Year 1

~ S T A T I S T I C S ~

Partitioning Data

Measures of Location & Spread

Representations of Data

A-Level Applied Mathematics

S T A T I S T I C S

Partitioning Data
Measures of Location & Spread
Representations of Data

Lesson 1

A-level Statistics : Year 1
Partitioning Data

1.1 Quartiles, Deciles, Percentiles

Given a large amount of numerical data, it's often helpful to;

- Put it in order.
- Partition it into parts, each part containing (roughly) the same number of items of data.

The most simple action would be to partition some data into just two parts. The location of the partition is then given by the statistic called the **median**. Half the data is below the median and half above.

Other popular partitioning involves dividing the data into,

- Quartiles : 4 parts : Q_1, Q_2, Q_3
- Deciles : 10 parts : $D_1, D_2, \ldots, D_9$
- Percentiles : 100 parts : $P_1, P_2, \ldots, P_{99}$

The following "partitioning rule" will make more sense once a couple of examples have been worked through.

1.2 The Partitioning Rule

The partitions that divides the n items of data up into x parts are located at;

$$\frac{1}{x} \times n, \quad \frac{2}{x} \times n, \quad \ldots, \quad \frac{x-1}{x} \times n$$

Each term in the partition list is further rounded as follows;

- If integer, take the mean of this and the next piece of data.
- If NOT integer, take the item of data one place further on.

1.3 Example involving Quartiles

A class of 17 school children contains 8 boys and 9 girls.
They all sit a test, marked out of 24.

Here are their scores for the boys.

Boys	15	15	16	17	19	19	20	21

(i) Work out Q_1, Q_2 and Q_3 for the boys.

(ii) Determine the Inter Quartile Range (IQR) of the boys' data.

(iii) Show the partitioning of the boys' data below;

Boys	15	15	16	17	19	19	20	21

Here are their scores for the girls.

Girls	8	17	18	21	22	22	23	24	24

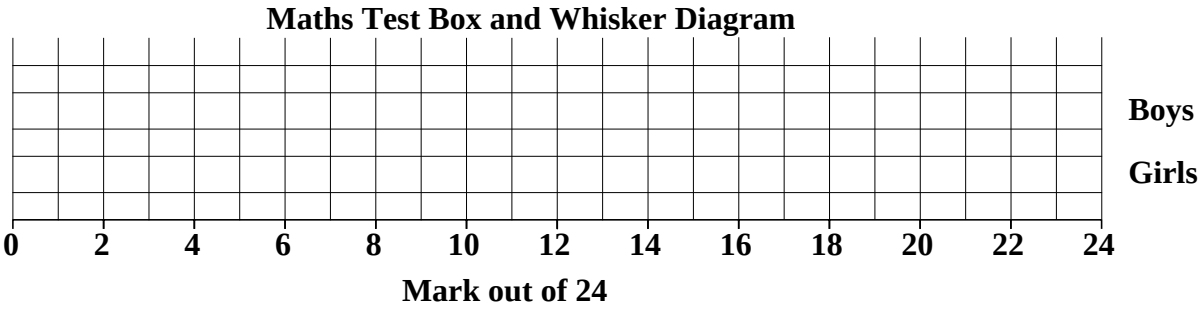
(iv) Work out Q_1, Q_2 and Q_3 for the girls.

(v) Determine the Inter Quartile Range (IQR) for the girls' data.

(vi) Show the partitioning of the girls' data below;

Girls	8	17	18	21	22	22	23	24	24

(vii) Plot a box and whisker diagram for the two sets of data.



(viii) Make two comments about the data based on the box and whisker diagram.

1.4 Example involving Percentiles

A shipment of 138 boxes of apples is examined.

The number of bad apples found in the boxes is given in the following table.

N° of bad apples	0	1	2	3	4	5	6	7	8	9	10
Frequency	55	14	21	12	9	15	3	5	0	3	1

(i) Calculate the 80^{th} percentile.

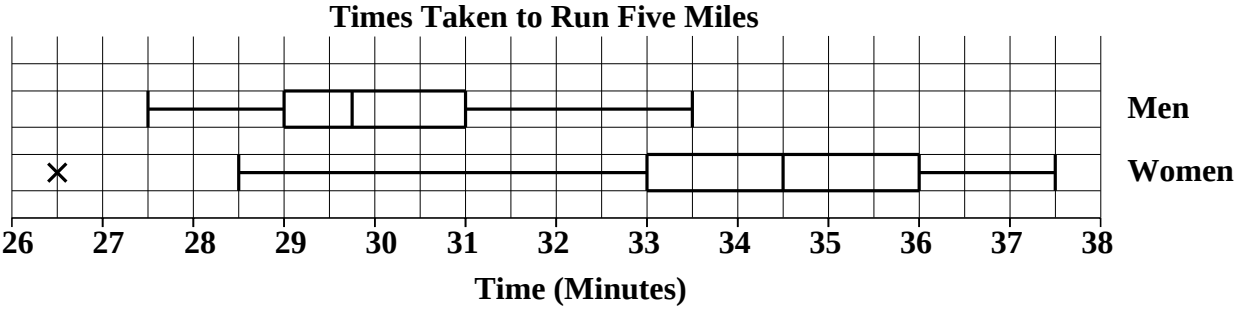
(ii) Calculate the 97^{th} percentile.

(iii) What is the median of the data ?

1.5 Exercise

Question 1

The box and whisker diagram shows the times taken by the members of a sports club to run a five mile course.



(i) For the men

(a) what was the shortest time taken by the fastest runner ?

(b) what was the inter quartile range of the times taken ?

(ii) For the women

(a) what was the average (median) time taken ?

(b) what percentage of runners took longer than 33 minutes ?

(iii) One runner is marked with a cross, ×
Explain what this tells you about this runner.

(iv) Make two comments in which you compare and contrast the data for the men with that of the women.

Question 2

Here is some data;

4	18	24	31	34	37	38	38	38	39
41	41	43	44	44	44	44	45	45	46
47	47	47	48	48	49	49	50	50	50
51	53	54	55	55	56	56	57	59	59
59	60	60	61	62	64	64	65	66	69
70	72	72	73	74	76	78	78	81	84
84	87	90	95						

Let the number of items of data be, n.

(i) State the value of n

(ii) The fourth decile, D_4 is the piece of data in the location given by

$$D_4 \text{ location } = \frac{4}{10} \times n$$

(a) Determine the location of D_4

(b) Hence state the value of D_4

(iii) Determine the value of D_7

(iv) What is the median of the data ?

Question 3

The shoe sizes in a class of 30 pupils are;

Shoe Size	2	2½	3	3½	4	4½	5	5½	6	6½	7
Frequency	3	2	7	5	9	2	1	0	0	0	1

(i) Determine Q_1, Q_2 and Q_3 for this data.

(ii) State the IQR.

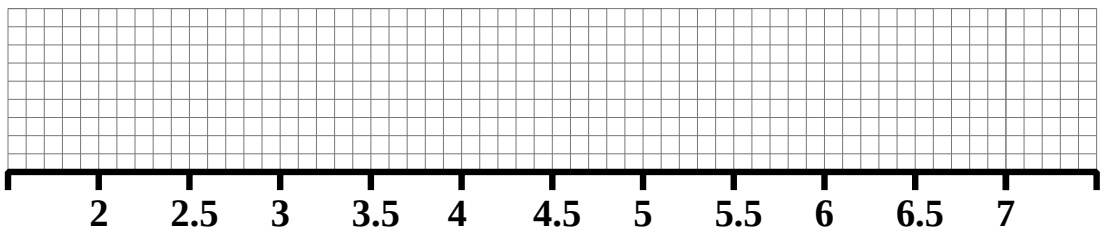
(iii) An outlier is any shoe size, s, for which

$$s > Q_3 + 1.5 \times \text{IQR}$$

Show that this definition classifies the shoe size of 7 as an outlier.

(iv) Plot a box and whisker diagram for the data.

Terminate the whisker that has an outlier at $Q_3 + 1.5 \times \text{IQR}$



Question 4

Partition the following data into four parts by finding the Quartiles, Q_1, Q_2 and Q_3.

7 10 9 7 2 2 8 8 3

Question 5

Find the partitions D_3 and D_9 for the data below.

(The 3^{rd} and 9^{th} deciles)

Number	0	1	2	3	4	5	6	7	8
Frequency	2	3	4	4	5	3	2	1	1

Question 6

Find the partitions P_{33} and P_{66} for the data below.

(The 33^{rd} and 66^{th} percentiles)

Number	10	11	12	13	14	15	16	17	18
Frequency	4	14	24	42	36	21	14	3	2

Question 7

For the data presented below find the Quartiles, Q_1, Q_2 and Q_3

11	15	22	24	26	26	32	35	36	37
37	40	40	41	44	46	48	48	50	51
51	55	62	64	64	67	78	82	82	

1.6 Answers

1.6.1 Solutions (1.3 Example)

(i)

$\frac{1}{4} \times 8 = 2$, which is an integer so $Q_1 = \frac{15 + 16}{2} = 15.5$

$\frac{2}{4} \times 8 = 4$, which is an integer so $Q_2 = \frac{17 + 19}{2} = 18$

$\frac{3}{4} \times 8 = 6$, which is an integer so $Q_3 = \frac{19 + 20}{2} = 19.5$

(ii) $IQR = Q_3 - Q_1$
$= 19.5 - 15.5$
$= 4$ marks

(iii)

Boys	15	15	16	17	19	19	20	21
			^		^		^	

(iv) ARU means “Always Round Up”

$\frac{1}{4} \times 9 = 2.25$, which is NOT an integer (ARU) $Q_1 = 18$

$\frac{2}{4} \times 9 = 4.5$, which is NOT an integer (ARU) $Q_2 = 22$

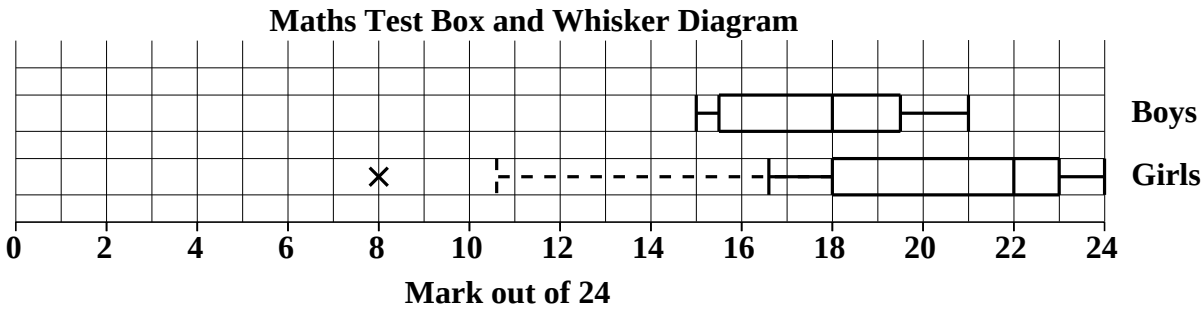
$\frac{3}{4} \times 9 = 6.75$, which is NOT an integer (ARU) $Q_3 = 23$

(v) $IQR = Q_3 - Q_1$
$= 23 - 18$
$= 5$ marks

(vi)

Girls	8	17	18	21	22	22	23	24	24
			^		^		^		

(vii)



(viii) On average (median) the girls have scored higher than the boys.
The girls' results are more spread out than the boys' results.

1.6.2 Solutions (1.4 Example)

N° of bad apples	0	1	2	3	4	5	6	7	8	9	10
Frequency	55	14	21	12	9	15	3	5	0	3	1
Cumulative	55	69	90	102	111	126	129	134	134	137	138

(i) 80^{th} percentile, P_{80}

$$\frac{80}{100} \times 138 = 110.4 \quad \text{(ARU)} \quad \text{Item 111 is 4 bad apples}$$

(ii) 97^{th} percentile, P_{13}

$$\frac{97}{100} \times 138 = 133.86 \quad \text{(ARU)} \quad \text{Item 134 is 7 bad apples}$$

(iii) 50^{th} percentile, P_{50}

$$\frac{50}{100} \times 138 = 69$$

The average (mean) of item 69 and iten 70 is 1.5 bad apples

1.6.3 Solutions (1.5 Exercise)

Answer 1

(i) **(a)** 27 minutes 30 seconds
(b) 2 minutes

(ii) **(a)** 34 minutes 30 seconds
(b) 75%

(iii) The cross is a woman who is an outlier
She ran the race in a remarkably short time of 26 minutes 30 seconds

(iv) On average the men ran faster and so took less time
Median for Men is 29 minutes 45 seconds, for women 34 minutes 30 seconds

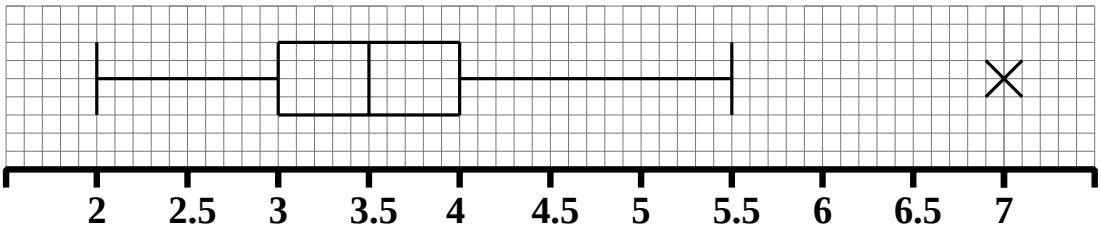
The men were less spread out than the women
The IQR for the men was 2 minutes, but for the women it was 3 minutes

Answer 2

(i) 64
(ii) **(a)** position is 26^{th} (25.6 rounded up)
(b) $D_4 = 49$ (The first of the two 49s)
(iii) position is 45^{th} (44.8 rounded up) so $D_7 = 62$
(iv) median is 53.5, which is also D_5

Answer 3

(i) Q_1 position is 8^{th} (7.5 rounded up) so $Q_1 = 3$
$Q_2 = 3½$
Q_3 position is 23^{rd} (22.5 rounded up) so $Q_3 = 4$
(ii) IQR = 1
(iii) $Q_3 + 1.5 \times IQR = 4 + 1.5 \times 1 = 5.5$ $\therefore$ 7 is an outlier □
(iv) Appropriate title and labelled axis added to the following...



Answer 4

Put the data in order !
$Q_1 = 3$, $Q_2 = 7$, $Q_3 = 8$

Answer 5

$D_3 = 2$, $D_9 = 6$

Answer 6

$P_{33} = 13$, $P_{66} = 14$

Answer 7

$Q_1 = 35$, $Q_2 = 44$, $Q_3 = 55$

Lesson 2

A-level Statistics : Year 1
Partitioning Data

2.1 Box and Whisker Diagrams with Outliers

A Geography class of 30 pupils, when given a test, yields %age scores as follows;

57	74	63	58	57	12	37	45	28
42	51	48	35	64	55	08	59	46
54	50	56	63	60	37	47	65	67
58	53	36						

Using a computer spreadsheet (Excel, for example) the data is put into order.

08	12	28	35	36	37	37	42	45
46	47	48	50	51	53	54	55	56
57	57	58	58	59	60	63	63	64
65	67	74						

(i) Which two, perhaps three, items of data are outliers ?

☞

(ii) Determine the quartiles Q_1, Q_2 and Q_3

☞

(iii) Calculate the IQR

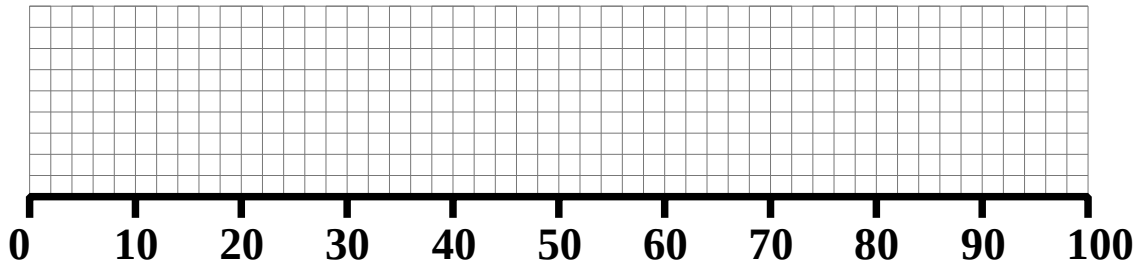
☞

(iv) An outlier is an item of data that is $1.5 \times$ IQR either above the Q_3 or below Q_1. Calculate these 'cut off' values.

☞

(v) Hence produce a box plot of the data with any outliers clearly marked.

☞



(vi) Is the skew positive, or negative or is the distribution symmetrical ?

☞

2.2 Exercise

Question 1

S1 Examination question from June 2005, Q4, (edited)

Aeroplanes fly from City *A* to City *B*. Over a long period of time the number of minutes delay in take-off from City *A* was recorded. the minimum delay was 5 minutes and the maximum delay was 63 minutes. A quarter of all delays were at most 12 minutes, half were at most 17 minutes and 75% were at most 28 minutes. Only one of the delays was longer than 45 minutes.

An outlier is an observation that falls either $1.5 \times$ (interquartile range) above the upper quartile or $1.5 \times$ (interquartile range) below the lower quartile.

(a) Calculate the outlier 'cut-off' values.

[2 marks]

(b) On graph paper draw a box plot to represent these data.

[4 marks]

(c) Calculate

(i) $Q_2 - Q_1$

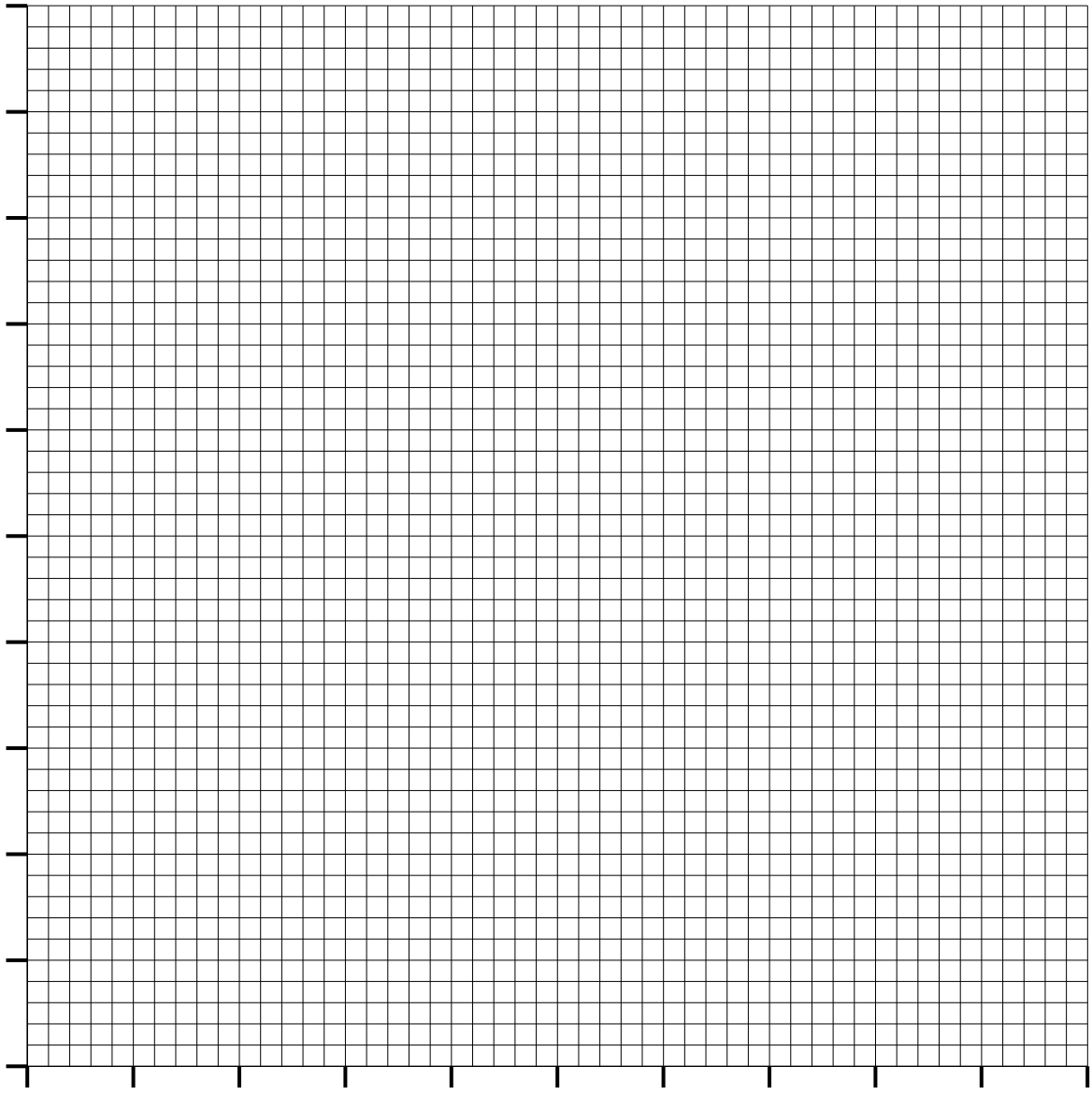
(ii) $Q_3 - Q_2$

Is the distribution negatively skewed or symmetrical or positively skewed ?

[3 marks]

(d) Suggest how the distribution might be interpreted by a passenger who frequently flies from City *A* to City *B*.

[1 mark]

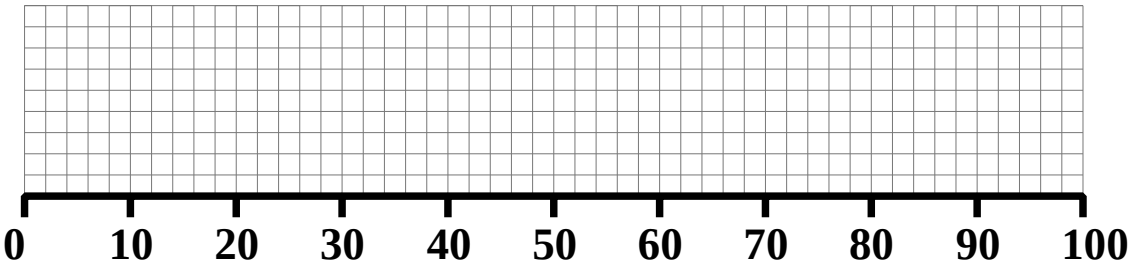


Question 2

A supermarket knows the age of each person who pays for their shopping using a loyalty card. They are interested in the 'age profile' of their customers.

The data reveals that the lowest age was 18 (persons below that age cannot own a loyalty card) and the highest was 98.
The lower quartile was 30 and the upper quartile was 68.
The median age of the shoppers was 44.

(i) Plot a box and whisker diagram to show this information.



(ii) The interquartile rage (IQR) = the upper quartile - the lower quartile.

$$IQR = Q_3 - Q_1$$

What is the IQR of the age profile of supermarket customers ?

(iii) Calculate

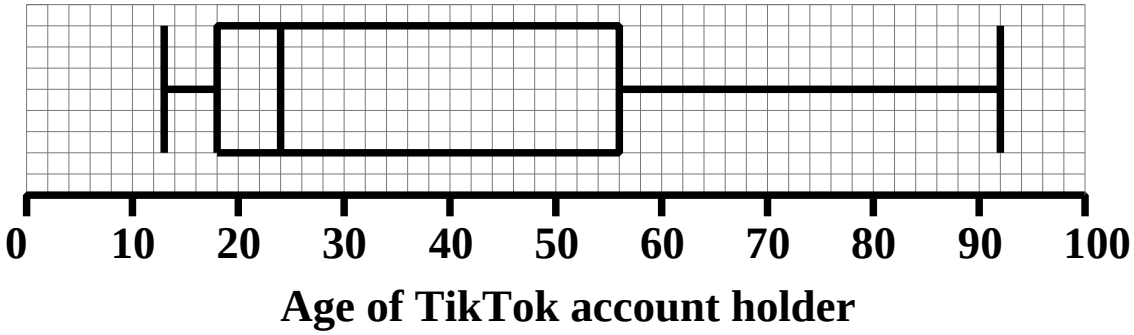
(a) $Q_2 - Q_1$

(b) $Q_3 - Q_2$

Is the distribution negatively skewed or symmetrical or positively skewed ?

Question 3

The box plot below shows the age profile of all persons with a TikTok account.



Age of TikTok account holder

State **(i)** The lower quartile, Q_1

(ii) The upper quartile, Q_3

(iii) The interquartile rage, IQR

(iv) The median age, Q_2

(v) $Q_2 - Q_1$

(vi) $Q_3 - Q_2$

Is the data is skewed ?

Justify your answer and give an interpretation of what this reveals in terms of the age profile of TikTok users.

Question 4

S1 examination question, January 2010, Q2 (edited)

The 19 employees of a company take an aptitude test.
The scores out of 40;

07	18	18	24	24	26	28	32	33	33
33	34	35	39	40	40	40	40	40	

Find

(a) the median score

[1 mark]

(b) the interquartile range

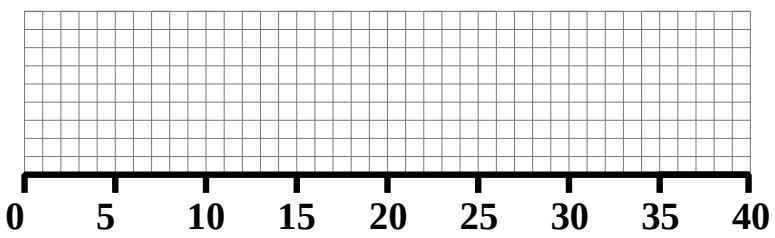
[3 marks]

The company director decides that any employees whose scores are so low that they are outliers will undergo retraining.
An outlier is an observation whose value is less than the lower quartile minus 1.0 times the interquartile range.

(c) Explain why there is only one employee who will undergo retraining.

[2 marks]

(d) On the graph paper below draw a box plot to illustrate the employee's scores



[3 marks]

Question 5

In a 100 metre sprint the running time (in seconds) of 23 runners was as follows.

15.8	14.2	15.7	15.3	17.0	15.2	12.9	13.8	15.7
16.1	17.6	16.4	12.9	13.4	13.6	14.8	14.6	14.1
16.4	15.7	13.4	14.4	15.2				

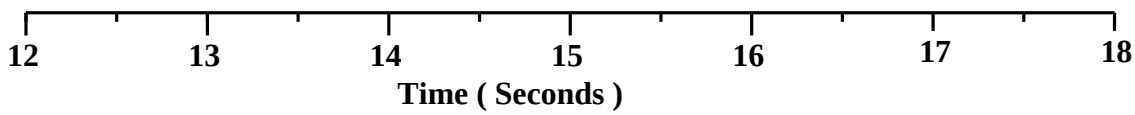
(i) Put the data in order.
Take care to get this correct and try to get a methodical system going.

(ii) Calculate the lower quartile.

(iii) Calculate the upper quartile.

(iv) Calculate the median.

(v) Plot a box and whisker diagram of the data.



(vi) What is the interquartile range for this data ?

2.2 Solutions

2.2.1 Answer to the Introductionary Problem

(i) The scores of 8% and 12% are well below the bulk of the data.
The score of 28% could also be considered an outlier.

(ii) Lower Quartile, Q_1
30 items of data $\Rightarrow 30 / 4 = 7.5$
Not an integer so take the 8^{th} item of data which is 42

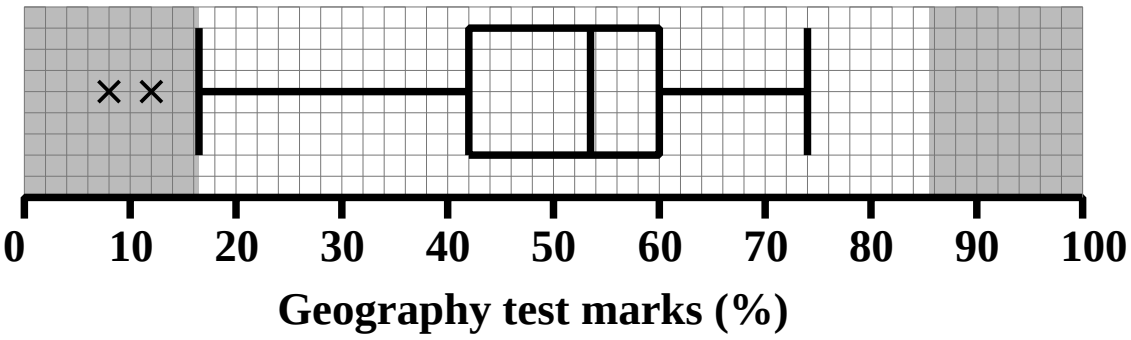
Median, Q_2
30 items of data $\Rightarrow 30 / 2 = 15$
Integer so take mean of 15^{th} and 16^{th} items of data which is 53.5

Upper Quartile, Q_3
30 items of data $\Rightarrow 3 \times 30 / 4 = 22.5$
Not an integer so take the 23^{rd} item of data which is 59

(iii) $IQR = Q_3 - Q_1 \Rightarrow IQR = 17$

(iv) $Q_3 + 1.5 \times IQR \Rightarrow 60 + 1.5 \times 17 = 85.5$
$Q_1 - 1.5 \times IQR \Rightarrow 42 - 1.5 \times 17 = 16.5$

(v)



(vi)
The distribution is negatively skewed
because
$Q_3 - Q_2 < Q_2 - Q_1$

2.2.1 Answers to the Exercise

Answer 1

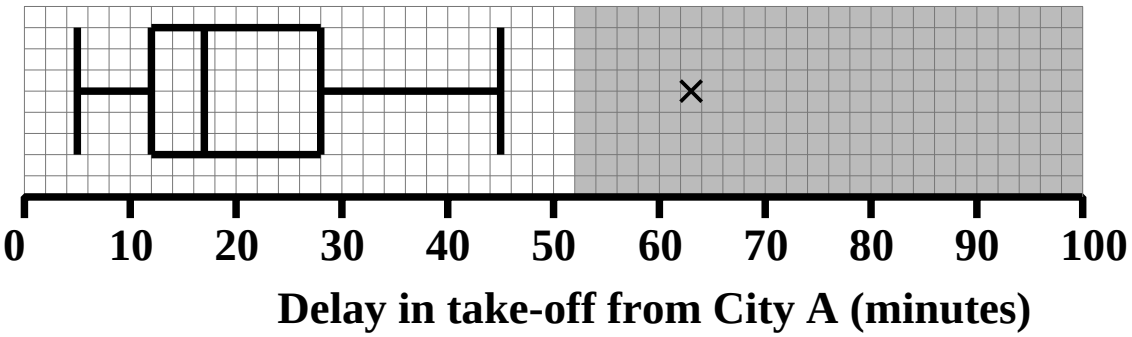
(a) Question is telling us that: $Q_1 = 12$ $Q_2 = 17$ $Q_3 = 28$ $\therefore$ IQR = 16

Cut off values are:

$12 - 1.5 \times 16 = -12$ no outliers at low end

$28 + 1.5 \times 16 = 52$ 63 is a high end outlier

(b)



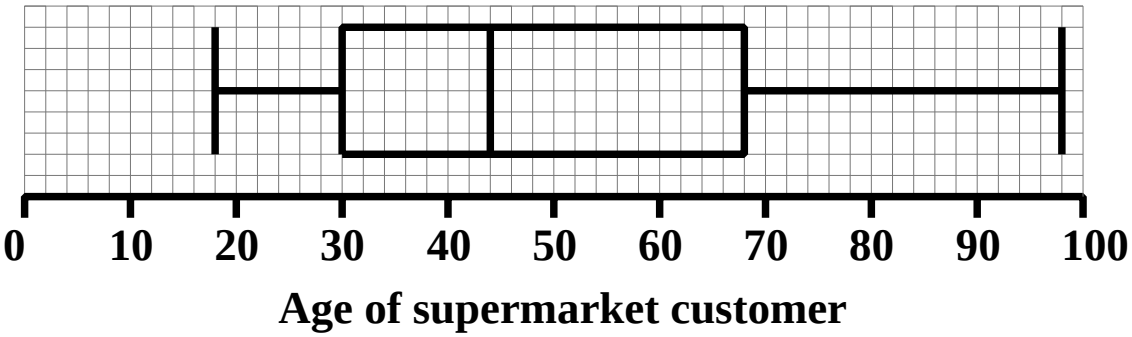
(c) **(i)** $Q_2 - Q_1 = 5$ **(ii)** $Q_3 - Q_2 = 11$

Distribution has positive skew because $Q_3 - Q_2 > Q_2 - Q_1$

(d) Many aircraft take off with only a short delay.
Most passengers are probably willing to accept such small delays.

Answer 2

(i)



(ii) IQR = 68 – 30
IQR = 38

(iii) **(a)** $Q_2 - Q_1 = 44 - 30 = 14$

(b) $Q_3 - Q_2 = 68 - 44 = 24$

As $Q_3 - Q_2 > Q_2 - Q_1$ the skew is positive.

Answer 3

(i) LQ, Q_1 is 18

(ii) UQ, Q_3 is 56

(iii) IQR is 38

(iv) Median, Q_2 is 24

(v) $Q_2 - Q_1 = 24 - 18 = 6$

(vi) $Q_3 - Q_2 = 56 - 24 = 32$

Skewed (positively) because $Q_3 - Q_2 > Q_2 - Q_1$

Lots of younger people have TikTok accounts

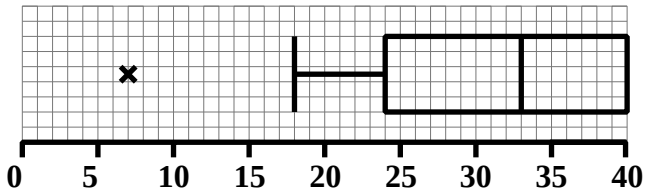
Answer 4

(a) There are 19 pieces of data.
$\therefore$ Median is piece of data N° 10 (9.5 with ARU applied)
Median in a score of 33 out of 40.

(b) Lower quartile is piece of data N° 5 (4.75 with ARU applied)
$Q_1 = 24$
Upper quartile is piece of data N° 15 (14.25 with ARU applied)
$Q_3 = 40$
$\therefore$ The interquartile range is 16

(c) 'cut off' is given by $24 - 1.0 \times 16 = 8$
Only one employee scored less than 8 out of 40 ☐

(d)



A whisker that goes down to the 'cut off' at 8 is also acceptable

Answer 5

(i) Data in order is;

12.9	12.9	13.4	13.4	13.6	13.8	14.1	14.2	14.4
14.6	14.8	15.2	15.2	15.3	15.7	15.7	15.7	15.8
16.1	16.4	16.4	17.0	17.6				

(ii) Lower Quartile, Q_1

23 items of data $\Rightarrow$ 23 / 4 = 5.75

Not an integer so take 6^{th} item of data which is 13.8 seconds

(iii) Upper Quartile, Q_3

23 items of data $\Rightarrow 3 \times 23 / 4 = 17.25$

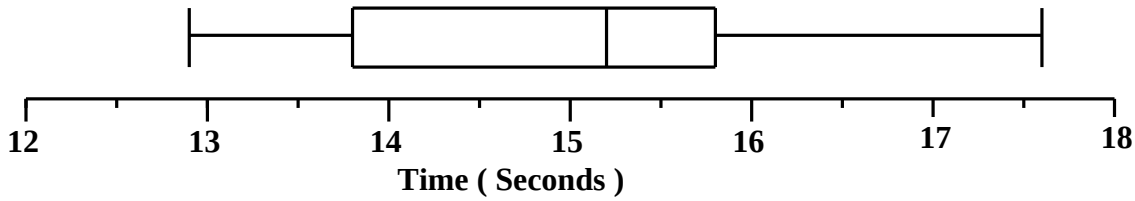
Not an integer so take 18th item of data which is 15.8 seconds.

(iv) Median, Q_2

23 items of data $\Rightarrow$ 23 / 2 = 11.5

Not an integer so take 12th item of data which is 15.2 seconds

(v)



(vi) IQR = 2s

3.1 Quartiles from Grouped Frequency Data

When the data under consideration is presented in a grouped frequency table, the method of calculating statistics such as Quartiles, Deciles, and Percentiles changes. The method in this case is linear interpolation.

3.2 Example

Determine the median mass of the following Red Admiral butterflies;

Mass (grams)	Frequency	Cumulative frequency
1.0 - 1.9	7	7
2.0 - 2.9	47	54
3.0 - 3.9	59	113
4.0 - 4.9	28	141

Give your answer in grams, correct to two decimal places.

3.3 Exercise

Question 1

In a survey, parents were asked how long they spent cooking the family evening meal on a particular day. The results are presented in summary form below.

time (to nearest minute)	Number of parents	
1 - 15	115	
16 - 25	46	
26 - 35	36	
36 - 55	22	
56 - 80	14	

(i) Add a column for the Cumulative Frequencies to the table.

(ii) Use linear interpolation to find the median of this data.
Give your answer in minutes and seconds.

Question 2

The mass of 62 adult Polar Bears were recorded to the nearest kilogram.
The data is summarised in the following grouped frequency table:

Mass kg	Frequency	Cumulative frequency
100 - 199	6	
200 - 299	12	
300 - 399	20	
400 - 499	14	
500 - 599	10	

(i) Complete the cumulative frequency column in the table.

(ii) Determine the Lower & Upper Quartiles, Q_1 and Q_3
Give your answers in kg correct to one decimal place.

(iii) State the interquartile range.

Question 3

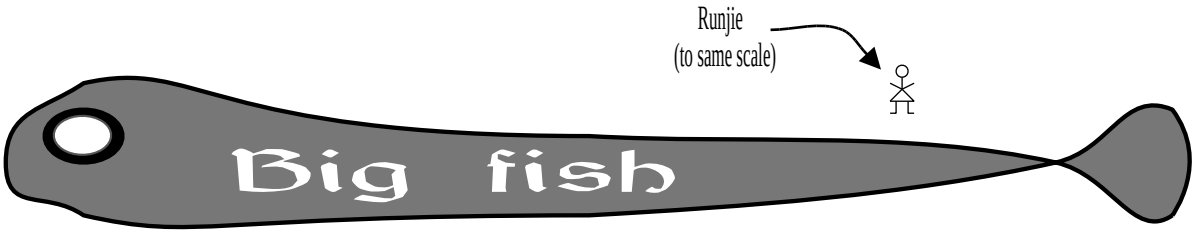
The grouped frequency distribution shown below gives the speed of service of the top fifty performers in men's professional tennis in 1992.

service speed (mph)	Number of serves	
90 - 94	2	
95 - 99	7	
100 - 104	9	
105 - 109	14	
110 - 114	9	
115 - 119	4	
120 - 124	3	
125 - 129	2	

(i) Find the third decile, D_3.

(ii) Find the thirty-nineth percentile, P_{39}.

Question 4



The owner of a pleasure fishing vessel records the lengths of the fish caught by his customers during a season.
The data is presented in summarised form in the following table;

length (to nearest cm)	Number of fish	Cumulative frequency
60 - 64	11	11
65 - 69	49	60
70 - 74	190	250
75 - 79	488	738
80 - 84	632	1370
85 - 89	470	1840
90 - 94	137	1977
95 - 99	23	2000

Find the quartiles, Q_1, Q_2 and Q_3 of this data.

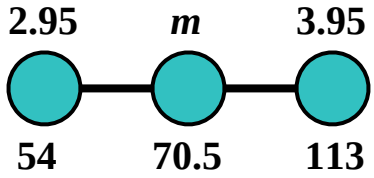
3.4 Answers

3.4.1 Solution to the 3.2 Example

To calculate which class the median is in;

$$\text{For the median : } \frac{141}{2} = 70.5 \text{ which is in the class } 3 - 3.9$$

Do not apply ARU (Always Round Up) as the individual pieces of data are not known.
The class sought is 3.0 - 3.9 but notice the gaps between the classes given in the table.
This means that the class we are interested in is slightly larger than given in the table.
(That is, 2.95 - 3.95)
Linear Interpolation :



$$\frac{m - 2.95}{3.95 - 2.95} = \frac{70.5 - 54}{113 - 54}$$

$$m = 3.23 \text{ g}$$

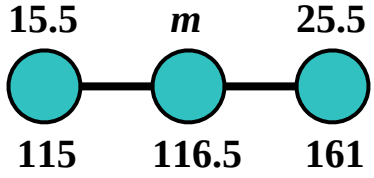
3.4.2 Solution to the 3.3 Exercise

Answer 1
(i)

time (to nearest minute)	Number of parents	Cumulative frequency
1 - 15	115	115
16 - 25	46	161
26 - 35	36	197
36 - 55	22	219
56 - 80	14	233

$$\text{For the median : } \frac{233}{2} = 116.5 \text{ which is in the class } 16 - 25$$

Linear Interpolation :



$$\frac{m - 15.5}{25.5 - 15.5} = \frac{116.5 - 115}{161 - 115}$$

$$m = 15.83 \text{ minutes} \quad (15 \text{ minutes } 50 \text{ seconds})$$

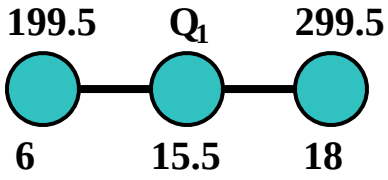
Answer 2

(i)

Mass kg	Frequency	Cumulative frequency
100 - 199	6	6
200 - 299	12	18
300 - 399	20	38
400 - 499	14	52
500 - 599	10	62

For Q_1 : $\frac{62}{4} = 15.5$ which is in the class 200 - 299

(ii) Linear interpolation :

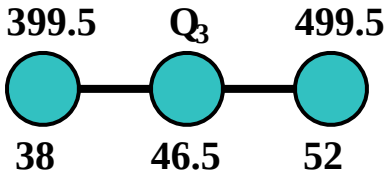


$$\frac{Q_1 - 199.5}{299.5 - 199.5} = \frac{15.5 - 6}{18 - 6}$$

$$Q_1 = 278.7 \text{ kg}$$

For Q_3 : $\frac{3 \times 62}{4} = 46.5$ which is in the class 400 - 499

Linear interpolation :



$$\frac{Q_3 - 399.5}{499.5 - 399.5} = \frac{46.5 - 38}{52 - 38}$$

$$Q_3 = 460.2 \text{ kg}$$

(iii) Thus the interquartile range is 181.5 kg

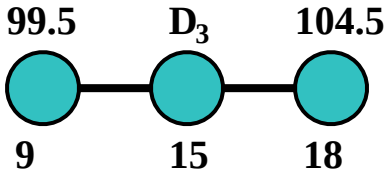
Answer 3

service speed (mph)	Number of serves	cumulative frequency
90 - 94	2	2
95 - 99	7	9
100 - 104	9	18
105 - 109	14	32
110 - 114	9	41
115 - 119	4	45
120 - 124	3	48
125 - 129	2	50

(i)

For D_3 : $\frac{3 \times 50}{10} = 15$ which is in the class 100 - 104

Linear interpolation :



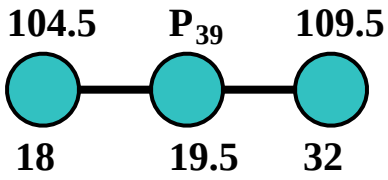
$$\frac{D_3 - 99.5}{104.5 - 99.5} = \frac{15 - 9}{18 - 9}$$

$$D_3 = 102.8 \text{ mph}$$

(ii)

For P_{39} : $\frac{39 \times 50}{100} = 19.5$ which is in the class 105 - 109

Linear interpolation :



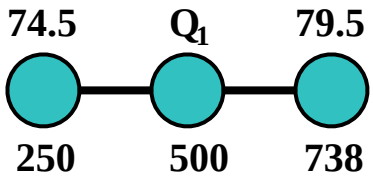
$$\frac{P_{39} - 104.5}{109.5 - 104.5} = \frac{19.5 - 18}{32 - 18}$$

$$D_3 = 105.0 \text{ mph}$$

Answer 4

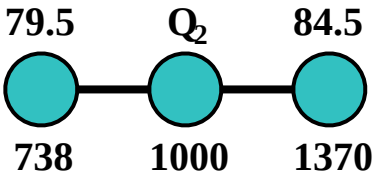
length (to nearest cm)	Number of fish	Cumulative frequency
60 - 64	11	11
65 - 69	49	60
70 - 74	190	250
75 - 79	488	738
80 - 84	632	1370
85 - 89	470	1840
90 - 94	137	1977
95 - 99	23	2000

For Q_1 : $\frac{2000 \times 1}{4} = 500$ which is in the class 75 - 79



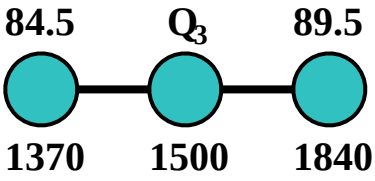
$$\frac{Q_1 - 74.5}{79.5 - 74.5} = \frac{500 - 250}{738 - 250} \Rightarrow Q_1 = 77.06 \text{ cm}$$

For Q_2 : $\frac{2000 \times 2}{4} = 1000$ which is in the class 80 - 84



$$\frac{Q_2 - 79.5}{84.5 - 79.5} = \frac{1000 - 738}{1370 - 738} \Rightarrow Q_2 = 81.57 \text{ cm}$$

For Q_3 : $\frac{2000 \times 3}{4} = 1500$ which is in the class 85 - 89



$$\frac{Q_3 - 84.5}{89.5 - 84.5} = \frac{1500 - 1370}{1840 - 1370} \Rightarrow Q_3 = 85.88 \text{ cm}$$

Lesson 4

A-level Statistics : Year 1
Partitioning Data

4.1 Skew & Standard Deviation

Fortune smiles on Lucky Lucy

Example
Over the course of a lucky year, Lucy finds the following lucky seven coins simply sitting on the ground.

10p 2p 1p 5p 2p 1p £2

(i) Find the median of her lucky finds.

(ii) Find the mean of her lucky finds.

(iii) Find the standard deviation of her lucky finds.

(iv) Evaluate the skew of the lucky finds, as given by;

$$skew = \frac{3\,(mean - median)}{standard\ deviation}$$

(v) Explain what your part **(iv)** answer is telling you about the distribution of the lucky finds.

4.2 Exercise

Question 1

S1 examination question from January 2006, Q1 (edited)

Over a period of time, the number of people x leaving a hotel each morning was recorded. These data are summarised below.

27	29	29	32	32	33	35	36	40	41
44	48	49	52	53	53	56	56	56	58
60	61	64	65	72	73	81			

For these data,

(a) write down the mode

[1 mark]

(b) find the values of the three quartiles

[3 marks]

Given that $\Sigma x = 1335$ and $\Sigma x^2 = 71801$, find

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

[4 marks]

One measure of skewness is found using

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

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

[2 marks]

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

[4 marks]

Question 2

S1 examination question from May 2008, Q1 (edited)

The age in years of the residents of two hotels are shown below.

Abbey Hotel

2	11	15	17	19	21	23	28	29	32
33	33	35	36	36	37	39	39	39	39
40	45	47	47	48	49	58			

Balmoral Hotel

26	34	34	37	40	40	45	45	46	49
50	50	50	50	51	53	56	56	57	62
63	63	64	65	67	70	71	75		

For the Balmoral Hotel,

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

[1 mark]

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

[3 marks]

(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 marks]

One measure of skewness is found using

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

(d) Evaluate this measure for the Balmoral Hotel

[2 marks]

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 marks]

Question 3

S1 examination question from June 2014, Q1 (edited)

A random sample of 35 homeowners was taken from each of the villages Greenslax and Penville and their ages were recorded. The results are summarised below.

Greenslax

27	28	37	38	39	40	44	44	44	52
52	55	56	56	61	61	62	64	65	66
68	71	71	73	74	76	76	76	78	82
83	84	88	89	94					

Penville

25	25	26	27	28	28	29	31	31	31
32	33	34	34	35	36	37	39	40	41
42	44	47	50	50	55	55	55	62	65
66	66	70	75	99					

Some of the quartiles for these two distributions are given in the table below.

	Greenslax	Penville
Lower quartile, Q_1	*a*	31
Median, Q_2	64	39
Upper quartile, Q_3	*b*	55

(a) Find the value of *a* and the value of *b*

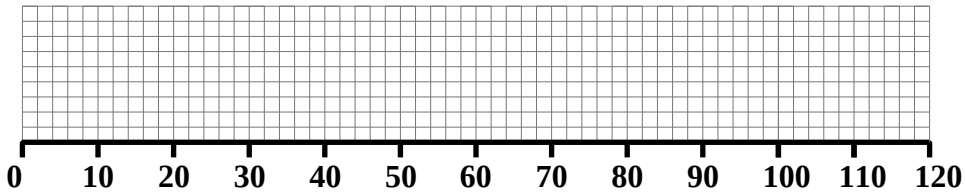
[2 marks]

An outlier is a value that falls either

more than $1.5 \times (Q_3 - Q_1)$ above Q_3

or more than $1.5 \times (Q_3 - Q_1)$ below Q_1

(b) Draw a box plot to represent the data from Penville.
Show clearly any outliers.



[4 marks]

(c) State the skewness of each distribution.
Justify your answers.

[3 marks]

4.3 Answers

4.3.1 Solution to *Fortune Smiles on Lucky Lucy*

Data in order is

1p 1p 2p 2p 5p 10p 200p

(i) Median : 7 items of data
$7 / 2 = 3.5^{th}$
Not an integer
$\therefore$ Round up (ARU) to 4^{th} which gives the median as 2p

(ii) Mean:

$$\mu = \frac{1 + 1 + 2 + 2 + 5 + 10 + 200}{7}$$

$$= \frac{221}{7}$$

$$= 31.6 \text{ pence}$$

(iii) Standard deviation:

$$\sigma^2 = \frac{1^2 + 1^2 + 2^2 + 2^2 + 5^2 + 10^2 + 200^2}{7} - \left(\frac{221}{7}\right)^2$$

$$= \frac{1 + 1 + 4 + 4 + 25 + 100 + 40000}{7} - \left(\frac{221}{7}\right)^2$$

$$= \frac{40135}{7} - \frac{48841}{49}$$

$$= \frac{280945 - 48841}{49}$$

$$= \frac{232104}{49}$$

$$\sigma = 68.8 \text{ pence}$$

(iv) Skew :

$$skew = \frac{3\,(31.6 - 2)}{68.8}$$

$$= 1.3$$

(v) Strong positive skew means that Lucy's finds are all piled up at the small coin values. The £2 find was an outlier (unusual)

4.3.2 Solutions to 4.2 Exercise

Answer 1

(a) mode = 56

(b)

$$\text{For } Q_1 : \frac{27 \times 1}{4} = 6.75 \Rightarrow \text{ARU} \Rightarrow 7^{th} \text{ which is } 35$$

$$\text{For } Q_2 : \frac{27 \times 2}{4} = 13.5 \Rightarrow \text{ARU} \Rightarrow 14^{th} \text{ which is } 52$$

$$\text{For } Q_3 : \frac{27 \times 3}{4} = 20.25 \Rightarrow \text{ARU} \Rightarrow 21^{st} \text{ which is } 60$$

(c)

$$\mu = \frac{\sum x}{\sum f}$$
$$= \frac{1335}{27}$$
$$= \frac{445}{9} \quad \text{(exact)}$$
$$= 49.4$$

$$\sigma = \sqrt{\frac{\sum x^2}{\sum f} - \mu^2}$$
$$= \sqrt{\frac{71\,801}{27} - \left(\frac{445}{9}\right)^2}$$
$$= 14.6472936...$$
$$= 14.6$$

(d)

$$skew = \frac{mean - mode}{standard\ deviation}$$
$$= \frac{49.4 - 56}{14.6}$$
$$= -0.452$$

That is, negative skew □

(e) mean < median < mode because $49.4 < 52 < 56$

$\therefore$ skew is negative

$Q_3 - Q_2 < Q_2 - Q_1$ because $60 - 52 < 52 - 35$

That is, $8 < 17$

$\therefore$ skew is negative

Answer 2

For the Balmoral Hotel,

(a) mode = 50

(b)

$$\text{For } Q_1 : \frac{28 \times 1}{4} = 7 \Rightarrow \text{Integer}$$

$$\Rightarrow \frac{7^{th} + 8^{th}}{2} = \frac{45 + 45}{2} \text{ which is } 45$$

$$\text{For } Q_2 : \frac{28 \times 2}{4} = 14 \Rightarrow \text{Integer}$$

$$\Rightarrow \frac{14^{th} + 15^{th}}{2} = \frac{50 + 51}{2} \text{ which is } 50.5$$

$$\text{For } Q_3 : \frac{28 \times 3}{4} = 28 \Rightarrow \text{Integer}$$

$$\Rightarrow \frac{21^{st} + 22^{nd}}{2} = \frac{63 + 63}{2} \text{ which is } 63$$

(c) **(i)**

$$mean = \frac{1469}{28} \text{ (exact)}$$

$$= 52.5$$

(ii)

$$standard\ deviation = \sqrt{\frac{81\,213}{28} - \left(\frac{1469}{28}\right)^2}$$

$$= 12.164\,004\,69$$

$$= 12.2$$

(d)

$$skew = \frac{mean - mode}{standard\ deviation}$$

$$= \frac{52.5 - 50}{12.2}$$

$$= 0.2049$$

(e) The average age at the Balmoral is higher than at the Abbey
(When using either mean, median or mode)
So older residents on average at the Balmoral Hotel
Lower standard deviation at Balmoral of 12.2 implies there is less variation in the ages of the residents there compared with the Addey, which has 12.7
Positive skew at Balmoral implies bunching at the younger end of their age range, whereas the opposite is true at the Abbey with it's negative skew.

Answer 3

(a) For Greenslax,

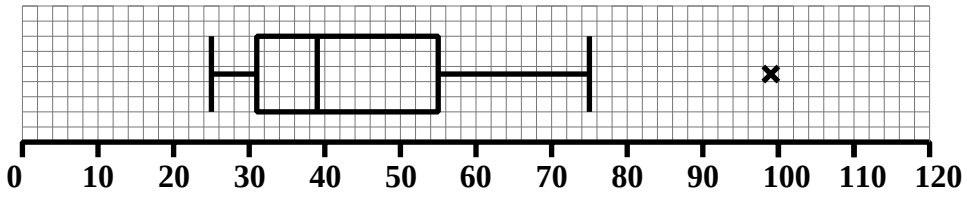
$$\text{For } Q_1 \ : \ \frac{35 \times 1}{4} \ = \ 8.75 \ \Rightarrow \ \text{ARU} \ \Rightarrow \ 9^{th} \text{ which gives } \ a \ = \ 44$$

$$\text{For } Q_3 \ : \ \frac{35 \times 3}{4} \ = \ 26.25 \ \Rightarrow \ \text{ARU} \ \Rightarrow \ 27^{th} \text{ which gives } \ b \ = \ 76$$

(b) For Penville, calculating the "cut offs",

$$\text{high end} \ : \ 55 + 1.5 \times (55 - 31) \ = \ 91$$

$$\text{low end} \ : \ 31 - \ 1.5 \times (55 - 31) \ = \ -5$$



Age of sample of homeowners in Penville

A whisker that goes up to the "cut off" at 91 is also acceptable

(c) From the information given the most convenient measure of skew to use is that based on the quartiles.

For Greenslax,

$$Q_3 - Q_2 \ = \ 76 - 64 \ = \ 12$$

$$Q_2 - Q_1 \ = \ 64 - 44 \ = \ 20$$

$$Q_3 - Q_2 \ < \ Q_2 - Q_1 \ \text{ indicates negative skew}$$

For Penville,

$$Q_3 - Q_2 \ = \ 55 - 39 \ = \ 16$$

$$Q_2 - Q_1 \ = \ 39 - 31 \ = \ 8$$

$$Q_3 - Q_2 \ > \ Q_2 - Q_1 \ \text{ indicates positive skew}$$

Lesson 5

5.1 Mean & Standard deviation from Grouped Frequency Tables

Example

The time taken for ten runners to complete a race are summarised in the table;

Time taken (minutes)	Number of runners
20 - 24	3
25 - 29	5
30 - 34	2

Find the mean and standard deviation of the times taken.
Give your answers in minutes and seconds.

5.2 Exercise

Question 1

The time taken by 15 shoppers in a supermarket is presented in the table;

Time in Shop (minutes)	Number of shoppers
10 - 16	8
17 - 23	5
24 - 30	2

Find the mean and standard deviation of the shopping times.
Give your answers in minutes and seconds.

In examination questions you are often told Σfx and $\Sigma f x^2$

This means that ***you don't need to work them out***, with considerable less work then being needed to answer the question.

The next question is such a question...

Question 2
The times it took a random sample of swimmers to complete a sponsored swim are summarised in the table;

Time (minutes)	20 - 29	30 - 39	40 - 49	50 - 59	60 - 69
Frequency	5	10	36	20	9

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

$$\Sigma fx = 3740 \qquad \Sigma f x^2 = 183040$$

Determine the mean and the standard deviation of the swim times.
Give your answers in minutes and seconds.

Question 3

The table below summarises data relating to the Average Rated Lifetime (ARL) of a random sample of 200 light bulbs taken from the production line of the Light Up You Life Forever™ LED lightbulb company.

Lifetime, x (to nearest 100 hours)	Number of light-bulbs, f	
700 - 719	10	
720 - 729	14	
730 - 739	16	
740 - 749	21	
750 - 754	35	
755 - 759	41	
760 - 764	38	
765 - 769	15	
770 - 779	7	
780 - 799	3	

(i) By linear interpolation, estimate the median and quartiles of these lifetimes. Give your answers in 100 hours to 1 decimal place.

The formulae for the mean, μ, and the standard deviation, σ, are;

$$u = \frac{\Sigma fx}{\Sigma f} \qquad \sigma = \sqrt{\frac{\Sigma f x^2}{\Sigma f} - \mu^2}$$

(ii) Given that $\Sigma fx = 150\,232.5$ and $\Sigma f x^2 = 112\,899\,573.8$

(b) determine the mean lifetime of a lightbulb.

(c) determine the standard deviation of the data.

(iii) One method of assessing the skewness of a distribution is to calculate

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

Evaluate this, to 1 decimal place, for the above distribution.

(iv) Use the quartiles to assess skewness and state whether or not you feel the result is compatible with your answer to part (iii).

Question 4

S1 examination question from January 2008, Q2

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 $\Sigma x^2 = 724\,961$]

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

[4 marks]

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

[3 marks]

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 marks]

Research suggests that cotinine levels in the blood form a skewed distribution.
One measure of skewness is found using

$$\frac{(Q_1 - 2Q_2 + Q_3)}{(Q_3 - Q_1)}$$

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

[3 marks]

5.3 Answers

5.3.1 Solutions to the 5.1 Example

Time taken (minutes)	Number of runners
20 - 24	3
25 - 29	5
30 - 34	2

Make a fresh table to organise the process,

Time taken, t (minutes)	Number of runners, f	Mid-interval value, x	fx	x^2	fx^2
$19.5 \leqslant t < 24.5$	3	22	66	484	1452
$24.5 \leqslant t < 29.5$	5	27	135	729	3645
$29.5 \leqslant t < 34.5$	2	32	64	1024	2048
	10		265		7145

$$\mu = \frac{\Sigma fx}{\Sigma f}$$

$$= \frac{265}{10}$$

$$= 26.5 \quad \text{(exact)}$$

That is, 26 minutes 30 seconds

$$\sigma = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2}$$

$$= \sqrt{\frac{7145}{10} - 26.5^2} \qquad \text{(using the exact value of the mean)}$$

$$= 3.5 \text{ minutes}$$

That is, 3 minutes 30 seconds

In examination questions, candidates are often told the numerical values of

$$\sum fx \qquad \text{and} \qquad \sum fx^2$$

5.3.2 Solutions to 5.2 Exercise

Answer 1

Time in Shop (minutes)	Number of shoppers
10 - 16	8
17 - 23	5
24 - 30	2

Make a fresh table to organise the process,

Time in shop, t (minutes)	Number of shoppers, f	Mid-interval value, x	xf	x^2	$x^2 f$
$9.5 \leqslant t < 16.5$	8	13	104	169	1352
$16.5 \leqslant t < 23.5$	5	20	100	400	2000
$23.5 \leqslant t < 30.5$	2	27	54	729	1458
	15		258		4810

$$\mu = \frac{\Sigma fx}{\Sigma f}$$

$$= \frac{258}{15}$$

$$= 17.2 \quad \text{(exact)}$$

That is, 17 minutes 12 seconds

$$\sigma = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2}$$

$$= \sqrt{\frac{4810}{15} - 17.2^2} \qquad \text{(using the exact value of the mean)}$$

$$= 4.9826... \text{ minutes}$$

That is, 4 minutes 59 seconds

Answer 2

$$\mu = \frac{\Sigma fx}{\Sigma f}$$

$$= \frac{3740}{80}$$

$$= 46.75 \quad \text{(exact)}$$

That is, 46 minutes 45 seconds

$$\sigma = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2}$$

$$= \sqrt{\frac{183\,040}{80} - 46.75^2} \qquad \text{(using the exact value of the mean)}$$

$$= 10.12114124... \text{ minutes}$$

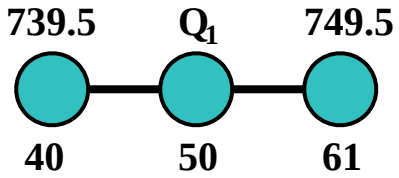
That is, 10 minutes 7 seconds

Answer 3

Lifetime, x (to nearest 100 hours)	Number of light-bulbs, f	Cumulative frequency
700 - 719	10	10
720 - 729	14	24
730 - 739	16	40
740 - 749	21	61
750 - 754	35	96
755 - 759	41	137
760 - 764	38	175
765 - 769	15	190
770 - 779	7	197
780 - 799	3	200

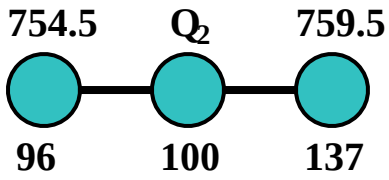
(i)

For Q_1 : $\frac{200 \times 1}{4} = 50$ which is in the class 740 - 749



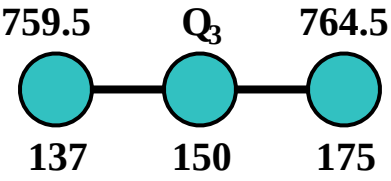
$$\frac{Q_1 - 739.5}{749.5 - 739.5} = \frac{50 - 40}{61 - 40} \Rightarrow Q_1 = 744.3 \text{ hundred hours}$$

For Q_2 : $\frac{200 \times 2}{4} = 100$ which is in the class 755 - 759



$$\frac{Q_2 - 754.5}{759.5 - 754.5} = \frac{100 - 96}{137 - 96} \Rightarrow Q_2 = 755.0 \text{ hundred hours}$$

For Q_3 : $\frac{200 \times 3}{4} = 150$ which is in the class 760 - 764



$$\frac{Q_3 - 759.5}{764.5 - 759.5.5} = \frac{150 - 137}{175 - 137} \Rightarrow Q_3 = 761.2 \text{ humdred hours}$$

(ii)

$$\mu = \frac{\Sigma fx}{\Sigma f}$$

$$= \frac{150\,232.5}{200}$$

$$= 751.1625 \quad \text{(exact) hundred hours}$$

$$\sigma = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2}$$

$$= \sqrt{\frac{112\,899\,573.8}{200} - 751.1625^2} \qquad \text{(using the exact mean)}$$

$$= 15.8987 \quad \text{hundred hours}$$

(iii)

$$skew = \frac{3\,(mean - median)}{standard\ deviation}$$

$$= \frac{3\,(751.1625 - 755)}{15.8987}$$

$$= -0.724 \Rightarrow \textit{Negative skew}$$

(iv)

$Q_3 - Q_2 = 761.2 - 755 = 6.2$: $Q_2 - Q_1 = 755 - 744.3 = 10.7$

$\therefore\ Q_3 - Q_2 < Q_2 - Q_1 \Rightarrow$ *Negative skew*

Thus, COMPATIBLE as both methods indicate negative skew

Answer 4

(a)

$$\begin{aligned} mean &= \frac{2757}{12} \\ &= 229.75 \quad \text{(exact)} \end{aligned}$$

$$\begin{aligned} standard\ deviation &= \sqrt{\frac{724\,961}{12} - 229.75^2} \qquad \text{(using exact mean)} \\ &= 87.34045 \end{aligned}$$

(b) The list needs to be put into order,

$$125,\ 160,\ 169,\ 171,\ 175,\ 186,\ 210,\ 243,\ 250,\ 258,\ 390,\ 420$$

$$\text{For } Q_1 : \frac{12 \times 1}{4} = 3 \Rightarrow \text{Integer} \Rightarrow \frac{169 + 171}{2} = 170$$

$$\text{For } Q_2 : \frac{12 \times 2}{4} = 6 \Rightarrow \text{Integer} \Rightarrow \frac{186 + 210}{2} = 198$$

$$\text{For } Q_3 : \frac{12 \times 3}{4} = 9 \Rightarrow \text{Integer} \Rightarrow \frac{250 + 258}{2} = 254$$

(c)

$$\begin{aligned} Q_3 + 1.5 \times \left(Q_3 - Q_1 \right) &= 254 + 1.5 \times (254 - 170) \\ &= 380 \end{aligned}$$

Patient *B* on Cotinine level 390, and patient *F* on Cotinine level 420 are thus outliers

(d)

$$\begin{aligned} skew &= \frac{(Q_1 - 2Q_2 + Q_3)}{(Q_3 - Q_1)} \\ &= \frac{(170 - 2 \times 198 + 254)}{(254 - 170)} \\ &= \frac{28}{84} \\ &= \frac{1}{3} \end{aligned}$$

The positive skew means that more patents are piled up at the lower end of the distribution than at the upper end.

Lesson 6

A-level Statistics : Year 1

Partitioning Data

6.1 Some Statistical Words

Qualitative Data

: Data described in words

For example: Colour of eyes, favourite sport

Quantitative Data

: Data described in numbers

For example: Heights in cm of students

Discrete Data

: Data which does not need rounding because it changes in steps

For example: Rolls of dice

Continuous Data

: Data which can take any value within an interval and which has to be rounded

For example: Weights of boys to the nearest 100 grams

6.2 Exercise

Question 1

Consider the following words used to describe eight items of data;

Gigantic	Colossal	Very Large	Enormous
Super-Sized	Massive	Huge	Titanic

Is this data Qualitative or Quantitative ?

Question 2

I have a list of the prices for which 100 houses were sold last month.

Is this data Discrete or Continuous ?

Question 3

I am interested in the the water consumption of 50 households in July.

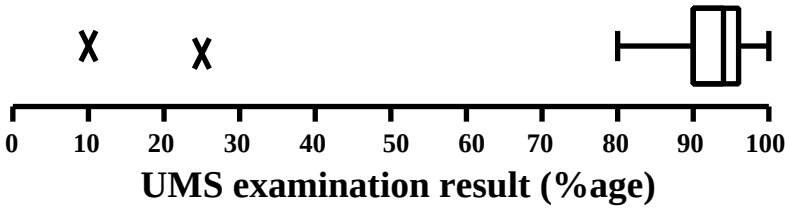
The water meters of the 50 households measure water consumption to the nearest litre.

(i) Am I dealing with Qualitative or Quantitative data ?

(ii) Is the data produced by the water meters Discrete or Continuous ?

Question 4

The box and whisker diagram shows the Statistics examination results of the 200 students at St Trinian's School last year.



(i) What is the lowest mark obtained by a student in the examination ?

[1 mark]

(ii) What is the lower quartile mark obtained by the students ?

[1 mark]

(iii) How many students scored 80% or more ?

[1 mark]

(iv) Describe the skew of the marks **giving a reason for your answer**.

[2 marks]

(v) In a press release The Ministry of Education has announced;
"St Trinian's is to close as pupils have scored less than 30% in exams".
Do you agree with the Ministry ?
Give a reason for your answer.

[2 marks]

Question 5

S1 Examination Question from January 2011 Q2

Keith records the amount of rainfall, in mm, at his school, each day for a week. The results are given below;

2.8 5.6 2.3 9.4 0.0 0.5 1.8

Jenny then records the amount of rainfall, x mm, at the school each day for the following 21 days. the results for the 21 days are summarised below;

$$\Sigma x = 84.6$$

(a) Calculate the mean amount of rainfall during the whole 28 days.

[2 marks]

Keith realises that he has transposed two of his figures.
The number 9.4 should have been 4.9 and the number 0.5 should have been 5.0

Keith corrects these figures.

(b) State, giving your reason, the effect this will have on the mean.

[2 marks]

Question 6

Give an example of data that is;

(a) both DISCRETE and QUALITATIVE

[2 marks]

(b) both CONTINUOUS and QUANTITATIVE

[2 marks]

Question 7

S1 Examination Question from January 2010 Q3

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 $\Sigma fx = 4841$ and $\Sigma fx^2 = 15\,889.5$]

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

[2 marks]

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

[2 marks]

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

[3 marks]

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

[2 marks]

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

[2 marks]

Question 8

S1 Examination Question from January 2011 Q5

On a randomly chosen day, each of the 32 students in a class record 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 marks]

Given that

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

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

[3 marks]

(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 marks]

Question 9

S1 Examination Question from May 2009 Q4

A researcher measures 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
$10 \leqslant l < 12$	5
$12 \leqslant l < 17$	53
$17 \leqslant l < 19$	29
$19 \leqslant l < 21$	15
$21 \leqslant l < 23$	11
$23 \leqslant l < 25$	7

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

[2 marks]

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

[6 marks]

One measure of skewness is given by

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

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

[3 marks]

6.3 Answers

Answer 1
Qualitative

Answer 2
Discrete

Answer 3
(i) Qualitative
(ii) Although a continuous flow of water is being measured, the data coming out of the water meter is to the nearest litre and so is discrete.

Answer 4
(i) 10 %
(ii) 80 %
(iii) 198 students
(iv) From the box and whisker diagram,

$$Q_3 - Q_2 < Q_2 - Q_1$$

and so the skew is negative, the students are piled up at the high end.
(v) Disagree because only two students scored less than 30% with the remaining 198 students scoring 80 % or more.

Answer 5
(a)

$$mean_{28} = \frac{2.8 + 5.6 + 2.3 + 9.4 + 0.0 + 0.5 + 1.8 + 84.6}{28}$$

$$= \frac{107}{28}$$

$$= 3.82 \text{ mm}$$

(b) As the 9.4 was 4.5 too much and ther 0.5 was 4.5 too little the two errors offset each other. Correcting the figures will have no effect on the mean.

Answer 6
(a) Rolls of a dice with each face a different colour.
(b) Heights of adult males in the UK.

Answer 7

(a)

Weight (kg)	Midpoint x kg	Frequency f	Cumulative frequency
0.0 - 1.0	0.50	1	1
1.0 - 2.0	1.50	6	7
2.0 - 2.5	2.25	60	67
2.5 - 3.0	**2.75**	280	347
3.0 - 3.5	3.25	820	1167
3.5 - 4.0	3.75	320	1187
4.0 - 5.0	4.50	10	1497
5.0 - 6.0	**5.50**	3	1500

(b)

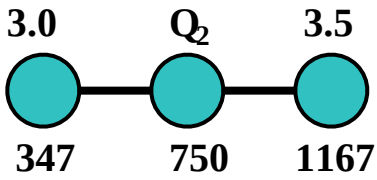
$$\textit{mean} = \frac{4841}{1500} \quad \text{(exeat)}$$
$$= 3.227\,333\,333\ldots$$
$$= 3.227 \text{ kg}$$

(c)

$$\textit{standard deviation} = \sqrt{\frac{15\,889.5}{1500} - \left(\frac{4841}{1500}\right)^2}$$
$$= 0.421\,093\,286\ldots$$
$$= 0.421 \text{ kg}$$

(d)

For Q_2 : $\frac{1500 \times 2}{4} = 750$ which is in the class 3.0 - 3.5



$$\frac{Q_2 - 3.0}{3.5 - 3.0} = \frac{750 - 347}{1167 - 347} \Rightarrow Q_2 = 3.245731707\ldots$$
$$\Rightarrow \textit{median} = 3.246 \text{ kg}$$

(e) Technically, the mean is less than the median and so the skew is negative.

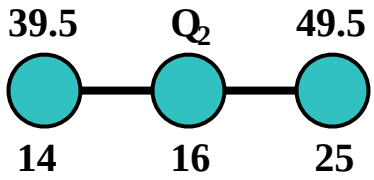
However, you could argue that the mean and the median are almost the same in which case there isn't much skew and the distribution is better described as being symmetrical.

Answer 8

Time, t	Number of students	Cumulative frequency
$9.5 \leqslant t < 19.5$	2	2
$19.5 \leqslant t < 29.5$	4	6
$29.5 \leqslant t < 39.5$	8	14
$39.5 \leqslant t < 49.5$	11	25
$49.5 \leqslant t < 69.5$	5	30
$69.5 \leqslant t < 79.5$	2	32

(a)

For Q_2 : $\dfrac{32 \times 2}{4} = 16$ which is in the class 39.5 - 49.5



$$\frac{Q_2 - 39.5}{49.5 - 39.5} = \frac{16 - 14}{25 - 14} \Rightarrow Q_2 = 41.31818182...$$

$$\Rightarrow \textit{median} = 41.3 \text{ minutes}$$

(b)

$$\textit{mean} = \frac{1414}{32}$$

$$= 44.1875 \text{ minutes} \quad \text{(exact)}$$

That is, 44 minutes 11 seconds

$$\textit{standard deviation} = \sqrt{\frac{69\,378}{32} - 44.1875^2}$$

$$= 14.68 \text{ minutes}$$

That is, 14 minutes 41 seconds

(c)

$$\textit{mean} > \textit{median}$$

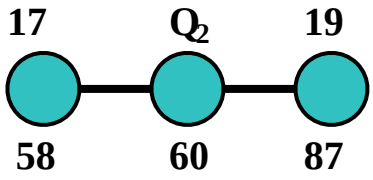
$\therefore$ positive skew

Answer 9

(a) No gaps in the table !

Foot length, l, (cm)	Number of children	Cumulative Frequency
$10 \leqslant l < 12$	5	5
$12 \leqslant l < 17$	53	58
$17 \leqslant l < 19$	29	87
$19 \leqslant l < 21$	15	102
$21 \leqslant l < 23$	11	113
$23 \leqslant l < 25$	7	120

For Q_2 : $\dfrac{120 \times 2}{4} = 60$ which is in the class 17 - 19



$$\frac{Q_2 - 17}{19 - 17} = \frac{60 - 58}{87 - 58} \Rightarrow Q_2 = 17.13793103 \Rightarrow \textit{median} = 17.1 \text{ cm}$$

(b)

Foot length, l, (cm)	Number of children, f	Mid-interval value, x	fx	x^2	$f x^2$
$10 \leqslant l < 12$	5	11	55	121	605
$12 \leqslant l < 17$	53	14.5	768.5	210.25	11143.25
$17 \leqslant l < 19$	29	18	522	324	9396
$19 \leqslant l < 21$	15	20	300	400	6000
$21 \leqslant l < 23$	11	22	242	484	5324
$23 \leqslant l < 25$	7	24	168	576	4032
	120		2055.5		36500.25

$$\textit{mean} = \frac{2055.5}{120}$$

$$= \frac{4111}{240} \quad \text{(exact)}$$

$$= 17.129\,166\,67...$$

$$= 17.1 \text{ cm}$$

$$\text{standard deviation} = \sqrt{\frac{36\,500.25}{120} - \left(\frac{4111}{240}\right)^2}$$
$$= 3.280\,304\,758...$$
$$= 3.28 \text{ cm}$$

(c)

$$\text{Coefficient of skewness} = \frac{3\,(\,\text{mean} - \text{median}\,)}{\text{standard deviation}}$$
$$= \frac{3 \times (\ 17.129\,166\,67 - 17.137\,931\,03\,)}{3.280\,304\,758}$$
$$= -0.008\,015\,438\,18$$

The skew is negligible.
This means that the distribution of the foot length data is symmetrical.

Notice the loss of significant figures due to the subtraction.
Students who use rounded off earlier answers can get very strange answers to this final part of the question.

7.1 Coding of Mean and Standard Deviation

Example

The number of pages in six randomly chosen novels was recorded as being;

The Shipping News	Annie Proulx	337
Deadly Décisions	Kathy Reichs	333
Birdsong	Sebastian Faulks	305
My Legendary GirlFriend	Mike Gayle	353
Enduring Love	Ian McEwan	249
Black Notice	Patricia Cornwell	413

(i) Calculate the Mean and Standard Deviation of this original data, μ_o and σ_o

	Data, x	x^2
	337	113569
	333	110889
	305	93025
	353	124609
	249	62001
	413	170569
Σ	1990	674662

$$\mu = \frac{\Sigma x}{f}$$ (Use the exact value in subsequent calculations)

$$\sigma = \sqrt{\frac{\Sigma x^2}{f} - \mu^2}$$

The idea of coding is to find a transformation of the form,

$$X_t = \frac{x_o - a}{b}$$

that can be applied to the original data.
In this equation

- x_o is the original data
- a and b are suitable constants
- X_t is the transformed data

To be useful, the transformed data should be easier to manipulate and, in particular, finding the mean and standard deviation of the transformed data μ_t and σ_t should be a relatively easy task.
These are then decoded to give the mean and standard deviation of the original data.

(ii) Code the example data using the transformation

$$X_t = \frac{x_o - 301}{4}$$

x_o	X_t	X_t^2
337		
333		
305		
353		
249		
413		
Σ		

(iii) For the coded (transformed) data find the mean and the standard deviation. That is, find μ_t and σ_t

(iv) Use the value of the the mean of the coded data, and the fact that it was coded using the coding

$$X_t = \frac{x_o - 301}{4}$$

to find the mean, μ_o, of the original data.

(v) Use the value of the the standard deviation of the coded data, and the fact that it for coded using the coding

$$X_t = \frac{x_o - 301}{4}$$

to find the standard deviation, σ_o, of the original data.

7.2 Summary

For data that has been coding using a transformation of the form,

$$X_t = \frac{x_o - a}{b}$$

and given the mean and standard deviation of the coded (transformed) data, μ_t and σ_t the original mean and standard deviation is retrieved using the facts that,

$$u_o = b\,\mu_t + a$$

$$\sigma_o = b\,\sigma_t$$

7.3 Exercise

Question 1

Data is coded using the transformation

$$y = \frac{x - 250}{80}$$

The mean of the coded data is 4.5

The standard deviation of the coded data is 3.2

Find the mean and the standard deviation of the original data.

Question 2

The annual income, *I*, of 100 IT consultants was recorded.

The data were coded using

$$y = \frac{I - 400}{10\,000}$$

and the following summations were obtained;

$$\Sigma y = 887 \qquad \Sigma y^2 = 8065$$

(i) Calculate the mean and the standard deviation of the coded data.

(ii) Calculate the mean and the standard deviation of the original data.

(iii) Most incomes will be within two standard deviations of the mean.
Calculate $\mu \pm 2 \times \sigma$ and so state the values between which most incomes lie.

Question 3

Mrs Crump gathers information about the times taken by 29 girls to get from Emma Darwin Hall to Kingsland Hall for breakfast one morning.

Time minutes	N° girls f	Mid-Interval minutes t	Coding $y = (t - 7) / 3$ y	y^2	fy	fy^2
$2 \leqslant t < 6$	5					
$6 \leqslant t < 8$	14					
$8 \leqslant t < 12$	8					
$12 \leqslant t < 20$	2					
		Σ				

(i) Fill in the empty cells on the table above, where the data is coded using

$$y = \frac{t - 7}{3}$$

(ii) Use Σfy to find mean and Σfy^2 to find standard deviation of the coded data.

(iii) Find the mean and standard deviation of the original data.

Question 4

S1 Examination Question from January 2012 Q4

The marks, x, of 45 students randomly selected from those students who sat a mathematics examination are shown below.

36	39	39	40	41	42	42	43	44	45
46	46	46	48	50	52	53	53	54	54
55	55	56	57	57	59	60	60	60	60
61	63	64	64	64	65	65	66	67	68
69	71	72	73	73					

(a) Write down the modal mark of these students.

[1 mark]

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

[3 marks]

For these students $\Sigma x = 2497$ and $\Sigma x^2 = 143\,369$

(c) Find the mean and the standard deviation of the marks of these students.

[3 marks]

(d) Describe the skewness of the marks of these students, giving a reason for your answer.

[2 marks]

The mean and the standard deviation of the marks of all the students who sat the examination where 55 and 10 respectively. The examiners decided that the total mark of each student should be scaled by subtracting 5 marks and then reducing the mark by a further 10 %.

(e) Find the mean and the standard deviation of the scaled marks of all the students.

[4 marks]

7.4 Answers

7.4.1 Solutions (Introductory Example)

(i) $\mu_o = \dfrac{\Sigma x}{f} = \dfrac{1990}{6} = \dfrac{995}{3} = 331\dfrac{2}{3}$

$$\sigma_o = \sqrt{\frac{\Sigma x^2}{f} - \mu^2} = \sqrt{\frac{674662}{6} - \left(\frac{995}{3}\right)^2} = 49.4$$

(ii)

x_o	X_t	X_t^2
337	9	81
333	8	64
305	1	1
353	13	169
249	− 13	169
413	28	784
Σ	46	1268

(iii)

$$\mu_o = \frac{\Sigma x}{f} = \frac{46}{6} = \frac{23}{3} = 7\frac{2}{3}$$

$$\sigma_o = \sqrt{\frac{\Sigma x^2}{f} - \mu^2} = \sqrt{\frac{1268}{6} - \left(\frac{23}{3}\right)^2} = 12.4$$

(iv)

$$\mu_o = 4\mu_t + 301$$
$$= 4 \times \frac{23}{3} + 301$$
$$= 331\frac{2}{3}$$

(v)

$$\sigma_o = 4\sigma_t$$
$$= 4 \times 12.4$$
$$= 49.4$$

7.4.2 Solutions to the Exercise (7.3 Exercise)

Answer 1

The original mean, $\mu_o = 610$

The original standard deviation, $\sigma_o = 256$

Answer 2

(i) $\mu_t = 8.87$ and $\sigma_t = 1.4046$

(ii) $\mu_o =$ £ 89 100 and $\sigma_o =$ £ 14 046

(iii) £ 61 008 and £ 117 192

Answer 3

(i)

Time minutes	N° girls	Mid-Interval minutes	Coding $y = (t - 7)/3$			
	f	t	y	y^2	fy	fy^2
$2 \leqslant t < 6$	5	4	–1	1	-5	5
$6 \leqslant t < 8$	14	7	0	0	0	0
$8 \leqslant t < 12$	8	10	1	1	8	8
$12 \leqslant t < 20$	2	16	3	9	6	18
		Σ	3	11	9	31

(ii) $\mu_{coded} = \frac{9}{29}$ (about 0.31) and $\sigma_{coded} = 0.986$

(iii) $\mu_{original} = 7.93$ and $\sigma_{original} = 2.93$

Answer 4

(a) 60

(b) $Q_1 = 46$ $Q_2 = 56$ $Q_3 = 64$

(c) mean is 55.48... or $\frac{2497}{45}$

standard deviation is

$$\sqrt{\frac{143369}{45} - \left(\frac{2497}{45}\right)^2}$$

$= 10.342...$ if rounded off mean used

$= 10.459..,$ if exact mean used

(d) skew is negative

reason requested may include

- $Q_2 - Q_1 > Q_3 - Q_2$ perhaps with part (b) numbers
- median closer to upper quartile than lower quartile
- mean < median < mode
- (mean-median) / standard deviation < 0

(e) scaled mean = 45, scaled standard deviation = 9

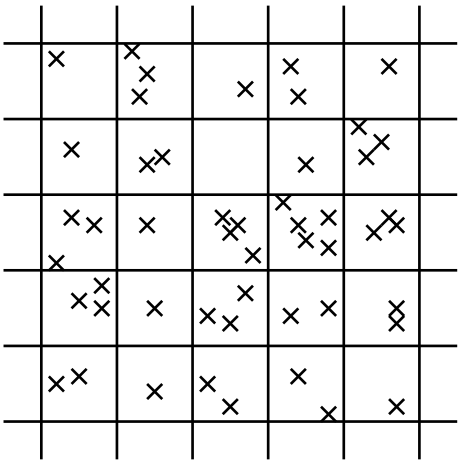
Lesson 8

A-level Statistics : Year 1
Partitioning Data

8.1 Partitioning Data : Topic Summary

Question 1

A Biologist has divided up an area of a field into 25 x 1 m^2 squares.
In each square he records the occurrence of a certain species of spider.
His results are shown below where each × indicates a spider found.



(i) Complete the table below to show how many times each Spider Count occurred;

Spider Count	0	1	2	3	4	5	6
Squares							

[4 marks]

(ii) Determine the mean number of spiders per m^2

[2 marks]

(iii) Find the standard deviation of spiders per m^2

[4 marks]

(iv) Any item of data that is more than two standard deviations away from the mean is considered to be an outlier.
On this basis, would you classify any item of the data as an outlier ?

[2 marks]

Question 2

The heights of 2000 adult males were measured and are tabulated below;

Height (cm)	160-164	165-169	170-174	175-179	180-184	185-189	190-194	195-199
Frequency	11	49	190	488	632	470	137	23

Calculate an estimate of the median of this data.
(Remember to use linear interpolation)

[6 marks]

Question 3

Some data is coded using

$$y = \frac{x - 14.5}{10}$$

For the coded data, the mean is 0.69 and the standard deviation 1.24

Determine the mean and the standard deviation for the original data.

[3 marks]

Question 4

The number of errors, x, on each of 50 mathematical Lesson tests was monitored. The results when summarised showed that:

$$\sum x = 230 \qquad \sum x^2 = 1258$$

Calculate the mean and the standard deviation of the number of errors per mathematical Lesson test.

[4 marks]

Question 5

The data below shows the number of letters received by a business each day for 5 weeks. (Sundays excluded)

02	06	07	07	08	08	13	14	15	15
15	17	17	19	21	22	23	25	28	28
32	33	35	36	36	36	41	41	42	43

(i) Calculate;

(a) Q_1 the lower quartile

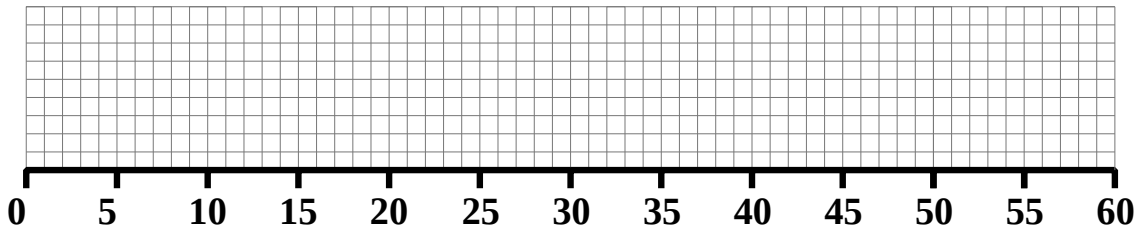
(b) Q_2 the median

(c) Q_3 the upper quartile

(d) IQR the inter quartile range

[7 marks]

(ii) On graph paper draw a box and whisker plot for the data.



[4 marks]

(iii) Comment on the skewness of the distribution.

[2 marks]

Question 6

Explain with examples the difference between Quantitative and Qualitative data.

[2 marks]

8.2 Answers

A-level Statistics : Year 1
Partitioning Data

Answer 1

(i)

Spider Count	0	1	2	3	4	5	6
Squares	1	9	7	6	1	1	0

(ii)

$$mean = \frac{50}{25}$$

$$= 2 \text{ spiders per m}^2$$

(iii)

$$standard\ deviation = \sqrt{\frac{132}{25} - 2^2}$$

$$= 1.13 \text{ spiders per m}^2$$

(iv)

$$mean + 2 \times standard\ deviation = 2 + 2 \times 1.13$$

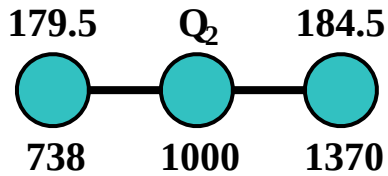
$$= 4.26$$

$\therefore$ The square with 5 spiders is an outlier

Answer 2

Height (cm)	160-164	165-169	170-174	175-179	180-184	185-189	190-194	195-199
Frequency	11	49	190	488	632	470	137	23
Cumulative	11	60	250	738	1370	1840	1977	2000

For Q_2 : $\frac{2000 \times 2}{4} = 1000$ which is in the class 175 - 179



$$\frac{Q_2 - 179.5}{184.5 - 179.5} = \frac{1000 - 738}{1370 - 738}$$

$$Q_2 = 181.5727848$$

$$median = 181.6 \text{ cm}$$

Answer 3

$$x_\mu = 10\,y_\mu + 14.5$$
$$= 10 \times 0.69 + 14.5$$
$$= 21.4$$

$$x_\sigma = 10\,y_\sigma$$
$$= 10 \times 1.24$$
$$= 12.4$$

Answer 4

$$mean = \frac{230}{50}$$
$$= 4.6 \text{ (exact)}$$

$$standard\ deviation = \sqrt{\frac{1258}{50} - 4.6^2}$$
$$= 2$$

Answer 5

(i) **(a)**

$$\text{For } Q_1 : \frac{30 \times 1}{4} = 7.5 \Rightarrow \text{ARU} \Rightarrow 8^{th} \text{ which gives } 14$$

(b)

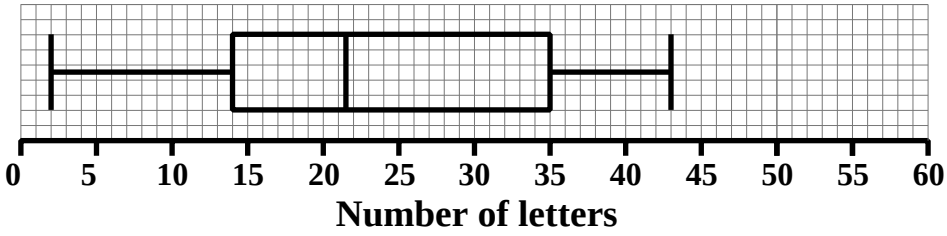
$$\text{For } Q_2 : \frac{30 \times 2}{4} = 15 \Rightarrow \text{Integer}$$
$$\Rightarrow \frac{15^{th} + 16^{th}}{2} = \frac{21 + 22}{2} \text{ which gives } 21.5$$

(c)

$$\text{For } Q_3 : \frac{30 \times 3}{4} = 22.5 \Rightarrow \text{ARU} \Rightarrow 23^{rd} \text{ which gives } 35$$

(d) $35 - 14 = 21$

(ii)



(iii)

$$Q_3 - Q_2 = 35 - 21.5 = 13.5$$

$$Q_2 - Q_1 = 21.5 - 14 = 7.5$$

$$\therefore\ Q_3 - Q_2 > Q_2 - Q_1 \implies \text{Positive skew}$$

Answer 6

Each item of quantitative data has a numerical value associated with it and so can be ordered according to that numerical value.
For example: The heights of adult males in the UK

Each item of qualitative data has no obvious numerical value associated with it and so there is no sense of relative importance or natural ordering.
For example: The colour of a person's eyes.

9.1 Partitioning Data : Homework (Topic Summary)

Question 1

A-Level Examination Question from June 2014, S1, Q2

The mark x, scored by each student who sat a statistics examination is coded using

$$y = 1.4x - 20$$

The coded marks have mean 60.8 and standard deviation 6.60

Find the mean and the standard deviation of x

[4 marks]

Question 2

A-Level Examination Question from June 2016, S1, Q3 (a)

Before going on holiday to Seapron, Tania records the weekly rainfall (x mm) at Seapron for 8 weeks during the summer. Her results are summarised as

$$\sum x = 86.8 \qquad \sum x^2 = 985.88$$

Find the standard deviation, σ_x, for these data.

[3 marks]

Question 3

A-Level Examination Question from June 2018, Applied, Q4

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 mark]

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

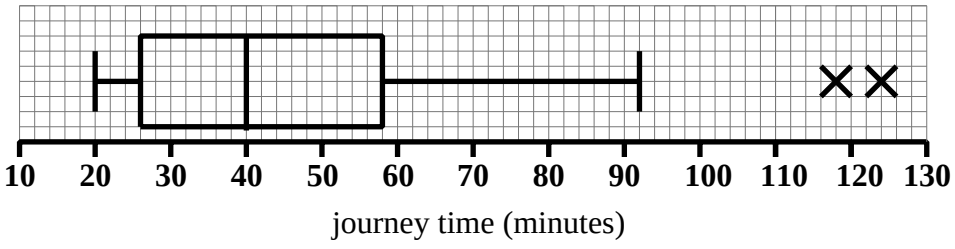
[2 marks]

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 mark]

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



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

(d) Write down the interquartile range for these data

[1 mark]

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

[3 marks]

(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 marks]

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 marks]

Question 4

AS-Level Examination Question from June 2018, Applied, Q4 (a) & (b)

Helen is studying the daily mean wind speed for Camborne using the large data set from 1987. The data for one month is summarised in Table 1 below.

Wind speed	n/a	6	7	8	9	11	12	13	14	16
Frequency	13	2	3	2	2	3	1	2	1	2

(a) Calculate the mean for these data.

[1 mark]

(b) Calculate the standard deviation for these data and state the units.

[2 marks]

9.2 Answers

Answer 1

$$x_{\mu} = \frac{60.8 + 20}{1.4}$$
$$= \frac{404}{7}$$
$$= 57.71 \quad \text{(mean)}$$

$$x_{\sigma} = \frac{6.60}{1.4}$$
$$= \frac{33}{7}$$
$$= 4.71 \quad \text{(standard deviation)}$$

Answer 2

$$\textit{standard deviation} = \sqrt{\frac{\sum x^2}{\sum f} - \left(\frac{\sum x}{\sum f}\right)^2}$$
$$= \sqrt{\frac{985.88}{8} - \left(\frac{86.8}{8}\right)^2}$$
$$= 2.347\,871\,376...$$
$$= 2.35 \text{ mm}$$

Answer 3

(a) Convenience sampling, opportunity sampling

(b) Charlie would need more than ten minutes to gather data on 40 people.
He could still stand by the door for a longer period and do Quota sampling.
Ask 8 people every ten minutes between, say, 08:30 and 09:20

(c) Taruni is carrying out a Census

(d) $Q_3 = 58$ and $Q_1 = 26$ so IQR is 32 minutes

(e)

$$\textit{mean} = \frac{4133}{95}$$
$$= 43.505\,263\,16...$$
$$= 43.5 \text{ minutes}$$

$$\textit{standard deviation} = \sqrt{\frac{202\,294}{95} - \left(\frac{4133}{95}\right)^2}$$
$$= 15.38514231...$$
$$= 15.4 \text{ minutes}$$

(f) From the box and whisker plot it can be seen that there are a couple of large value outliers and these will distort the values of the mean and the standard deviation. Therefore, the median and interquartile range will better reflect the average and spread of the majority of the data.

(g) The median will need to be decreased because two values that were above the previous median of 40 are now below it.
Similarily, the upper quartile will need to de decreased because two values that were above the previous upper quartile of 58 are now below it.
All else on the box plot remains the same.

Answer 4
See the notes following the following answer about what the Examiners were expecting in answer to this question.

Windspeed, w (knots)	Frequency f	Mid-interval value, x	fx	x^2	fx^2
$5.5 \leqslant w < 6.5$	2	6	12	36	72
$6.5 \leqslant w < 7.5$	3	7	21	49	147
$7.5 \leqslant w < 8.5$	2	8	16	64	128
$8.5 \leqslant w < 9.5$	2	9	18	81	162
$9.5 \leqslant w < 10.5$	0	10	0	100	0
$10.5 \leqslant w < 11.5$	3	11	33	121	363
$11.5 \leqslant w < 12.5$	1	12	12	144	144
$12.5 \leqslant w < 13.5$	2	13	26	169	338
$13.5 \leqslant w < 14.5$	1	14	14	196	196
$14.5 \leqslant w < 15.5$	0	15	0	225	0
$15.5 \leqslant w < 16.5$	2	16	32	256	512
	18		184		2062

(a)

$$\begin{aligned} mean &= \frac{184}{18} \\ &= 10.222... \\ &= 10.2 \end{aligned}$$

(b)

$$\begin{aligned} standard\ deviation &= \sqrt{\frac{2062}{18} - \left(\frac{184}{18}\right)^2} \\ &= 3.172... \text{ knots} \end{aligned}$$

Note : The examiners report for this question says,

"Another skill students are required to have under the new specifications is to be able to calculate standard statistics on their calculators.
In this case we expected students to enter the data into their calculators and write down the mean (part a) and standard deviation (part (b).
It was clear many students are not familiar with this requirement and we often saw several lines of working to derive these results which unfortunately, under the new specifications cannot receive any marks"

Lesson 10

A-level Statistics : Year 1
Partitioning Data

10.1 Histograms #1

A Histogram is **NOT** the same as a Bar Chart.

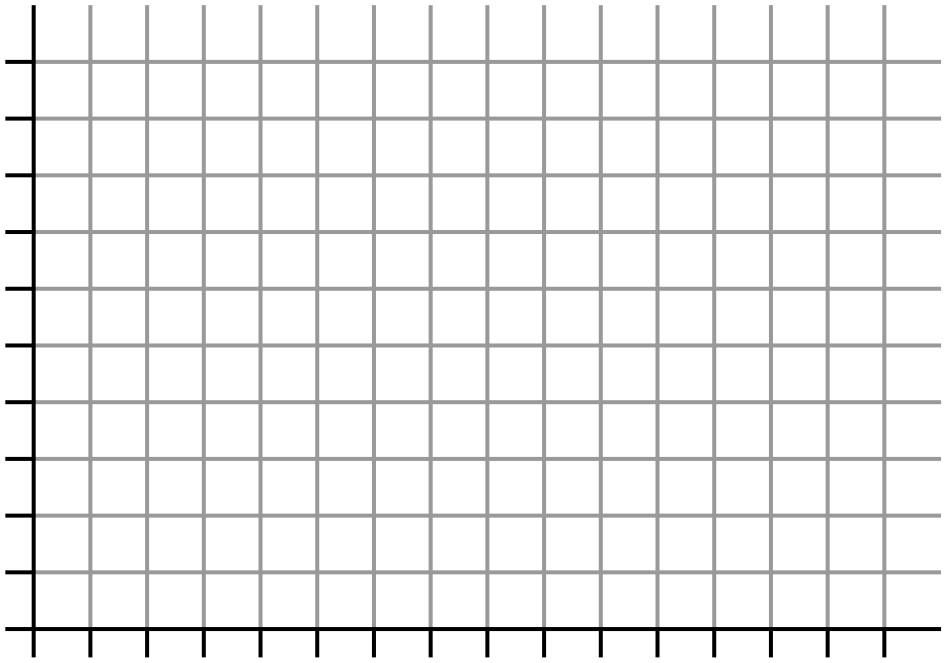
Bar Chart : ***Height*** is directly proportional to frequency
Histogram : ***Area*** is directly proportional frequency

10.2 Example

At a call centre the times taken to deal with 1000 calls were as follows:

Duration of call (minutes)	Number of calls Frequency = Area	Width	Height
$0 \leqslant t < 2$	200		
$2 \leqslant t < 6$	300		
$6 \leqslant t < 10$	200		
$10 \leqslant t < 30$	300		

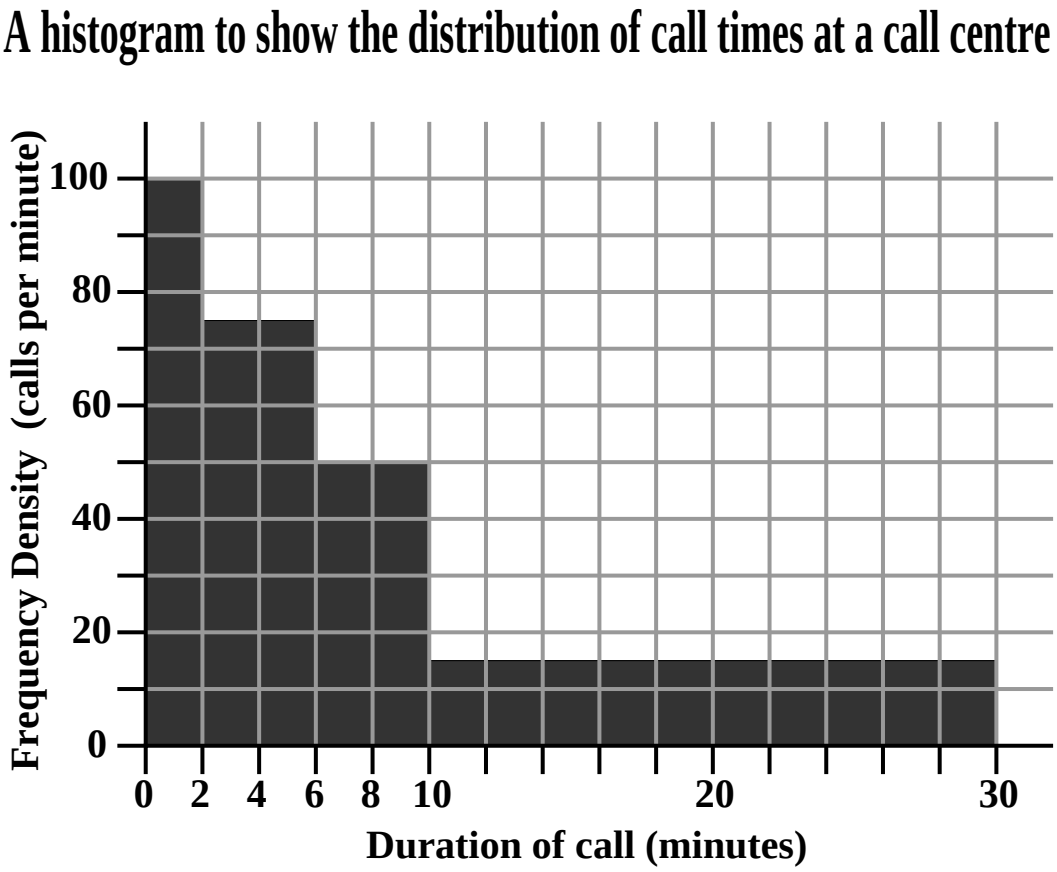
Plot a histogram to show the distribution of the 1000 call times at the centre.
Be sure to number and label the axes and give the histogram a title.



10.3 Solution to the example

Duration of call (minutes)	Number of calls Frequency = Area	Width	Height
$0 \leqslant t < 2$	200	2	100
$2 \leqslant t < 6$	300	4	75
$6 \leqslant t < 10$	200	4	50
$10 \leqslant t < 30$	300	20	15

[2 marks]



Suitable title **[1 mark]**
An *x*-axis that is numbered, labelled, and with the units clearly stated **[1 mark]**
A *y*-axis that is numbered, labelled, and with the units clearly stated **[1 mark]**
For each bar that is of the correct width and height **[3 marks]**

Total : 8 marks

10.4 Exercise

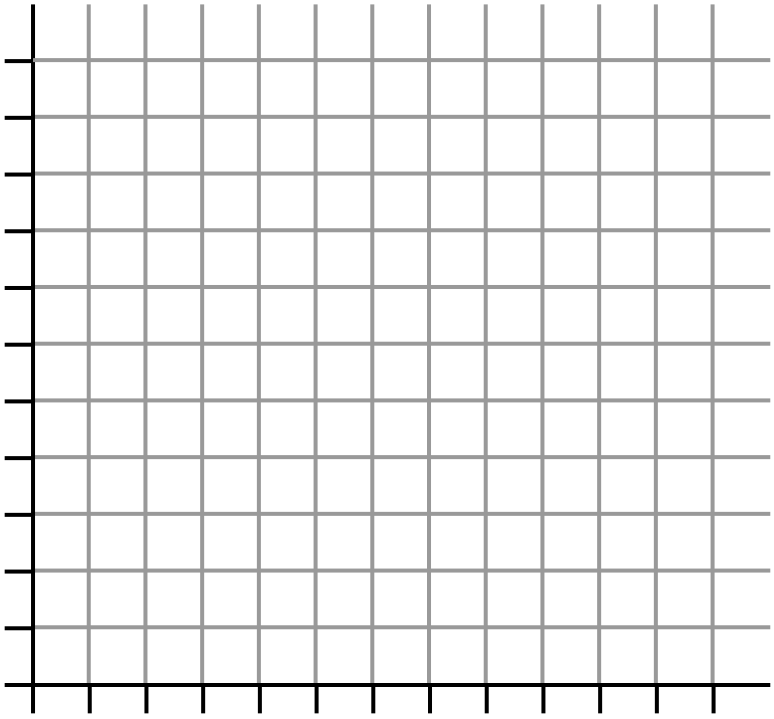
Question 1

In a government survey, a questionnaire is emailed to 15000 people who have previously agreed to take part.
The time taken, in hours, to return the questionnaire is logged.
The following table presents a summary.

Response time (hours)	Number of questionnaires Frequency = Area	Width	Height
$0 \leqslant h < 2$	700		
$2 \leqslant h < 4$	1100		
$4 \leqslant h < 6$	2200		
$6 \leqslant h < 8$	1700		
$8 \leqslant h < 15$	3500		
$15 \leqslant h < 24$	2700		

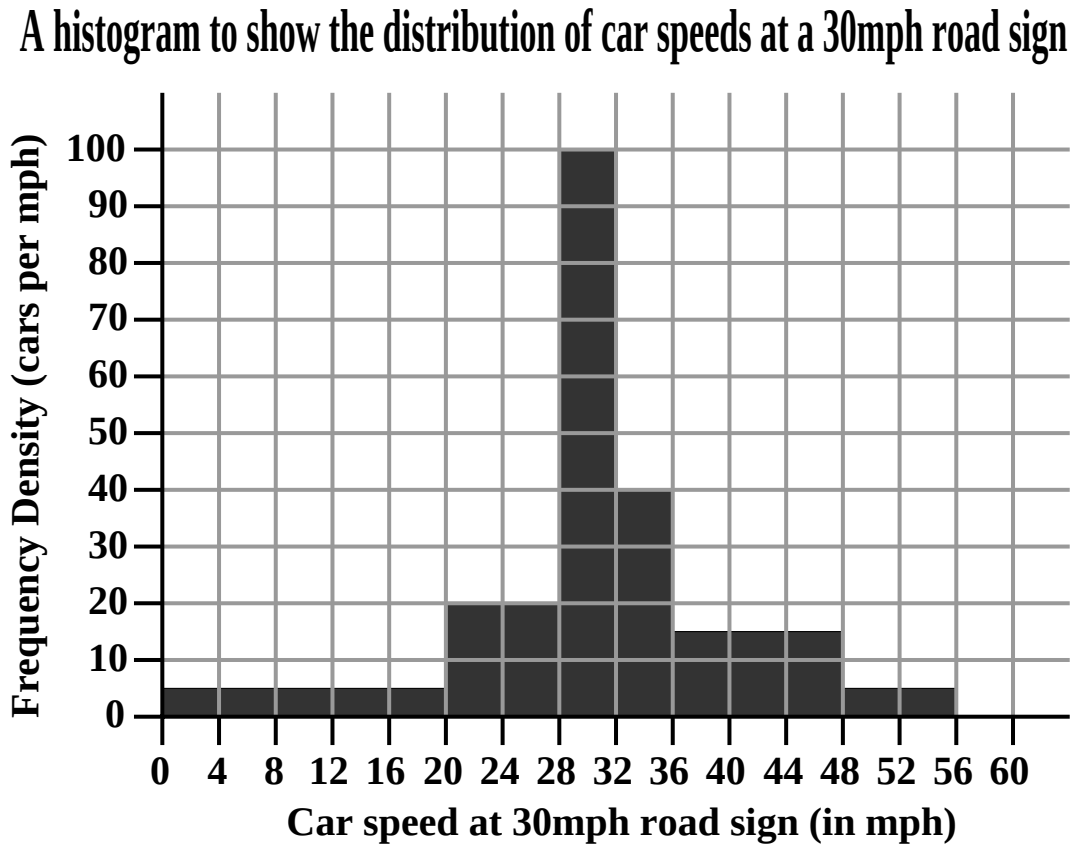
(i) How many questionnaires have not generated a response ?

(ii) Plot a histogram to show the distribution of the times taken in responding to the questionnaire. Complete the columns headed Width and Height in the table above, to help you do this.



Question 2

In trying to decide if a speed camera is required in the town of Numberville, a consultant provides the following histogram which shows the distribution of car speeds as cars pass the 30mph sign heading into Numberville.



(i) Use the histogram to complete the following table :

Car speed (miles per hour, mph)	Number of cars Frequency = Area	Width	Height
$0 \leqslant m < 20$			
$20 \leqslant m < 28$			
$28 \leqslant m < 32$			
$32 \leqslant m < 36$			
$36 \leqslant m < 48$			
$48 \leqslant m < 56$			

(ii) How many cars in total had their speed recorded entering Numberville ?

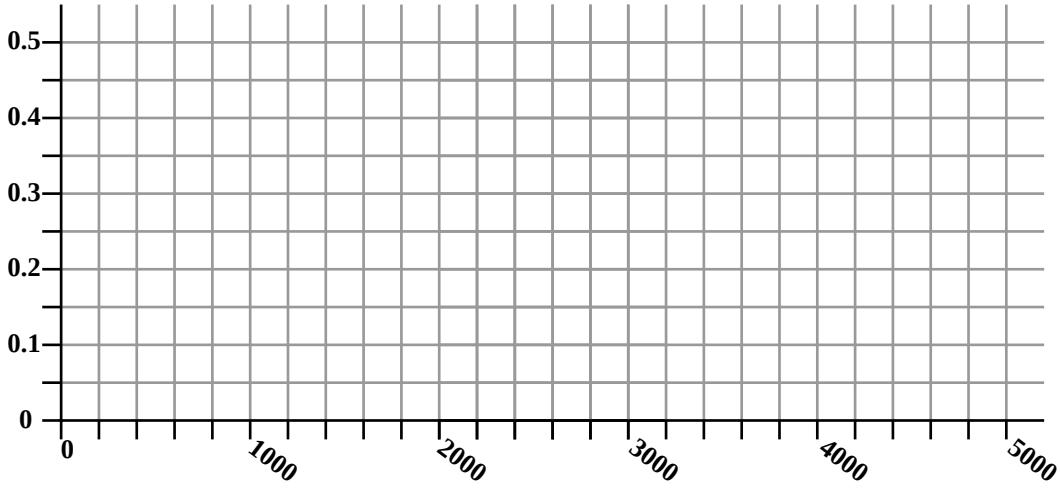
(iii) What percentage of cars were travelling within 2mph of the speed limit ?

Question 3

As part of a quality control test, 1100 light bulbs are left on continuously and the time taken before they fail is recorded.
The results are summarised in the table below.

Bulb lifetime (in hours)	Number of bulbs Frequency	Width	Height
$0 \leqslant t < 100$	50		
$100 \leqslant t < 1000$	180		
$1000 \leqslant t < 2000$	100		
$2000 \leqslant t < 4000$	500		
$4000 \leqslant t < 4500$	250		
$4500 \leqslant t < 4600$	20		

Plot a histogram to show the distribution of the life-times of the bulbs. Complete the columns headed Width and Height in the table above, to help you do this.



Be sure to number and label the axes and give the histogram a title.

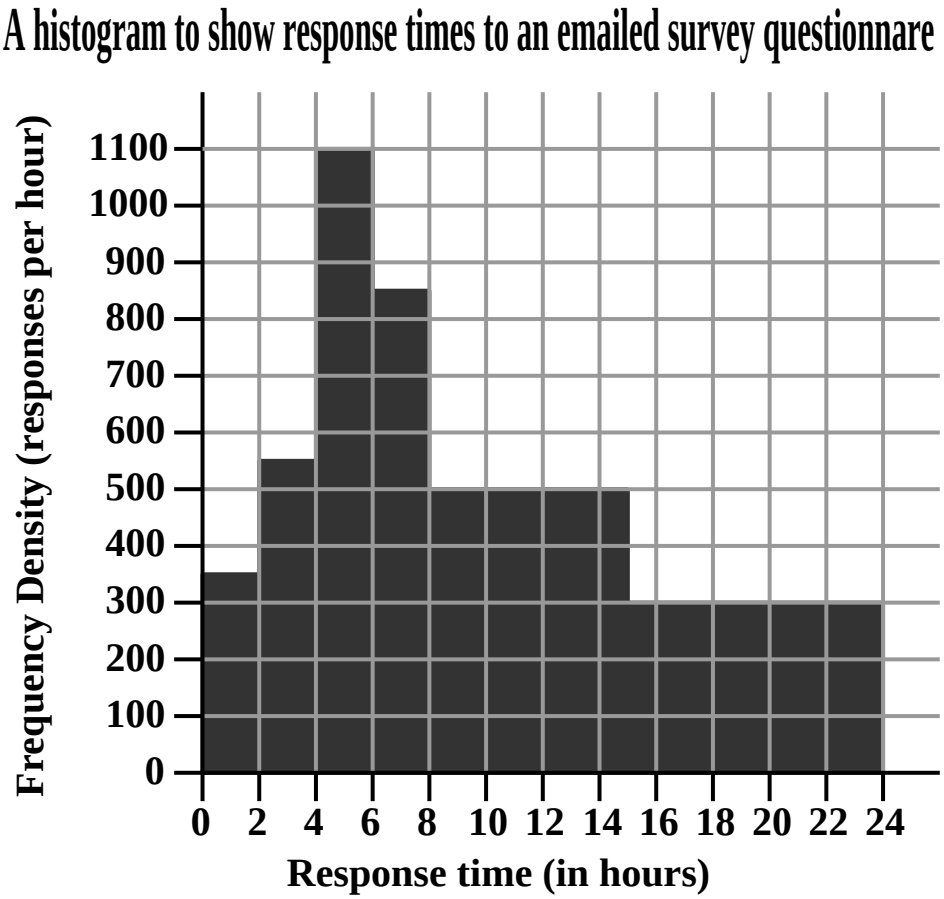
10.5 Solutions (10.4 Exercise)

A-level Statistics : Year 1
Partitioning Data

Answer 1
(i) 3100
(ii)

Response time (hours)	Number of questionnaires Frequency = Area	Width	Height
$0 \leqslant h < 2$	700	2	350
$2 \leqslant h < 4$	1100	2	550
$4 \leqslant h < 6$	2200	2	1100
$6 \leqslant h < 8$	1700	2	850
$8 \leqslant h < 15$	3500	7	500
$15 \leqslant h < 24$	2700	9	300

[2 marks]



Suitable title **[1 mark]**
An *x*-axis that is numbered, labelled, and with the units clearly stated **[1 mark]**
A *y*-axis that is numbered, labelled, and with the units clearly stated **[1 mark]**
For each bar that is of the correct width and height. **[3 marks]**

Total : 8 marks

Answer 2

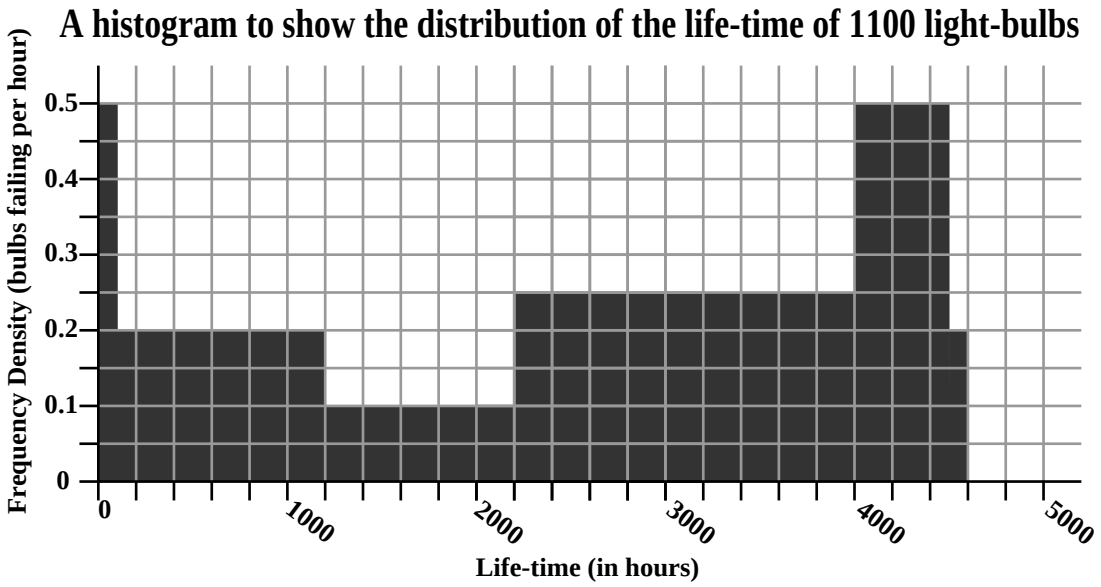
Car speed (miles per hour, mph)	Number of cars Frequency = Area	Width	Height
$0 \leqslant m < 20$	100	20	5
$20 \leqslant m < 28$	160	8	20
$28 \leqslant m < 32$	400	4	100
$32 \leqslant m < 36$	160	4	40
$36 \leqslant m < 48$	180	12	15
$48 \leqslant m < 56$	40	8	5

(ii) 1040 cars

(iii) $\frac{400}{1040} \times 100 = 38.5\%$

Answer 3

Bulb lifetime (in hours)	Number of bulbs Frequency	Width	Height
$0 \leqslant t < 100$	50	100	0.5
$100 \leqslant t < 1000$	180	900	0.2
$1000 \leqslant t < 2000$	100	1000	0.1
$2000 \leqslant t < 4000$	500	2000	0.25
$4000 \leqslant t < 4500$	250	500	0.5
$4500 \leqslant t < 4600$	20	100	0.2



11.1 Histograms #2

Having looked at some simple histograms in Lesson 10, the moment has come to look at some of the more subtle aspects of their underlying mathematics. Firstly, note that only continuous data should be represented by a histogram. Secondly, recall that the key idea behind a histogram is that Area is directly proportional to frequency.

$$Area \propto Frequency$$

$$\therefore\ Area = k \times Frequency$$

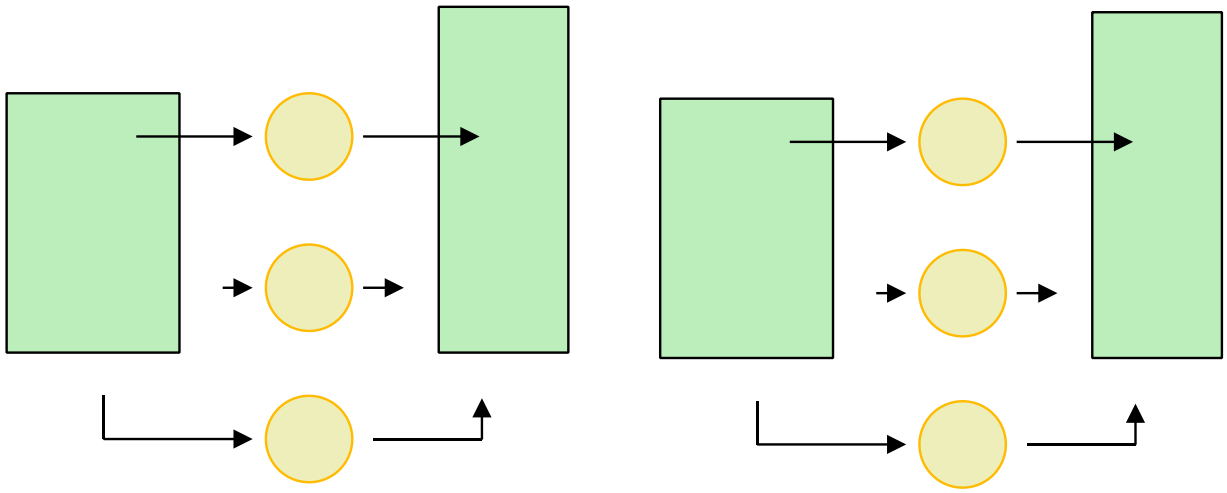
The constant of this proportionality need not be 1 which can give rise to some quite tricky problems.

11.2 Example

The number of hours of sunshine each day, h, over a seven week period in the summer was recorded at Heathrow Airport;

Hours	$0 \leqslant h < 5$	$5 \leqslant h < 8$	$8 \leqslant h < 11$	$11 \leqslant h < 12$	$12 \leqslant h < 14$
Frequency	18	9	12	6	4

A histogram was drawn to represent these data.
The $8 \leqslant h < 11$ group was represented by a bar of width 1.5 cm and height 12 cm.
Find the width and height of the $0 \leqslant h < 5$ group.



11.3 Exercise

Question 1

The DMG (daily maximum gust) is measured (in knots) at Hurn (Bournmouth Airport) throughout May and June 2019.

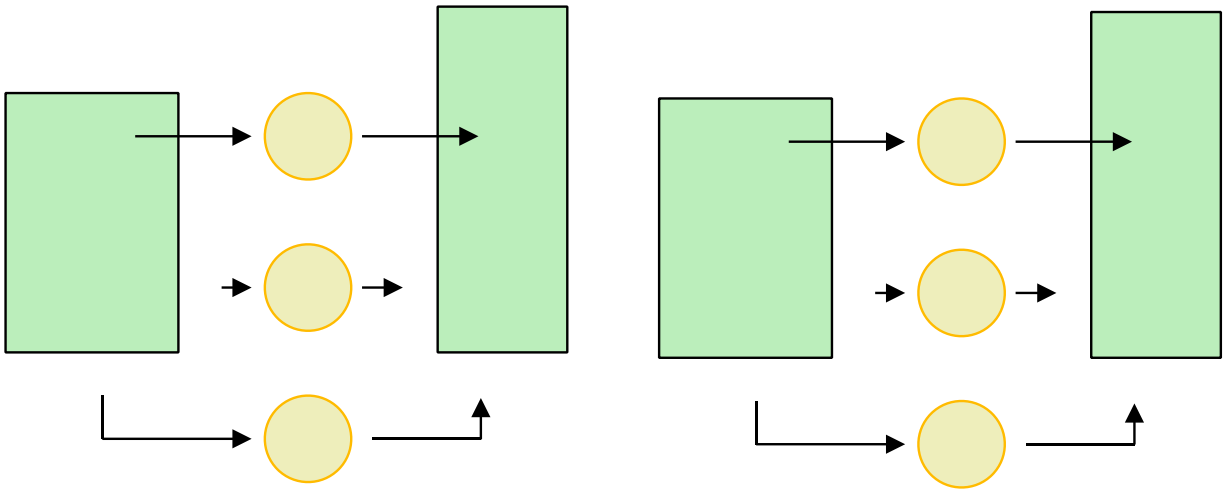
The data is summarised in the table;

DMG, g	$10 \leqslant g < 15$	$15 \leqslant g < 18$	$18 \leqslant g < 20$	$20 \leqslant g < 25$	$25 \leqslant g < 50$
Frequency	3	9	9	20	16

Marta draws a histogram to represent this data.

She draws the bar representing the $10 \leqslant g < 15$ class is 2.5 cm wide and 1.8 cm high.

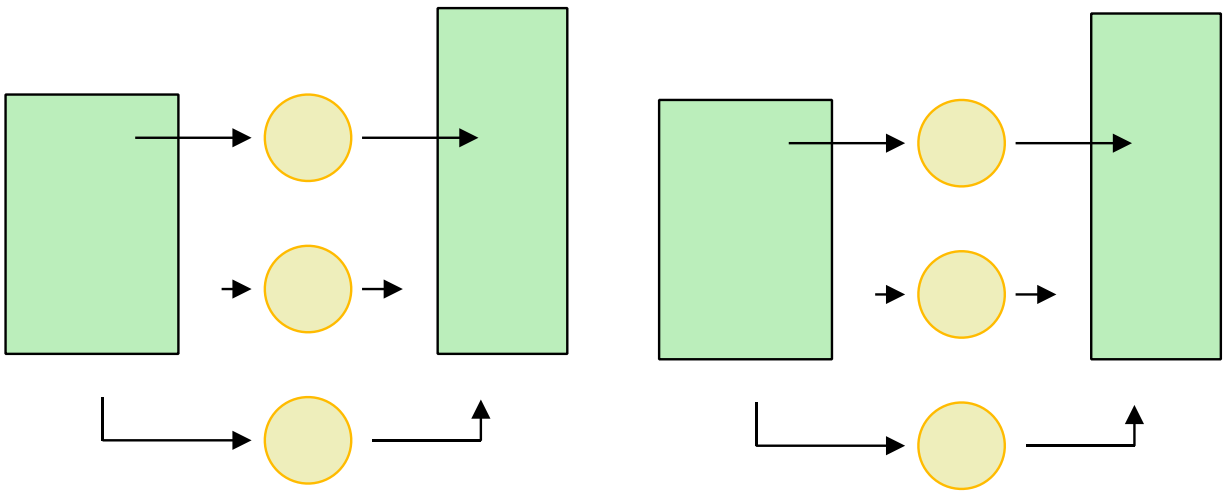
(i) Calculate the width and height of the bar representing the $18 \leqslant g < 20$ class.



Cathy also draws a histogram to represent the same data.

She draws the bar representing the $10 \leqslant g < 15$ class is 2 cm wide and 3 cm high.

(ii) Calculate the width and height of the bar representing the $18 \leqslant g < 20$ class.



Question 2

The lifetime in thousand hours, L, of 36 LED lightbulbs is given in the following table;

Lifetime (1000 Hours)	$0 \leqslant L < 45$	$45 \leqslant L < 55$	$55 \leqslant L < 80$
Frequency	9	12	15

A histogram was drawn to represent these data.

The $45 \leqslant L < 55$ group was represented by a bar of width 2 cm and height 9 cm.

(i) Find the width and height of the $55 \leqslant L < 80$ group.

(ii) Find the width and height of the $0 \leqslant L < 45$ group.

Question 3

A survey of 100 households in Shrewsbury, Autumn 2019, gave the following results for weekly income £y

Income y (£)	Frequency f
$0 \leqslant y < 400$	12
$400 \leqslant y < 480$	28
$480 \leqslant y < 640$	22
$640 \leqslant y < 800$	18
$800 \leqslant y < 1200$	12
$1200 \leqslant y < 1600$	8

(i) An IT consultant who took part in the survey has an annual salary of £54000.
In which class interval was her income placed ?

A histogram represented $400 \leqslant y < 480$ by a rectangle of width 2cm and height 7 cm

(ii) Calculate the width and the height of the rectangle representing $640 \leqslant y < 800$

(iii) Use linear interpolation to determine the median annual income of a household in Shrewsbury, Autumn 2019.
Assume 52.143 weeks in a year, and give your answer to the nearest pound.

(iv) Toby claims that a histogram should not be drawn for this data.
Do you agree with him ?
Justify your answer.

Question 4

S1 A-level Examination Question from June 2017, Q2 (edited)

An estate agent is studying the cost of office space in London.

He takes a random sample of 90 offices and calculates the cost, £x per square foot.

His results are given in the table below.

Cost (£x)	Frequency	Midpoint (£y)
$20 \leqslant x < 40$	12	30
$40 \leqslant x < 45$	13	42.5
$45 \leqslant x < 50$	25	47.5
$50 \leqslant x < 60$	32	55
$60 \leqslant x < 80$	8	70

The discrete data (money) is modelled as continuous data and a histogram is drawn.

The bar representing $50 \leqslant x < 60$ is 2 cm wide and 8 cm high.

(a) Calculate the width and height of the bar representing $20 \leqslant x < 40$

[3 marks]

(b) Use linear interpolation to estimate the median cost.

[2 marks]

(c) Estime the mean cost of office space for these data.

[2 marks]

(d) Estimate the standard deviation for these data.
You may use the fact that,

$$\sum fy^2 = 226\,687.5$$

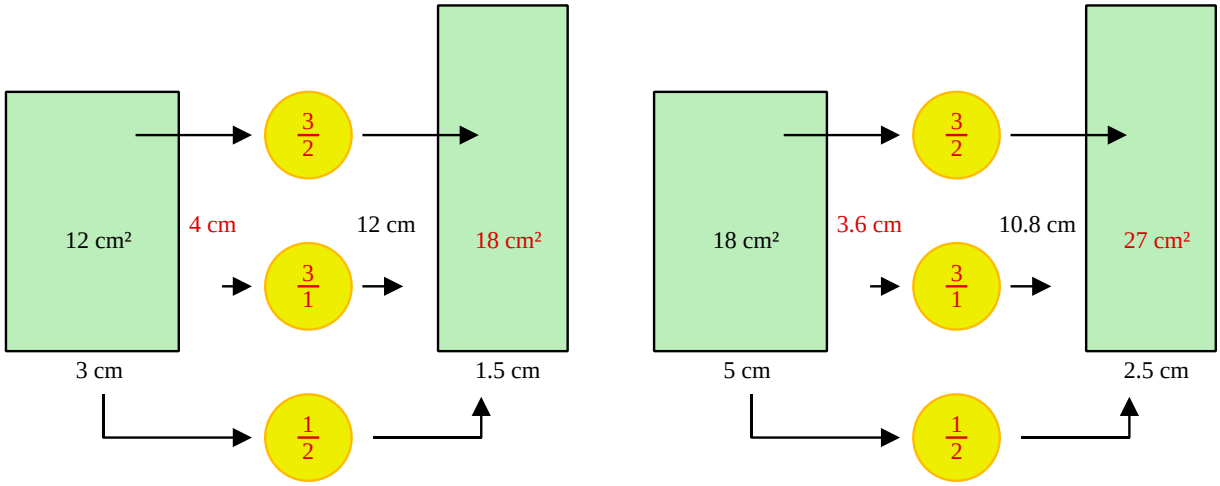
(e) Describe, giving a reason, the skewness.

[1 mark]

11.4 Answers

11.4.1 Solution to 11.2 Example

Complete the diagram with Area Scale Factor, separate Length Scale Factors for the x and y axis and then work through the various numbers for the lengths and areas.

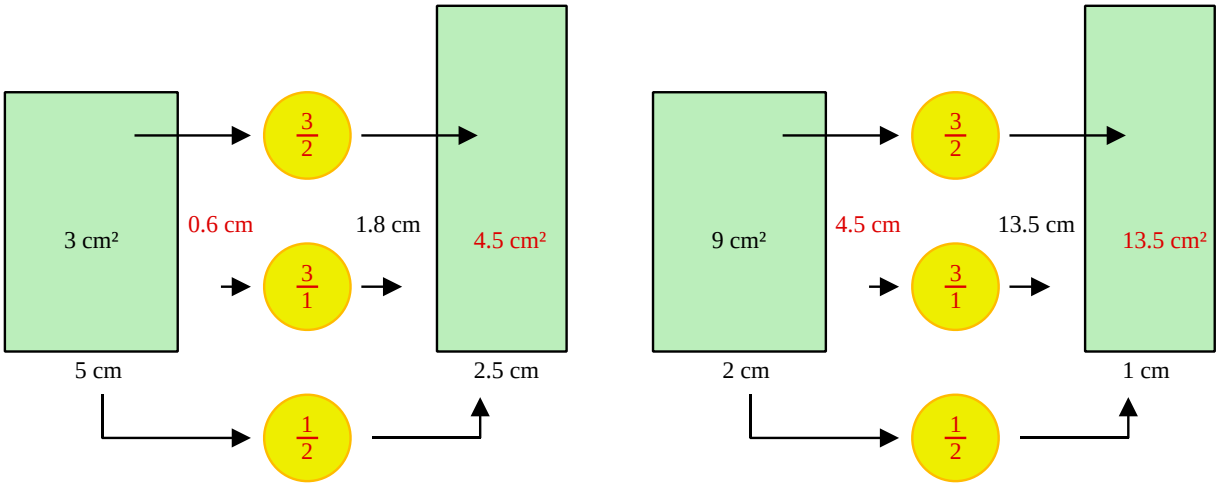


Thus, the $0 \leqslant h < 5$ group is represented by a bar of width 2.5 cm and height 10.8 cm

11.4.2 Solutions to 11.3 Exercise

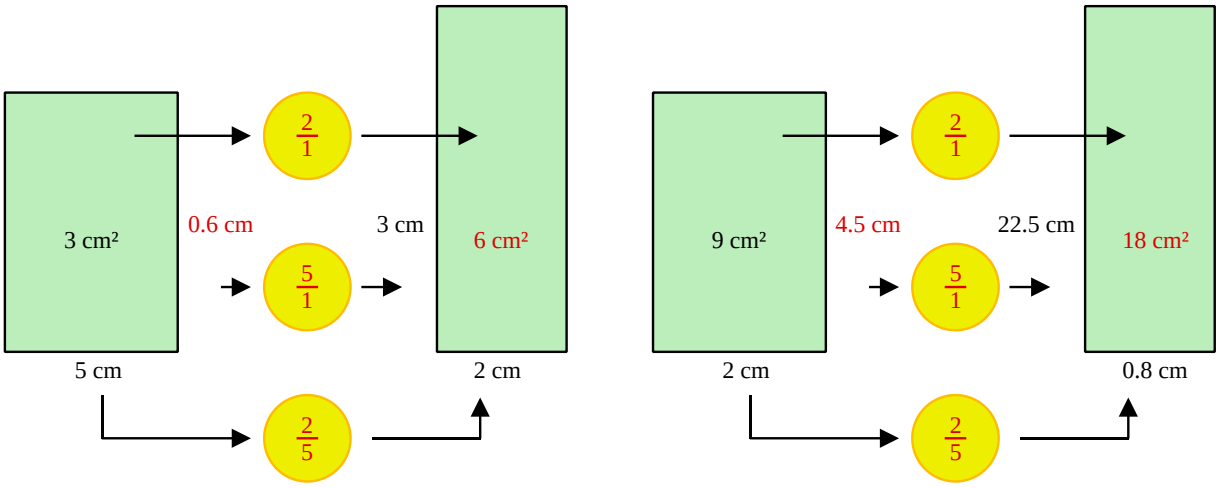
Answer 1

(i)



Marta's $18 \leqslant g < 20$ class is represented by a bar of width 1 cm and height 13.5 cm

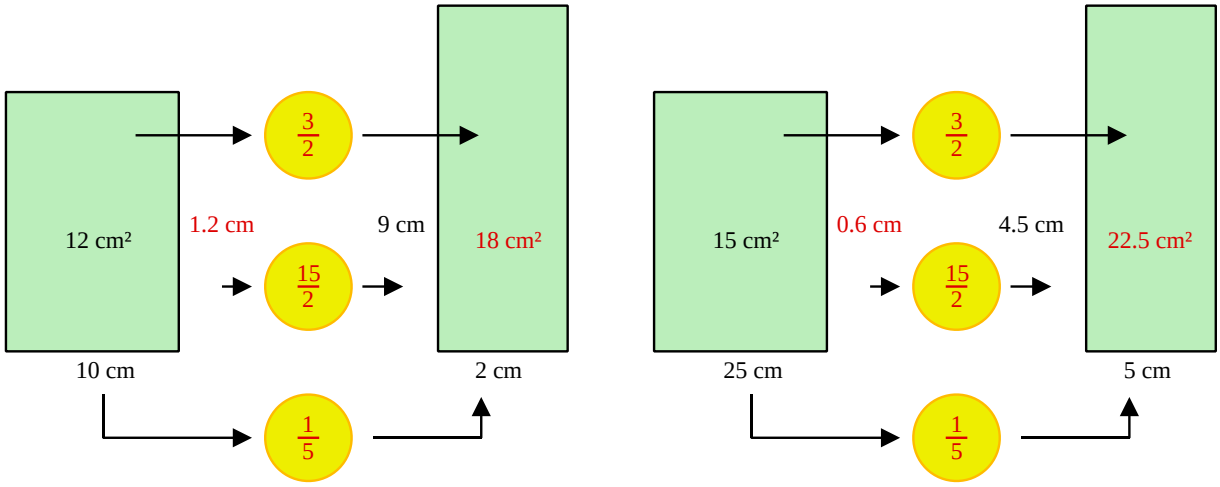
(ii)



Cathy's $18 \leqslant g < 20$ class is represented by a bar of width 0.8 cm and height 22.5 cm

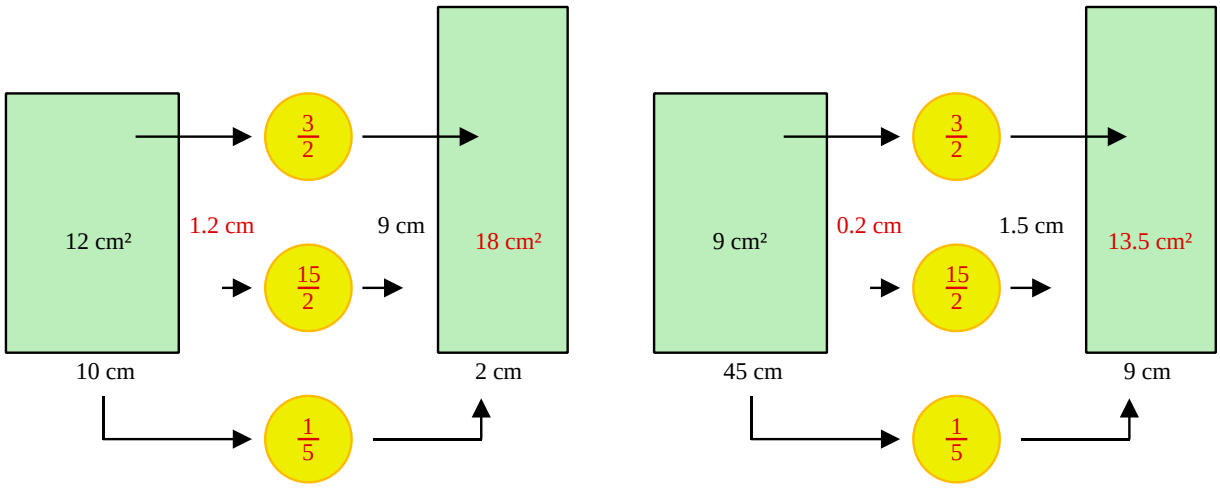
Answer 2

(i)



The $55 \leqslant L < 80$ group will have a width of 5 cm and height 4.5 cm

(ii)

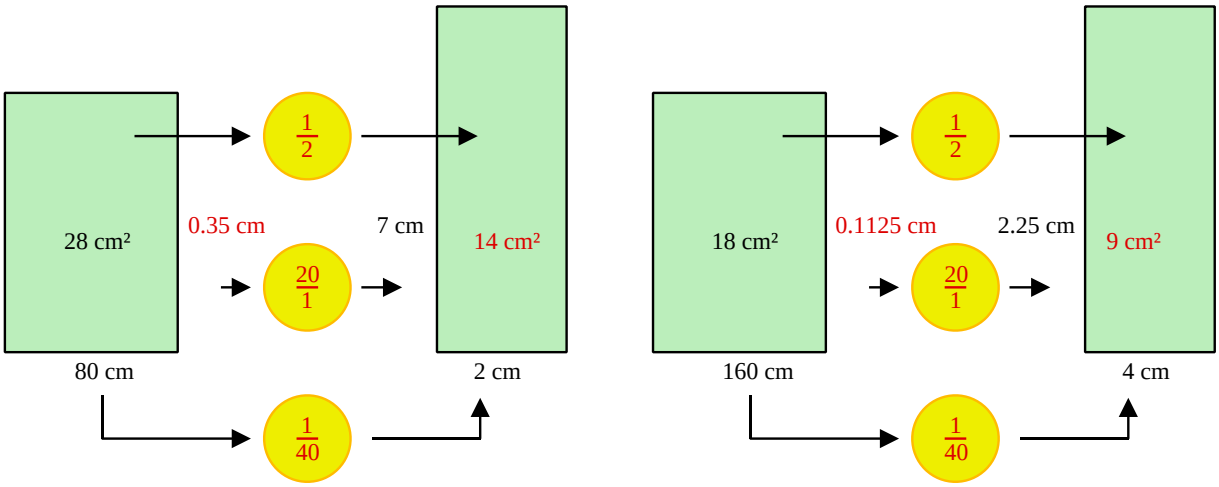


The $0 \leqslant L < 45$ group will have a width of 9 cm and height 1.5 cm

Answer 3

(i) A weekly income of £1038.46 will be in the $800 \leqslant y < 1200$ class interval.

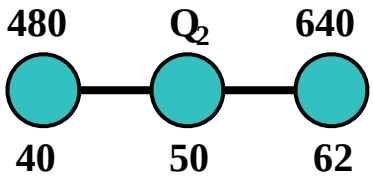
(ii)



The $640 \leqslant y < 800$ group will have a width of 4 cm and height 2.25 cm

(iii)

For Q_2 : $\dfrac{100 \times 2}{4} = 50$ which is in the class $480 \leqslant y < 640$



$$\frac{Q_2 - 480}{640 - 480} = \frac{50 - 40}{62 - 40}$$

$$Q_2 = 181.5727848$$

$$median = £552.73 \text{ per week}$$

$$\therefore \text{ About } £28821 \text{ per annum}$$

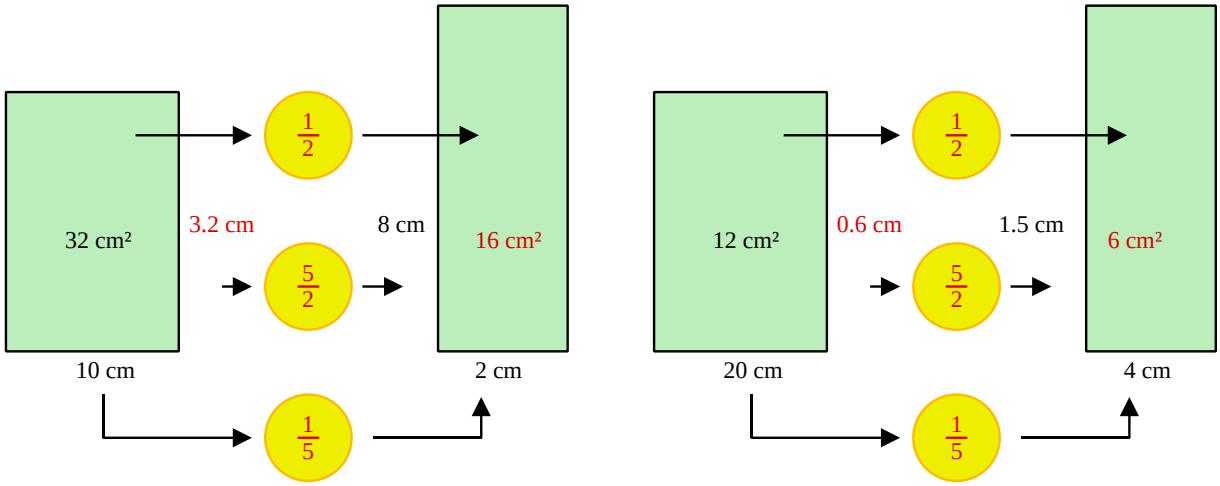
(iv) You can either agree or disagree with Toby and get full marks provided you justify your viewpoint !

A key point is to identify that a histogram should only be drawn for continuous data and as money is a discrete quanity, Toby has a valid point. Thus you could agree with Toby and get full marks.

However, the granulation of the data is so fine, relative to the quantities of money being discusses, that you could continue by arguing that, whilst Toby is technically correct, the data can be modellel accurately as being continuous. So, from a pragmatic point of view, the histogram is indeed a useful representation of the data.
Thus you could disagree with Toby, and get full marks.

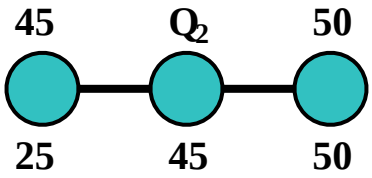
Answer 4

(a)



The $20 \leqslant y < 40$ group will have a width of 4 cm and height 1.5 cm

(b) For Q_2 : $\dfrac{90 \times 2}{4} = 45$ which is in the class $45 \leqslant y < 6 = 50$



$$\frac{Q_2 - 45}{50 - 45} = \frac{45 - 25}{50 - 25}$$

$$Q_2 = 49$$

$$median = £49 \text{ per square foot}$$

(c)

$$mean = \frac{12 \times 30 + 13 \times 42.5 + 25 \times 47.5 + 32 \times 55 + 8 \times 70}{90}$$

$$= \frac{4420}{90}$$

$$= \frac{442}{9} \qquad \text{(About 49.11...)}$$

(d)

$$standard\ deviation = \sqrt{\frac{226\,687.5}{90} - \left(\frac{442}{90}\right)^2}$$

$$= 10.3368$$

(e) As *mean* ≈ median the distribution is symmetric (negligible skew)