mean time spent on their conversations.

(2)

During the following 25 weeks they monitored their weekly conversation and found that at the end of the 80 weeks their overall mean length of conversation was 21 minutes.

(*c*) Find the mean time spent in conversation during these 25 weeks.

(4)

(*d*) Comment on these two mean values.

(2)
Q2, May 2006

10. Summarised below are the distances, to the nearest mile, travelled to work by a random sample of 120 commuters.

Distance (to the nearest mile)	Number of commuters
0 – 9	10
10 – 19	19
20 – 29	43
30 – 39	25
40 – 49	8
50 – 59	6
60 – 69	5
70 – 79	3
80 – 89	1

For this distribution,

(*a*) describe its shape,

(1)

(*b*) use linear interpolation to estimate its median.

(2)

The mid-point of each class was represented by x and its corresponding frequency by f giving

$$\Sigma fx = 3550 \text{ and } \Sigma fx^2 = 138020$$

(*c*) Estimate the mean and standard deviation of this distribution.

(3)

One coefficient of skewness is given by

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

(*d*) Evaluate this coefficient for this distribution.

(3)

(*e*) State whether or not the value of your coefficient is consistent with your description in part (*a*). Justify your answer.

(2)

(*f*) State, with a reason, whether you should use the mean or the median to represent the data in this distribution. **(2)**

(*g*) State the circumstance under which it would not matter whether you used the mean or the median to represent a set of data. **(1)**

Q4, Jan 2007

11. A teacher recorded, to the nearest hour, the time spent watching television during a particular week by each child in a random sample. The times were summarised in a grouped frequency table and represented by a histogram.

One of the classes in the grouped frequency distribution was 20–29 and its associated frequency was 9. On the histogram the height of the rectangle representing that class was 3.6 cm and the width was 2 cm.

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

(1)

(*b*) Write down the underlying feature associated with each of the bars in a histogram.

(1)

(*c*) Show that on this histogram each child was represented by 0.8 cm^2.

(3)

The total area under the histogram was 24 cm^2.

(*d*) Find the total number of children in the group.

(2)

Q5, Jan 2007

12.

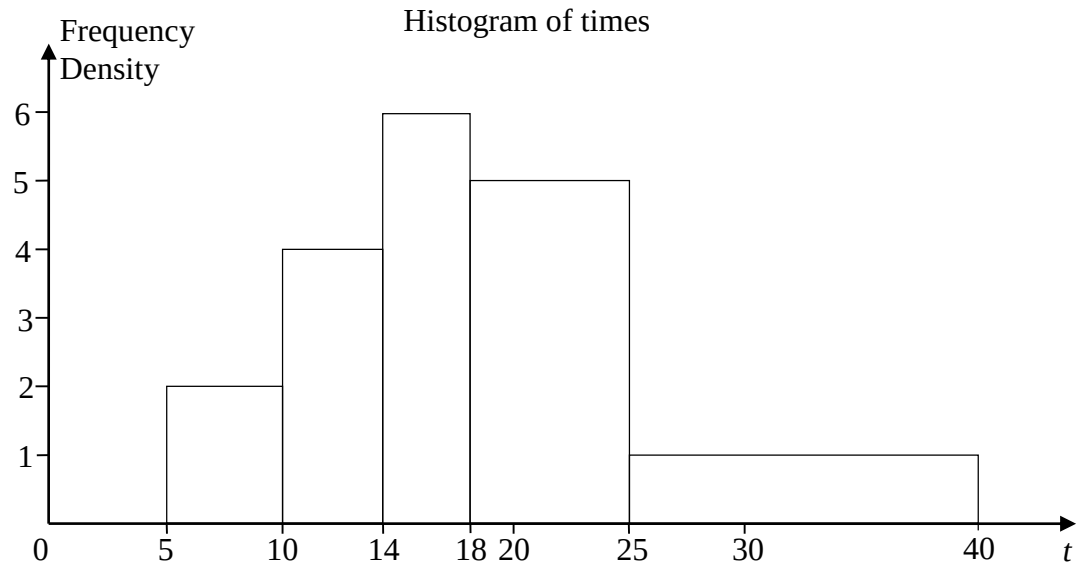


Figure 2

Figure 2 shows a histogram for the variable *t* which represents the time taken, in minutes, by a group of people to swim 500 m.

(*a*) Copy and complete the frequency table for *t*.

t	5 – 10	10 – 14	14 – 18	18 – 25	25 – 40
Frequency	10	16	24		

(2)

(*b*) Estimate the number of people who took longer than 20 minutes to swim 500 m.

(2)

(*c*) Find an estimate of the mean time taken.

(4)

(*d*) Find an estimate for the standard deviation of *t*.

(3)

(*e*) Find the median and quartiles for *t*.

(4)

One measure of skewness is found using $\frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$.

(*f*) Evaluate this measure and describe the skewness of these data.

(2)

Q5, June 2007

13. The histogram in Figure 1 shows the time taken, to the nearest minute, for 140 runners to complete a fun run.

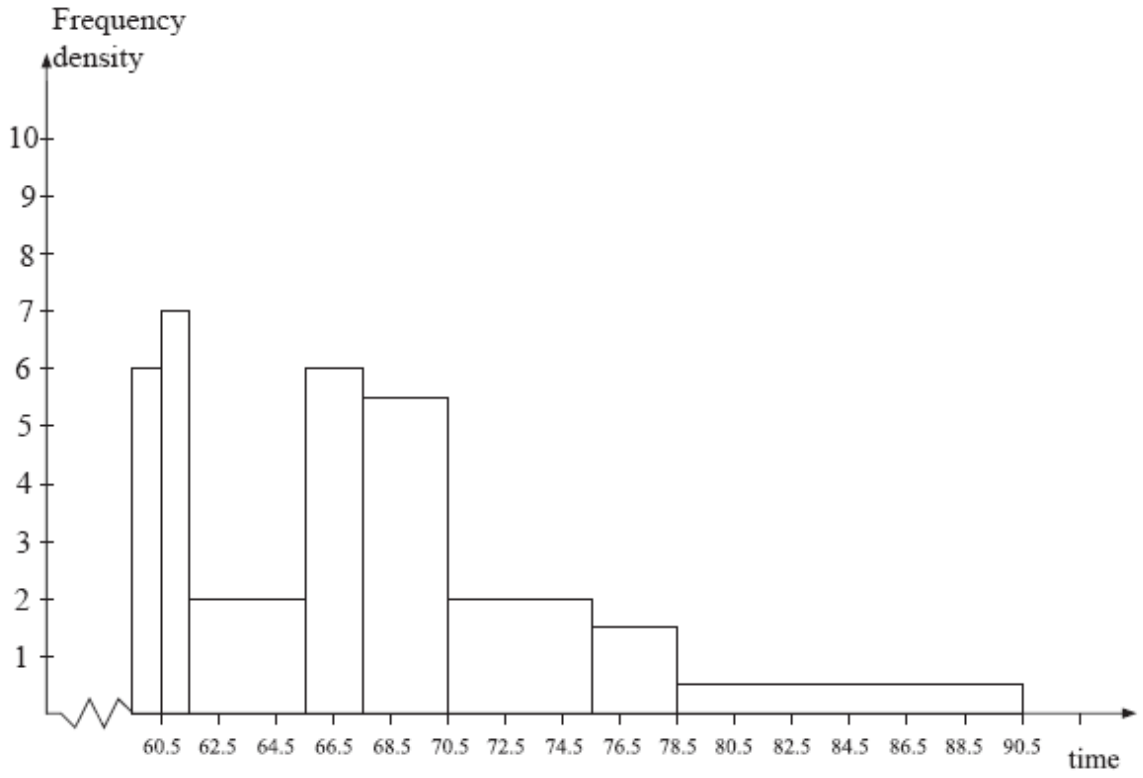


Figure 1

Use the histogram to calculate the number of runners who took between 78.5 and 90.5 minutes to complete the fun run.

(5)

Q3, Jan 2008

14. In a shopping survey a random sample of 104 teenagers were asked how many hours, to the nearest hour, they spent shopping in the last month. The results are summarised in the table below.

Number of hours	Mid-point	Frequency
0 – 5	2.75	20
6 – 7	6.5	16
8 – 10	9	18
11 – 15	13	25
16 – 25	20.5	15
26 – 50	38	10

A histogram was drawn and the group (8 – 10) hours was represented by a rectangle that was 1.5 cm wide and 3 cm high.

(*a*) Calculate the width and height of the rectangle representing the group (16 – 25) hours.

(3)

(*b*) Use linear interpolation to estimate the median and interquartile range.

(5)

(*c*) Estimate the mean and standard deviation of the number of hours spent shopping.

(4)

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

(2)

(*e*) State, giving a reason, which average and measure of dispersion you would recommend to use to summarise these data.

(2)

Q5, Jan 2009

15. The variable x was measured to the nearest whole number. Forty observations are given in the table below.

x	10 – 15	16 – 18	19 –
Frequency	15	9	16

A histogram was drawn and the bar representing the 10 – 15 class has a width of 2 cm and a height of 5 cm. For the 16 – 18 class find

(*a*) the width, **(1)**

(*b*) the height **(2)**

of the bar representing this class.

Q3, May 2009

16. A researcher measured the foot lengths of a random sample of 120 ten-year-old children. The lengths are summarised in the table below.

Foot length, *l*, (cm)	**Number of children**
$10 \leq l < 12$	5
$12 \leq l < 17$	53
$17 \leq l < 19$	29
$19 \leq l < 21$	15
$21 \leq l < 23$	11
$23 \leq l < 25$	7

(*a*) Use interpolation to estimate the median of this distribution.

(2)

(*b*) Calculate estimates for the mean and the standard deviation of these data.

(6)

One measure of skewness is given by

$$\text{Coefficient of skewness} = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$$

(*c*) Evaluate this coefficient and comment on the skewness of these data.

(3)

Greg suggests that a normal distribution is a suitable model for the foot lengths of ten-year-old children.

(*d*) Using the value found in part (*c*), comment on Greg's suggestion, giving a reason for your answer.

(2)

Q4, May 2009

17. The birth weights, in kg, of 1500 babies are summarised in the table below.

Weight (kg)	Midpoint, x kg	Frequency, f
0.0 – 1.0	0.50	1
1.0 – 2.0	1.50	6
2.0 – 2.5	2.25	60
2.5 – 3.0		280
3.0 – 3.5	3.25	820
3.5 – 4.0	3.75	320
4.0 – 5.0	4.50	10
5.0 – 6.0		3

[You may use $\sum fx = 4841$ and $\sum fx^2 = 15\,889.5$]

(*a*) Write down the missing midpoints in the table above.

(2)

(*b*) Calculate an estimate of the mean birth weight.

(2)

(*c*) Calculate an estimate of the standard deviation of the birth weight.

(3)

(*d*) Use interpolation to estimate the median birth weight.

(2)

(*e*) Describe the skewness of the distribution. Give a reason for your answer.

(2)

Q3, Jan 2010

18. A teacher selects a random sample of 56 students and records, to the nearest hour, the time spent watching television in a particular week.

Hours	1–10	11–20	21–25	26–30	31–40	41–59
Frequency	6	15	11	13	8	3
Mid-point	5.5	15.5		28		50

(*a*) Find the mid-points of the 21−25 hour and 31−40 hour groups.

(2)

A histogram was drawn to represent these data. The 11−20 group was represented by a bar of width 4 cm and height 6 cm.

(*b*) Find the width and height of the 26−30 group.

(3)

(*c*) Estimate the mean and standard deviation of the time spent watching television by these students.

(5)

(*d*) Use linear interpolation to estimate the median length of time spent watching television by these students.

(2)

The teacher estimated the lower quartile and the upper quartile of the time spent watching television to be 15.8 and 29.3 respectively.

(*e*) State, giving a reason, the skewness of these data.

(2)

Q5, May 2010

19. On a randomly chosen day, each of the 32 students in a class recorded the time, t minutes to the nearest minute, they spent on their homework. The data for the class is summarised in the following table.

Time, t	Number of students
10 – 19	2
20 – 29	4
30 – 39	8
40 – 49	11
50 – 69	5
70 – 79	2

(*a*) Use interpolation to estimate the value of the median.

(2)

Given that

$$\sum t = 1414 \qquad \text{and} \qquad \sum t^2 = 69\,378,$$

(*b*) find the mean and the standard deviation of the times spent by the students on their homework.

(3)

(*c*) Comment on the skewness of the distribution of the times spent by the students on their homework. Give a reason for your answer.

(2)

Q5, Jan 2011

20. A class of students had a sudoku competition. The time taken for each student to complete the sudoku was recorded to the nearest minute and the results are summarised in the table below.

Time	Mid-point, x	Frequency, f
2 – 8	5	2
9 – 12		7
13 – 15	14	5
16 – 18	17	8
19 – 22	20.5	4
23 – 30	26.5	4

(You may use $\sum fx^2 = 8603.75$)

(*a*) Write down the mid-point for the 9 – 12 interval.

(1)

(*b*) Use linear interpolation to estimate the median time taken by the students.

(2)

(*c*) Estimate the mean and standard deviation of the times taken by the students.

(5)

The teacher suggested that a normal distribution could be used to model the times taken by the students to complete the sudoku.

(*d*) Give a reason to support the use of a normal distribution in this case.

(1)

On another occasion the teacher calculated the quartiles for the times taken by the students to complete a different sudoku and found

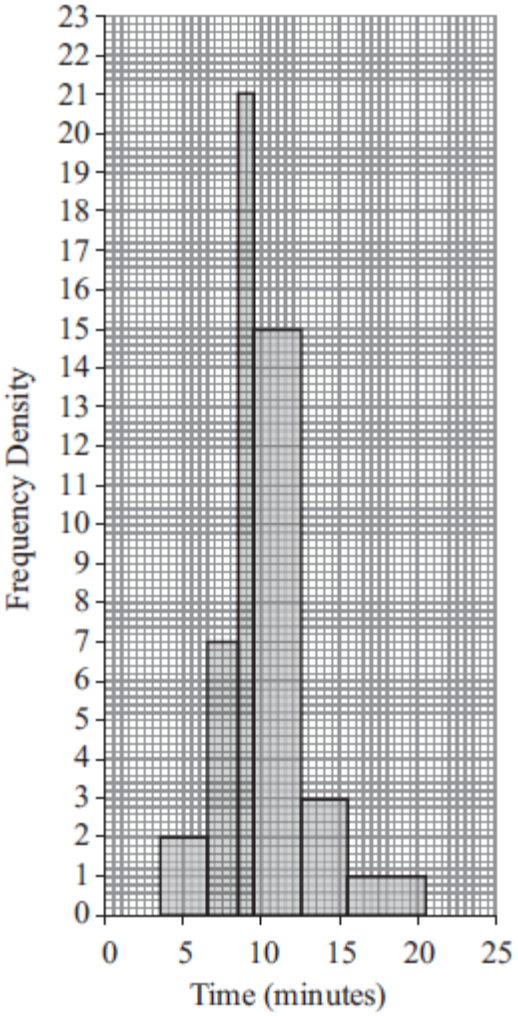
$Q_1 = 8.5$ $Q_2 = 13.0$ $Q_3 = 21.0$

(*e*) Describe, giving a reason, the skewness of the times on this occasion.

(2)

Q5, May 2011

21. The histogram in Figure 1 shows the time, to the nearest minute, that a random sample of 100 motorists were delayed by roadworks on a stretch of motorway.



(a) Complete the table.

Delay (minutes)	Number of motorists
4 – 6	6
7 – 8	
9	21
10 – 12	45
13 – 15	9
16 – 20	

(2)

(*b*) Estimate the number of motorists who were delayed between 8.5 and 13.5 minutes by the roadworks.

(2)

Q1, Jan 2012

22.

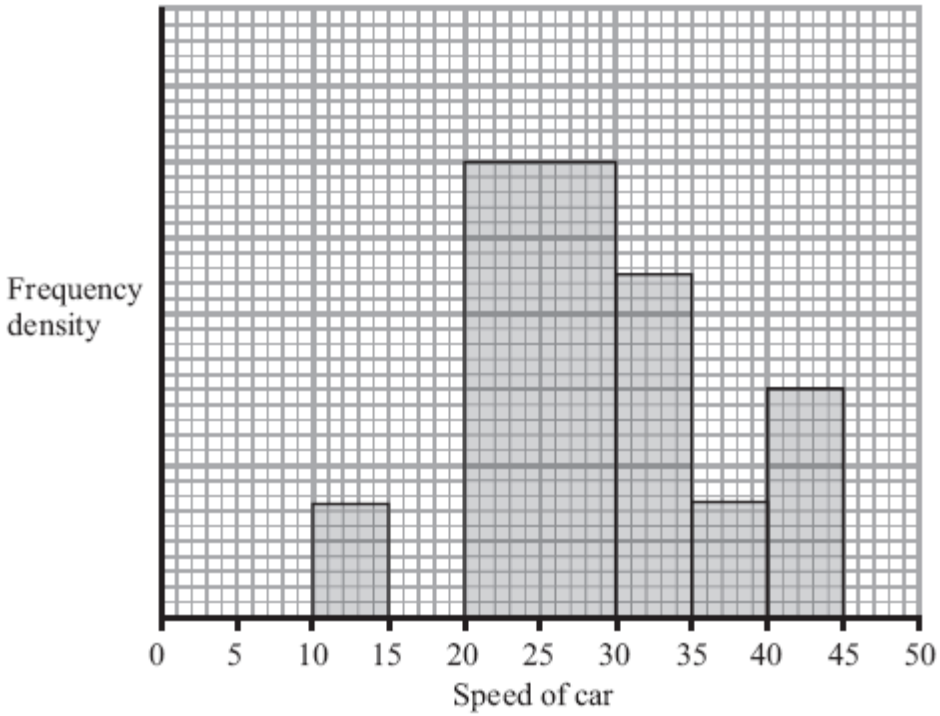


Figure 2

A policeman records the speed of the traffic on a busy road with a 30 mph speed limit.

He records the speeds of a sample of 450 cars. The histogram in Figure 2 represents the results.

(*a*) Calculate the number of cars that were exceeding the speed limit by at least 5 mph in the sample.

(4)

(*b*) Estimate the value of the mean speed of the cars in the sample.

(3)

(*c*) Estimate, to 1 decimal place, the value of the median speed of the cars in the sample.

(2)

(*d*) Comment on the shape of the distribution. Give a reason for your answer.

(2)

(*e*) State, with a reason, whether the estimate of the mean or the median is a better representation of the average speed of the traffic on the road.

(2)

Q5, May 2012

23. A survey of 100 households gave the following results for weekly income £y.

Income y (£)	Mid-point	Frequency f
$0 \leq y < 200$	100	12
$200 \leq y < 240$	220	28
$240 \leq y < 320$	280	22
$320 \leq y < 400$	360	18
$400 \leq y < 600$	500	12
$600 \leq y < 800$	700	8

(You may use $\sum fy^2 = 12\,452\,800$)

A histogram was drawn and the class $200 \leq y < 240$ was represented by a rectangle of width 2 cm and height 7 cm.

(a) Calculate the width and the height of the rectangle representing the class $320 \leq y < 400$.

(3)

(b) Use linear interpolation to estimate the median weekly income to the nearest pound.

(2)

(c) Estimate the mean and the standard deviation of the weekly income for these data.

(4)

One measure of skewness is $\dfrac{3(\text{mean} - \text{median})}{\text{standard deviation}}$.

(d) Use this measure to calculate the skewness for these data and describe its value.

(2)

Katie suggests using the random variable X which has a normal distribution with mean 320 and standard deviation 150 to model the weekly income for these data.

(e) Find $P(240 < X < 400)$.

(2)

(f) With reference to your calculations in parts (d) and (e) and the data in the table, comment on Katie's suggestion.

(2)

Q5, Jan 2013

2**4.** The following table summarises the times, t minutes to the nearest minute, recorded for a group of students to complete an exam.

Time (minutes) t	11 – 20	21 – 25	26 – 30	31 – 35	36 – 45	46 – 60
Number of students f	62	88	16	13	11	10

[You may use $\sum ft^2 = 134281.25$]

(*a*) Estimate the mean and standard deviation of these data.

(5)

(*b*) Use linear interpolation to estimate the value of the median.

(2)

(*c*) Show that the estimated value of the lower quartile is 18.6 to 3 significant figures.

(1)

(*d*) Estimate the interquartile range of this distribution.

(2)

(*e*) Give a reason why the mean and standard deviation are not the most appropriate summary statistics to use with these data.

(1)

The person timing the exam made an error and each student actually took 5 minutes less than the times recorded above. The table below summarises the actual times.

Time (minutes) t	6 – 15	16 – 20	21 – 25	26 – 30	31 – 40	41 – 55
Number of students f	62	88	16	13	11	10

(*f*) Without further calculations, explain the effect this would have on each of the estimates found in parts (*a*), (*b*), (*c*) and (*d*).

(3)

Q4, May 2013

25. An agriculturalist is studying the yields, y kg, from tomato plants. The data from a random sample of 70 tomato plants are summarised below.

Yield (y kg)	**Frequency** (f)	**Yield midpoint** (x kg)
$0 \leq y < 5$	16	2.5
$5 \leq y < 10$	24	7.5
$10 \leq y < 15$	14	12.5
$15 \leq y < 25$	12	20
$25 \leq y < 35$	4	30

(You may use $\sum fx = 755$ and $\sum fx^2 = 12\,037.5$)

A histogram has been drawn to represent these data.

The bar representing the yield $5 \leq y < 10$ has a width of 1.5 cm and a height of 8 cm.

(*a*) Calculate the width and the height of the bar representing the yield $15 \leq y < 25$.

(3)

(*b*) Use linear interpolation to estimate the median yield of the tomato plants.

(2)

(*c*) Estimate the mean and the standard deviation of the yields of the tomato plants.

(4)

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

(2)

(*e*) Estimate the number of tomato plants in the sample that have a yield of more than 1 standard deviation above the mean.

(2)

Q3, May 2013_R

26. The times, in seconds, spent in a queue at a supermarket by 85 randomly selected customers, are summarised in the table below.

Time (seconds)	Number of customers, f
0 – 30	2
30 – 60	10
60 – 70	17
70 – 80	25
80 – 100	25
100 – 150	6

A histogram was drawn to represent these data. The 30 – 60 group was represented by a bar of width 1.5 cm and height 1 cm.

(*a*) Find the width and the height of the 70 – 80 group.

(3)

(*b*) Use linear interpolation to estimate the median of this distribution.

(2)

Given that x denotes the midpoint of each group in the table and

$$\Sigma fx = 6460 \qquad \Sigma fx^2 = 529\,400$$

(*c*) calculate an estimate for

(i) the mean,

(ii) the standard deviation,

for the above data.

(3)

One measure of skewness is given by

$$\text{coefficient of skewness} = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$$

(*d*) Evaluate this coefficient and comment on the skewness of these data.

(3)

Q6, May 2014

27. The table shows the time, to the nearest minute, spent waiting for a taxi by each of 80 people one Sunday afternoon.

Waiting time (in minutes)	Frequency
2–4	15
5–6	9
7	6
8	24
9–10	14
11–15	12

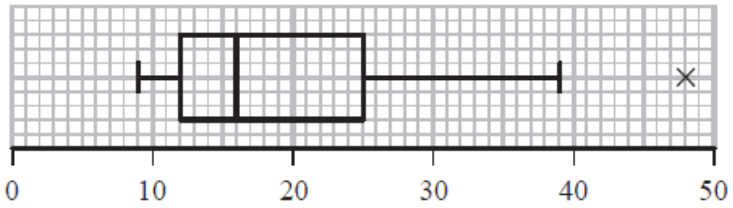
(*a*) Write down the upper class boundary for the 2–4 minute interval.**(1)**

A histogram is drawn to represent these data. The height of the tallest bar is 6 cm.

(*b*) Calculate the height of the second tallest bar.**(3)**
(*c*) Estimate the number of people with a waiting time between 3.5 minutes and 7 minutes.**(2)**
(*d*) Use linear interpolation to estimate the median, the lower quartile and the upper quartile of the waiting times.**(4)**
(*e*) Describe the skewness of these data, giving a reason for your answer.**(2)**

Q5, May ,2014_R

1. Each of 60 students was asked to draw a 20° angle without using a protractor. The size of each angle drawn was measured. 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)

The students were then asked to draw a 70° angle.
The results are summarised in the table below.

Angle, a, (degrees)	Number of students
$55 \leq a < 60$	6
$60 \leq a < 65$	15
$65 \leq a < 70$	13
$70 \leq a < 75$	11
$75 \leq a < 80$	8
$80 \leq a < 85$	7

(*c*) Use linear interpolation to estimate the size of the median angle drawn. Give your answer to 1 decimal place.

(2)

(*d*) Show that the lower quartile is 63°.

(2)

For these data, the upper quartile is 75°, the minimum is 55° and the maximum is 84°.

An outlier is an observation 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 no outliers for these data.

(ii) On graph paper, draw a box plot for these data.

(5)

(*f*) State which angle the students were more accurate at drawing. Give reasons for your answer.

(3)

Q1, June 2015