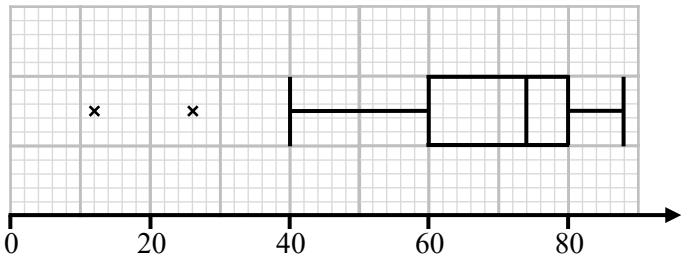
Outliers, Box Plots and Cumulative Frequency Diagrams

SCAN ME

REVISE THIS TOPIC

1 180 students completed a maths assessment. The maximum mark for the assessment was 100. The results are summarised in the box plot below.



(a) Write down the median test score. **(1)**

(b) Find the range of the test scores. **(1)**

(c) Find the interquartile range of the test scores. **(1)**

(d) State the meaning of the × symbols shown on the box plot. **(1)**

One of the students is selected at random. Eric says

"the probability that the student scored between 40 and 60 marks is the same the probability that the student scored between 60 and 80 marks."

(e) Explain how you know that Eric must be incorrect. **(1)**

(a) 74

(b) 88 – 12 = 76

(c) 80 – 60 = 20

(d) Outliers

(e) 60 is the lower quartile and 80 is the upper quartile so 50% of the data is between these values.

Only 25% of the data is below the lower quartile so less than 25% must be between 40 and 60.

(Total for Question 1 is 5 marks)

2 The table below shows the race times for the 100 metre final at the 2009 World Championships.

Race Position	1	2	3	4	5	6	7	8
Time (seconds)	9.58	9.71	9.84	9.93	9.93	10.00	10.00	10.34

(a) Calculate the interquartile range of the race times. **(1)**

(b) Calculate, to 3 decimal places, the mean race time. **(1)**

(c) Calculate, to 3 decimal places, the standard deviation of the race times. **(1)**

Hannah defines a race time an outlier if it falls either

more than 1.5 × (interquartile range) above the upper quartile or
more than 1.5 × (interquartile range) below the lower quartile.

(d) Determine if any of the race times are considered outliers using Hannah's definition. **(2)**

Ross defines a race time an outlier if it falls either

more than 2 × (standard deviation) above the mean or
more than 2 × (standard deviation) below the mean.

(e) Determine if any of the race times are considered outliers using Ross' definition. **(2)**

(a) n = 8 so lower quartile is 2.5th value and upper quartile is 6.5th value

Lower quartile = 9.775 Upper quartile = 10.00 Interquartile range = 10.00 – 9.775

= 0.225

(b) Mean = 9.916 seconds (from calculator)

(c) Standard Deviation = 0.211 (from calculator)

(d) 10 + 1.5 × 0.225 = 10.3375

9.775 – 1.5 × 0.225 = 9.4375

10.34 > 10.3375 therefore the time of runner number 8 is an outlier.

(e) 9.916 + 2 × 0.211 = 10.338

9.916 – 2 × 0.211 = 9.494

10.34 > 10.338 therefore the time of runner number 8 is an outlier.

(Total for Question 2 is 7 marks)

3 The table below shows the ages of 30 players in a football squad.

15	17	19	19	20	21	21	22	23	23	23	23	23	23	24
24	24	24	24	25	25	26	27	28	29	29	31	33	35	36

An outlier is any value that falls either

more than 1.5 × (interquartile range) above the upper quartile or
more than 1.5 × (interquartile range) below the lower quartile.

(a) Show that the ages two of the players in the football squad are considered outliers. **(3)**

(b) Draw a box plot for the ages of the players. **(2)**
You should indicate outliers with a ×

(a) n = 30 so lower quartile is 8th value and upper quartile is 23rd value

Lower quartile = 22 Upper quartile = 27 Interquartile range = 27 – 22
= 5

27 + 1.5 × 5 = 34.5

22 – 1.5 × 5 = 14.5

35 and 36 are both greater than 34.5 therefore they are both outliers.

(b) Minimum = 15

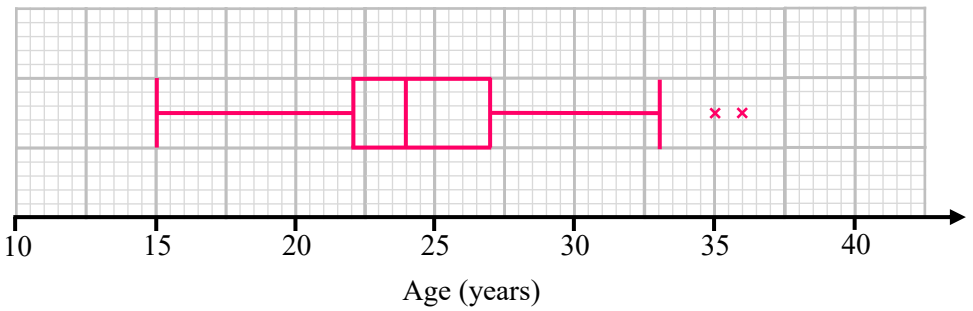
Lower Quartile = 22

Median = 24

Upper Quartile = 27

Maximum = 33 (non outlier)

Outliers: 35, 36



(Total for Question 3 is 5 marks)

4 A college tracks lateness to school to the nearest minute.
Students that are early or on time are considered to have 0 minutes of lateness.

A sample of 20 students from Year 12 and 20 students from Year 13 were taken on Monday.

The number of minutes lateness to school for the Year 12 students (x) is summarised below.

$$\sum x = 72 \qquad \sum x^2 = 1048$$

(a) Calculate, to 1 decimal place, the mean and standard deviation of the lateness for Year 12 students. **(2)**

An outlier is any value that falls either

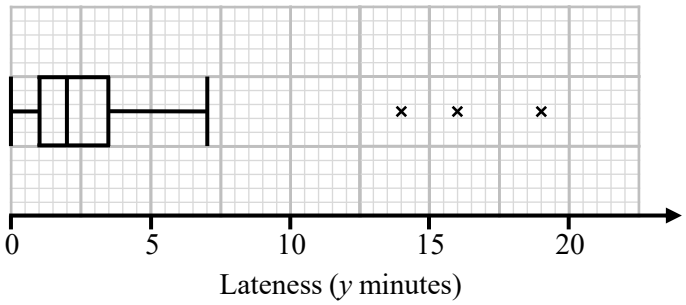
more than 2 × (standard deviation) above the mean or
more than 2 × (standard deviation) below the mean.

The students in Year 12 who were the latest to school are

Merry (13 minutes late) and Perry (26 minutes late)

(b) Show that Perry's lateness is the **only** outlier from the sample of 20 students in Year 12. **(3)**

The box plot below shows the lateness to school for the Year 13 students (y)



(c) State the meaning of the × symbols shown on the box plot. **(1)**

(d) Find the range of the lateness to school for the Year 13 students. **(1)**

Students who arrive to school 6 or more minutes late receive a detention.
A teacher claims that 5 of the Year 13 students received a detention on Monday for lateness.

(e) Show that the teacher must be incorrect. **(3)**

(a) $\bar{x} = \frac{72}{20} = 3.6$ minutes

$\sigma_x = \sqrt{\frac{1048}{20} - 3.6^2}$

$= 6.3$ minutes

(b) $3.6 + 2 \times 6.3 = 16.2$

$3.6 - 2 \times 6.3 = -9$

$26 > 16.2$ therefore Perry's lateness is an outlier.

$13 < 16.2$ therefore Merry's lateness is NOT an outlier.

It is not possible to have a time less than 0 therefore none of the other times are outliers.

(c) Outliers

(d) $19 - 0 = 19$ minutes

(e) Since there are 20 students, Q_3 (3.5 minutes) is located between the 15th and 16th value.

From the box plot we can see the 17th, 18th, 19th and 20th values are 7, 14, 16 and 19 minutes

If 5 students were late then the 16th value must be either 6 or 7 minutes.

If the 16th value was a 6, the 15th value must be a 1 since $Q_3 = (1 + 6)/2 = 3.5$

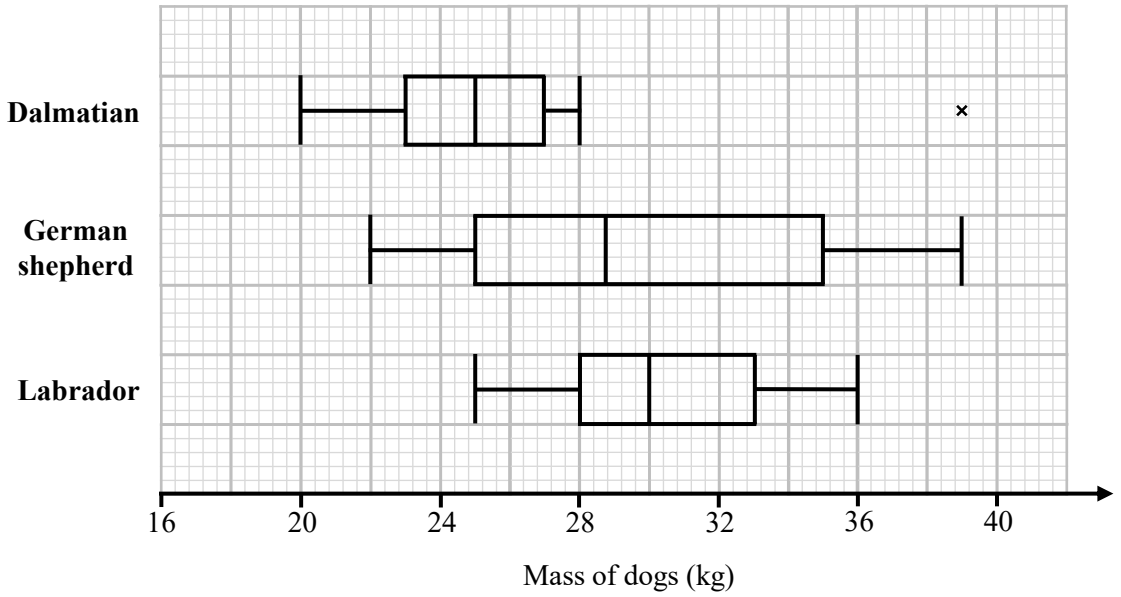
If the 16th value was a 7, the 15th value must be a 0 since $Q_3 = (0 + 7)/2 = 3.5$

The 15th value cannot be a 0 or a 1, as this is below the median (10.5th) value of 2.

Therefore only 4 students were late.

(Total for Question 4 is 10 marks)

5 A vet measured the masses of 40 dogs of three different breeds.
The results are summarised in the box plots below.



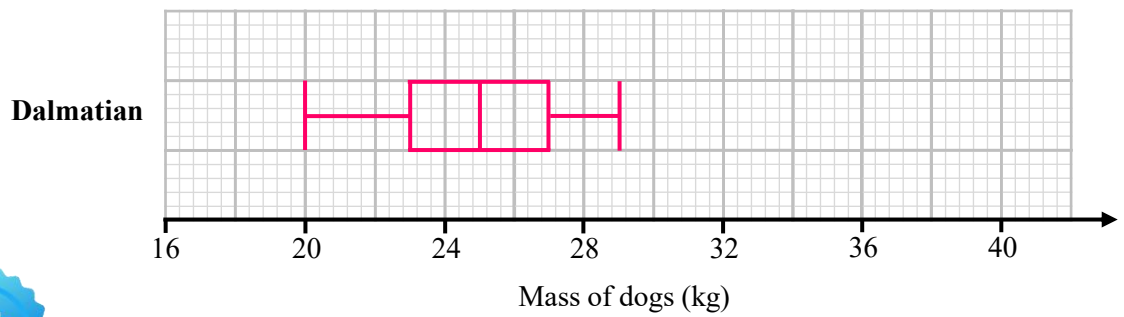
(a) Compare the masses of the German shepherds to the labradors. **(2)**

The heaviest dalmatian had a mass of 29 kg but was incorrectly logged at 39 kg.

(b) Given that the mass of 29 kg dalmatian is not considered an outlier, draw a corrected box plot for the masses of the 40 dalmatians on the grid below. **(2)**

(a) The median mass of the German shepherds is 28.5kg, which is lower than the median mass of the labradors, which is 30 kg. This means the labradors were heavier, on average.

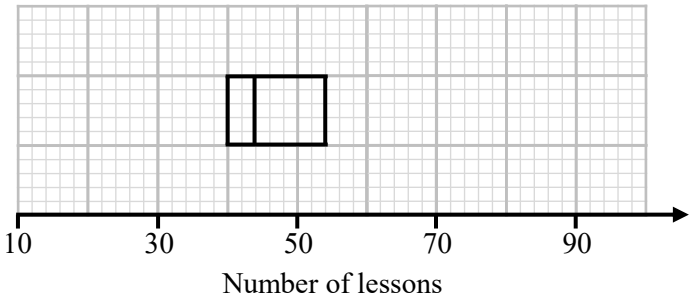
The interquartile range of the German shepherds masses is 35 – 25 = 10 kg, which is greater than the interquartile range of the masses of the labradors, which is 33 – 28 = 5 kg. This means that the masses of the German shepherds were more varied/less consistent.



(Total for Question 5 is 4 marks)

6 A driving instructor recorded the number of lessons before each of their learners were ready to take their driving test in 2023.

The box plot below shows the lower quartile, median and upper quartile for this data.



An outlier is any value that falls either

more than 1.5 × (interquartile range) above the upper quartile or
more than 1.5 × (interquartile range) below the lower quartile.

The learners with the greatest and least number of lessons are shown below.

Learner	A	B	C	D	E	F
Number of Lessons	14	24	25	72	80	85

(a) (i) Determine which of the values above are outliers. **(2)**

(ii) Hence, complete the box plot on the next page for the number of lessons. **(2)**

The driving instructor also recorded the number of lessons their learners had before they were ready to take their test in 2024. The 2024 data is summarised below.

Q_1	Q_2	Q_3
38	41	50

(b) Compare the number of lessons before learners were ready to take their driving test in 2023 and 2024. **(2)**

(a) Lower quartile = 40 Upper quartile = 54 Interquartile range = 54 – 40
= 14

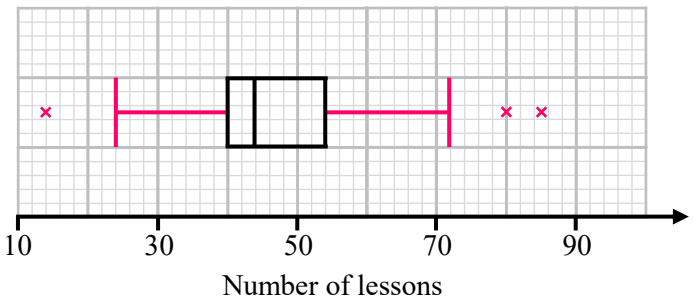
54 + 1.5 × 14 = 75

40 – 1.5 × 14 = 19

80 > 75 and 85 > 75, therefore there are both outliers.

14 < 19, therefore it is an outlier.

Outliers: 14, 80, 85



(b) The median number of lessons in 2024 is 41, which is lower than the median number of lessons in 2023, which is 44. This means that on average, learners in 2024 had fewer lessons.

The interquartile range of the number of lessons in 2024 is 50 – 38 = 12, which is less than the interquartile range of the number of lessons in 2023, which is 54 – 40 = 14. This means that the number of lessons for learners in 2024 is less varied/more consistent.

(Total for Question 6 is 6 marks)

7 A student recorded the sound level (S, dB) in a town centre every minute between 7am and 7pm. The results are summarised in the table below.

Sound level (S, dB)	Frequency (f)	Midpoint (S, dB)	C. Freq
$40 \leq S < 50$	99	45	99
$50 \leq S < 60$	162	55	261
$60 \leq S < 70$	132	65	393
$70 \leq S < 80$	129	75	522
$80 \leq S < 90$	120	85	642
$90 \leq S < 100$	78	95	720

You may use $\sum fS = 49230$ and $\sum fS^2 = 3544800$

(a) Calculate, to 3 decimal places, an estimate for the mean sound level. **(1)**

(b) Calculate, to 3 decimal places, an estimate for the standard deviation of the sound levels. **(2)**

An outlier is any value that falls either

more than 2 × (standard deviation) above the mean or
more than 2 × (standard deviation) below the mean.

The minimum sound level recorded was 41 dB.
The maximum sound level recorded was 98 dB.

(c) Use your answers to parts (a) and (b) to show that none of the sound levels are outliers. **(2)**

(d) Use linear interpolation to calculate estimates for the lower quartile, median and upper quartile for the sound levels. **(4)**

Another way to define an outlier is any value that falls either

more than 1.5 × (interquartile range) above the upper quartile or
more than 1.5 × (interquartile range) below the lower quartile.

(e) (i) Show that none of the sound levels were outliers using this definition. **(2)**

(ii) Draw a box plot for the sound levels. **(2)**

(a) $\bar{S} = \frac{49230}{720} = 68.375$ dB

(b) $\sigma_S = \sqrt{\frac{3544800}{720} - 68.375^2} = 15.754$ dB

(c) $68.375 + 2 \times 15.754 = 99.883$

$68.375 - 2 \times 15.754 = 36.867$

$41 > 36.867$ and $98 < 99.883$, therefore there are no outliers.

(d) $\frac{720}{2} = 360^{\text{th}}$ position

261		360	393
60		$60 + x$	70

$\frac{360 - 261}{393 - 261} = \frac{x}{70 - 60}$

$x = 7.5$

$\boldsymbol{Q_2 = 67.5}$ **dB**

$\frac{720}{4} = 180^{\text{th}}$ position

99		180	261
50		$50 + x$	60

$\frac{180 - 99}{261 - 99} = \frac{x}{60 - 50}$

$x = 5$

$\boldsymbol{Q_1 = 55}$ **dB**

$\frac{3 \times 720}{4} = 540^{\text{th}}$ position

522	540		642
80	$80 + x$		90

$\frac{540 - 522}{642 - 522} = \frac{x}{90 - 80}$

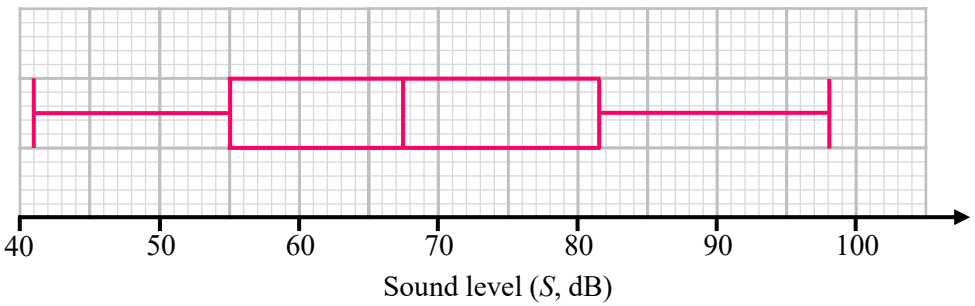
$x = 1.5$

$\boldsymbol{Q_3 = 81.5}$ **dB**

(e) $81.5 + 1.5 \times (81.5 - 55) = 121.25$

$55 - 1.5 \times (81.5 - 55) = 15.25$

$41 > 15.25$ and $98 < 121.25$, therefore there are no outliers.

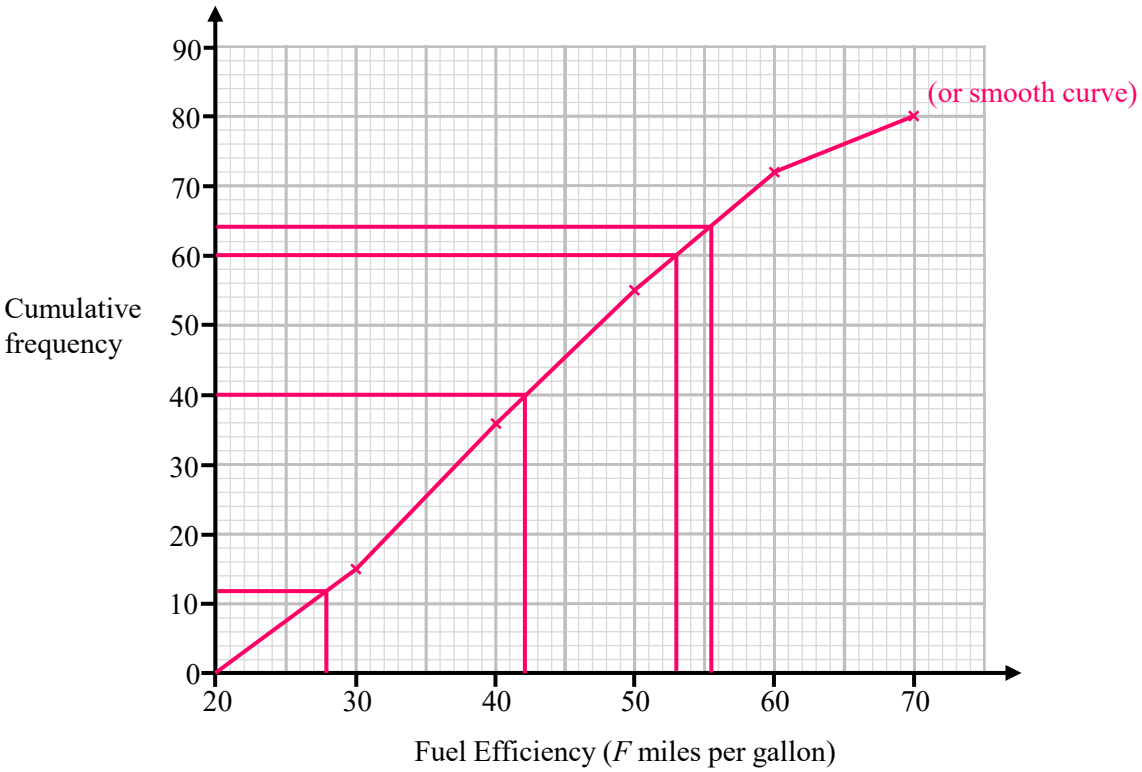


(Total for Question 7 is 13 marks)

8 The grouped frequency table gives information about the fuel efficiency (F, miles per gallon) of 80 different vehicles.

Fuel Efficiency (F miles per gallon)	$20 \leq F < 30$	$30 \leq F < 40$	$40 \leq F < 50$	$50 \leq F < 60$	$60 \leq F < 70$
Frequency	15	21	19	17	8
C. Freq	15	36	55	72	80

(a) On the grid, draw the cumulative frequency graph for this information. **(2)**



(b) Use your cumulative frequency diagram to estimate
(i) the median
(ii) the upper quartile
(iii) the 80th percentile
(iv) the 15th percentile **(4)**

(b) (i) $80 \times 0.5 = 40^{th}$ value = 42 miles per gallon

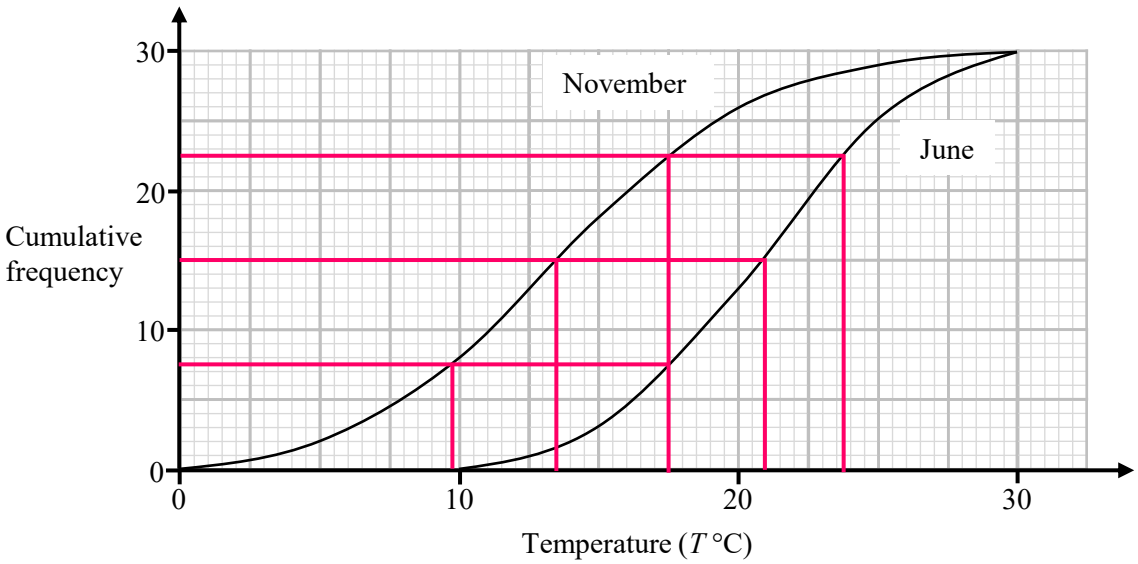
(ii) $80 \times 0.75 = 60^{th}$ value = 53 miles per gallon

(iii) $80 \times 0.8 = 64^{th}$ value = 55.5 miles per gallon

(iv) $80 \times 0.15 = 12^{th}$ value = 28 miles per gallon

(Total for Question 8 is 6 marks)

9 The cumulative frequency diagram below shows the maximum temperatures (T °C) for a city for each of the days during June and November.



(a) Work out estimates for the median and interquartile range of the temperatures in June. **(2)**

(b) Work out estimates for the median and interquartile range of the temperatures in November. **(2)**

(c) Compare the maximum temperatures in June and November. **(2)**

(a) Median = $30 \times 0.5 = 15^{th}$ value = 21 °C

Lower Quartile = $30 \times 0.25 = 7.5^{th}$ value = 17.5 °C

Upper Quartile = $30 \times 0.75 = 22.5^{th}$ value = 23.75°C

Interquartile range = 23.75 – 17.5 = 6.25 °C

(b) Median = $30 \times 0.5 = 15^{th}$ value = 13.5 °C

Lower Quartile = $30 \times 0.25 = 7.5^{th}$ value = 9.75°C

Upper Quartile = $30 \times 0.75 = 22.5^{th}$ value = 17.5°C

Interquartile range = 17.5 – 9.75 = 7.75 °C

(c) The median temperature for June is 21 °C, which is higher than the median temperature for November, which is 13.5° C. This means that on average, temperatures in June were higher.

The interquartile range for the temperatures in June 6.25 °C, which is less than the interquartile range for the temperatures in November, which 7.75°C. This means that the temperatures in June are less varied/more consistent.

(Total for Question 9 is 6 marks)