Edexcel GCE

Core Mathematics S1

Representing Data

Materials required for examination
Mathematical Formulae (Green)

Items included with question papers
Nil

Advice to Candidates

You must ensure that your answers to parts of questions are clearly labelled.
You must show sufficient working to make your methods clear to the Examiner. Answers without working may gain no credit.

1. (*a*) Give two reasons to justify the use of statistical models.

(2)

It has been suggested that there are 7 stages involved in creating a statistical model. They are summarised below, with stages 3, 4 and 7 missing.

Stage 1. The recognition of a real-world problem.

Stage 2. A statistical model is devised.

Stage 3.

Stage 4.

Stage 5. Comparisons are made against the devised model.

Stage 6. Statistical concepts are used to test how well the model describes the real-world problem.

Stage 7.

(*b*) Write down the missing stages.

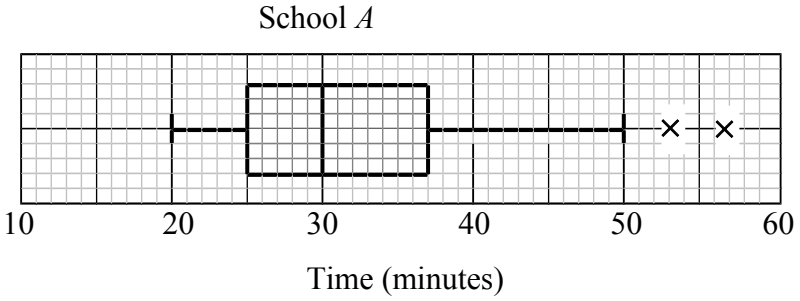
(3)

2. (*a*) Describe the main features and uses of a box plot.

(3)

Children from schools *A* and *B* took part in a fun run for charity. The times, to the nearest minute, taken by the children from school *A* are summarised in Figure 1.

Figure 1



(*b*) (i) Write down the time by which 75% of the children in school *A* had completed the run.

(ii) State the name given to this value.

(2)

(*c*) Explain what you understand by the two crosses (×) on Figure 1.

(2)

For school *B* the least time taken by any of the children was 25 minutes and the longest time was 55 minutes. The three quartiles were 30, 37 and 50 respectively.

(*d*) On graph paper, draw a box plot to represent the data from school *B*.

(4)

(*e*) Compare and contrast these two box plots.

(4)

3. 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)

4. 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)***

5. 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)

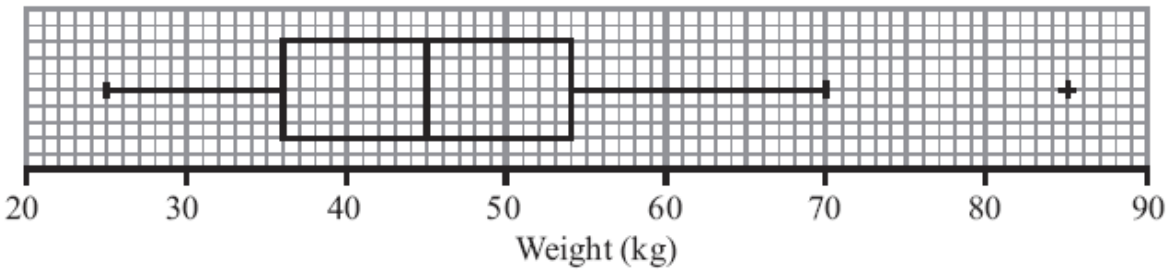
The total area under the histogram was 24 cm^2.

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

(2)

6. The box plot in Figure 1 shows a summary of the weights of the luggage, in kg, for each musician in an orchestra on an overseas tour.

Figure 1



The airline's recommended weight limit for each musician's luggage was 45 kg.

Given that none of the musician's luggage weighed exactly 45 kg,

(*a*) state the proportion of the musicians whose luggage was below the recommended weight limit.

(1)

A quarter of the musicians had to pay a charge for taking heavy luggage.

(*b*) State the smallest weight for which the charge was made.

(1)

(*c*) Explain what you understand by the + on the box plot in Figure 1, and suggest an instrument that the owner of this luggage might play.

(2)

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

(2)

One musician of the orchestra suggests that the weights of the luggage, in kg, can be modelled by a normal distribution with quartiles as given in Figure 1.

(*c*) Find the standard deviation of this normal distribution.

(4)

7.

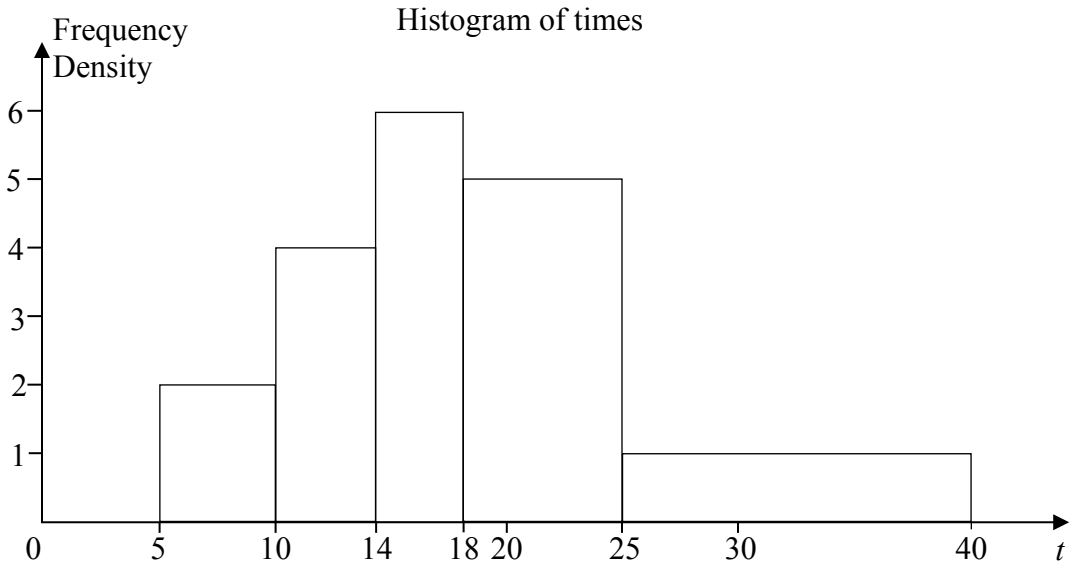


Figure 2

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)

8. Cotinine is a chemical that is made by the body from nicotine which is found in cigarette smoke. A doctor tested the blood of 12 patients, who claimed to smoke a packet of cigarettes a day, for cotinine. The results, in appropriate units, are shown below.

Patient	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*
Cotinine level, x	160	390	169	175	125	420	171	250	210	258	186	243

[You may use $\sum x^2 = 724\,961$]

(*a*) Find the mean and standard deviation of the level of cotinine in a patient's blood.

(4)

(*b*) Find the median, upper and lower quartiles of these data.

(3)

A doctor suspects that some of his patients have been smoking more than a packet of cigarettes per day. He decides to use $Q_3 + 1.5(Q_3 - Q_1)$ to determine if any of the cotinine results are far enough away from the upper quartile to be outliers.

(*c*) Identify which patient(s) may have been smoking more than a packet of cigarettes a day. Show your working clearly.

(4)

Research suggests that cotinine levels in the blood form a skewed distribution.

One measure of skewness is found using $\dfrac{(Q_1 - 2Q_2 + Q_3)}{(Q_3 - Q_1)}$.

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

(3)

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

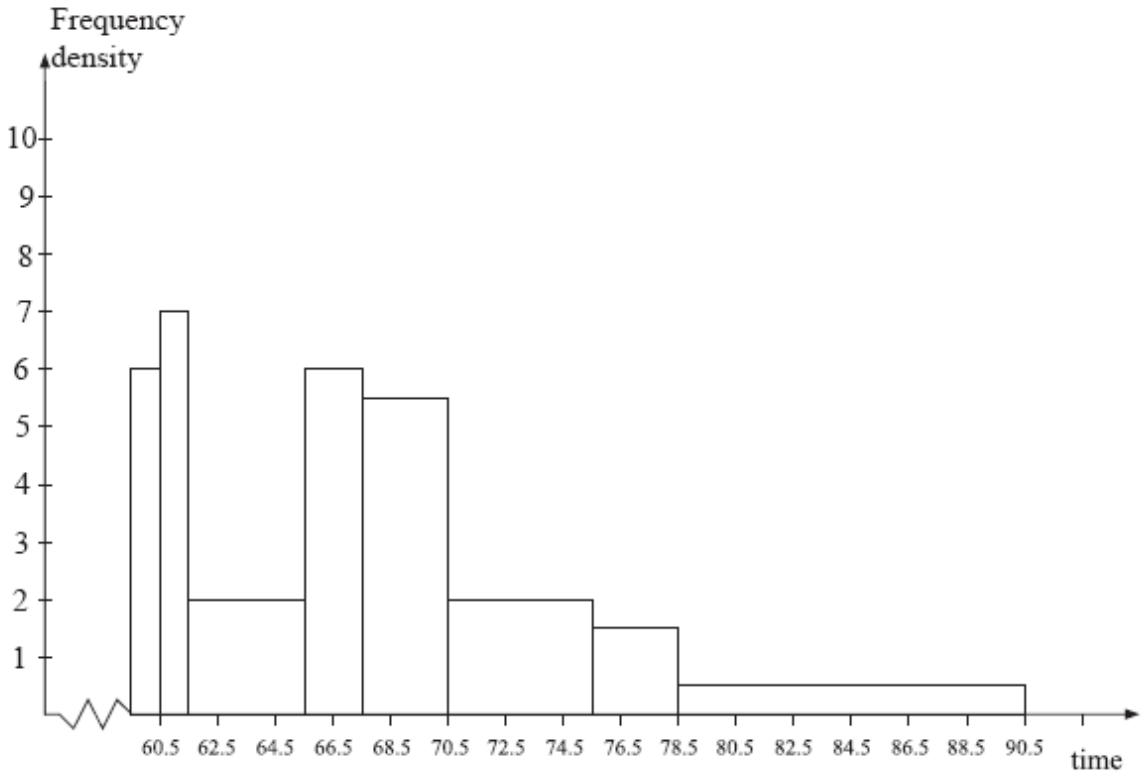


Figure 1

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

(5)

10. The age in years of the residents of two hotels are shown in the back to back stem and leaf diagram below.

Abbey Hotel	8 | 5 | 0 means 58 years in Abbey Hotel and 50 years in Balmoral Hotel		Balmoral Hotel
(1)	2	0	
(4)	9 7 5 1	1	
(4)	9 8 3 1	2	6 (1)
(11)	9 9 9 9 7 6 6 5 3 3 2	3	4 4 7 (3)
(6)	9 8 7 7 5 0	4	0 0 5 5 6 9 (6)
(1)	8	5	0 0 0 0 1 3 6 6 7 (9)
		6	2 3 3 4 5 7 (6)
		7	0 1 5 (3)

For the Balmoral Hotel,

(*a*) write down the mode of the age of the residents,

(1)

(*b*) find the values of the lower quartile, the median and the upper quartile.

(3)

(*c*) (i) Find the mean, $\bar{x}$, of the age of the residents.

(ii) Given that $\Sigma x^2 = 81\,213$, find the standard deviation of the age of the residents.

(4)

One measure of skewness is found using

$$\frac{\text{mean} - \text{mode}}{\text{standard deviation}}$$

(*d*) Evaluate this measure for the Balmoral Hotel.

(2)

For the Abbey Hotel, the mode is 39, the mean is 33.2, the standard deviation is 12.7 and the measure of skewness is –0.454.

(*e*) Compare the two age distributions of the residents of each hotel.

(3)

11. In a study of how students use their mobile telephones, the phone usage of a random sample of 11 students was examined for a particular week.

The total length of calls, y minutes, for the 11 students were

17, 23, 35, 36, 51, 53, 54, 55, 60, 77, 110

(*a*) Find the median and quartiles for these data.

(3)

A value that is greater than $Q_3 + 1.5 \times (Q_3 - Q_1)$ or smaller than $Q_1 - 1.5 \times (Q_3 - Q_1)$ is defined as an outlier.

(*b*) Show that 110 is the only outlier.

(2)

(*c*) Draw a box plot for these data indicating clearly the position of the outlier.

(3)

The value of 110 is omitted.

(*d*) Show that S_{yy} for the remaining 10 students is 2966.9

(3)

These 10 students were each asked how many text messages, x, they sent in the same week. The values of S_{xx} and S_{xy} for these 10 students are $S_{xx} = 3463.6$ and $S_{xy} = -18.3$.

(*e*) Calculate the product moment correlation coefficient between the number of text messages sent and the total length of calls for these 10 students.

(2)

A parent believes that a student who sends a large number of text messages will spend fewer minutes on calls.

(*f*) Comment on this belief in the light of your calculation in part (*e*).

(1)

12. 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)

13. 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.

14. Over a period of time, the number of people x leaving a hotel each morning was recorded. These data are summarised in the stem and leaf diagram below.

Number leaving	3	2 means 32	Totals
2	7 9 9		(3)
3	2 2 3 5 6		(5)
4	0 1 4 8 9		(5)
5	2 3 3 6 6 6 8		(7)
6	0 1 4 5		(4)
7	2 3		(2)
8	1		(1)

For these data,

(*a*) write down the mode,

(1)

(*b*) find the values of the three quartiles.

(3)

Given that $\Sigma x = 1335$ and $\Sigma x^2 = 71\,801$, find

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

(4)

One measure of skewness is found using

$$\frac{\text{mean} - \text{mode}}{\text{standard deviation}}.$$

(*d*) Evaluate this measure to show that these data are negatively skewed.

(2)

(*e*) Give two other reasons why these data are negatively skewed.

(4)

15. 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)