Chapter 3 - Statistics

Representations of data

Chapter Overview

1. Box plots and outliers

2. Cumulative frequency diagrams

3. Histograms

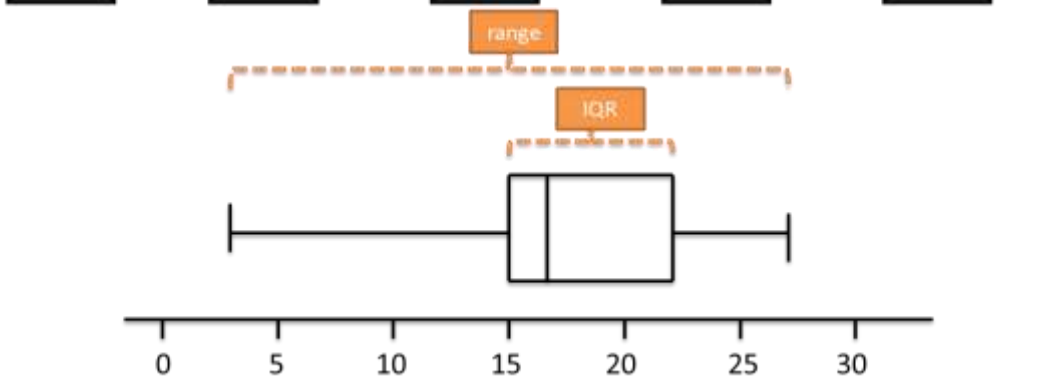
2 **Data presentation and interpretation**	2.1	**Interpret diagrams for single-variable data, including understanding that area in a histogram represents frequency.** **Connect to probability distributions.**	**Students should be familiar with histograms, frequency polygons, box and whisker plots (including outliers) and cumulative frequency diagrams.**

Topics	What students need to learn:		
	Content		Guidance
2 **Data presentation and interpretation** *continued*	2.4	**Recognise and interpret possible outliers in data sets and statistical diagrams.**	**Any rule needed to identify outliers will be specified in the question.** **For example, use of $Q_1 - 1.5 \times$ IQR and $Q_3 + 1.5 \times$ IQR or mean $\pm$ 3 $\times$ standard deviation.**
		Select or critique data presentation techniques in the context of a statistical problem.	**Students will be expected to draw simple inferences and give interpretations to measures of central tendency and variation.** Significance tests, other than those mentioned in Section 5, will not be expected.
		Be able to clean data, including dealing with missing data, errors and outliers.	**For example, students may be asked to identify possible outliers on a box plot or scatter diagram.**

Box Plots

Box Plots allow us to visually represent the distribution of the data.

Minimum	Lower Quartile	Median	Upper Quartile	Maximum
3	15	17	22	27
Sketch	Sketch	Sketch	Sketch	Sketch

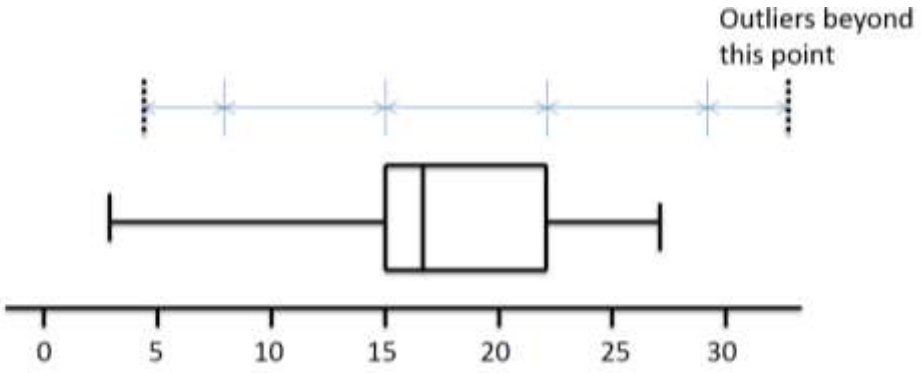


How is the IQR represented in this diagram?

How is the range represented in this diagram?

Outliers

An outlier is an extreme value.



One common definition of an outlier is when we're 1.5 IQRs beyond the lower and upper quartiles.

Examples

1. The diameters of 11 different Roman coins are measured in centimetres:

   2.2 2.5 2.7 2.7 2.8 3.0 3.1 3.1 3.2 4.0 4.7

   Determine the quartiles and hence any outliers.

2. [Textbook] The lengths, in cm, of 12 giant African land snails are given below:

   17 18 18 19 20 20 20 20 21 23 24 32

   Calculate the mean and standard deviation, given that $\Sigma x = 252$ and $\Sigma x^2 = 5468$. An outlier is an observation which lies ± 2 standard deviations from the mean. Identify any outliers for this data.

Test Your Understanding

The ages of 15 Lib Dem MPs are given:

11 18 20 27 30 31 32 32 35 36 37 58 63 78 104.5

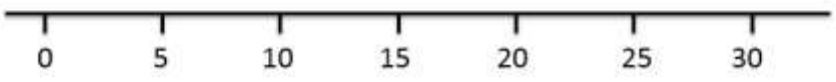
a) If an outlier is considered to be 1.5 interquartile ranges below the lower quartile or above the upper quartile, determine any outliers.

b) If instead an outlier is considered to be outside 2 standard deviations within the mean, determine any outliers. Note that $\Sigma x = 612$ and $\Sigma x^2 = 33606$

Box Plot Example

Smallest values	Largest values	Lower Quartile	Median	Upper Quartile
0, 3	21, 27	8	10	14

Draw a box plot to represent the above data.



Test Your Understanding

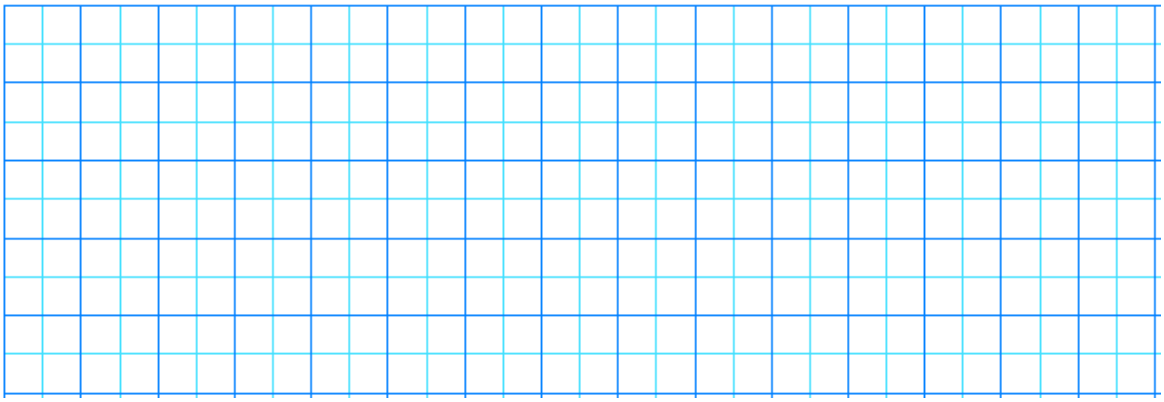
[Jan 2011 Q3] Over a long period of time a small company recorded the amount it received in sales per month. The results are summarised below.

	Amount received in sales (£1000s)
Two lowest values	3, 4
Lower quartile	7
Median	12
Upper quartile	14
Two highest values	20, 25

An outlier is an observation that falls
either 1.5 × interquartile range above the upper quartile
or 1.5 × interquartile range below the lower quartile.

(*a*) On the graph paper below, draw a box plot to represent these data, indicating clearly any outliers. **(5)**

(c) The company claims that for 75% of the months, the amount received per month is greater than £10 000. Comment on this claim, giving a reason for your answer. **(2)**

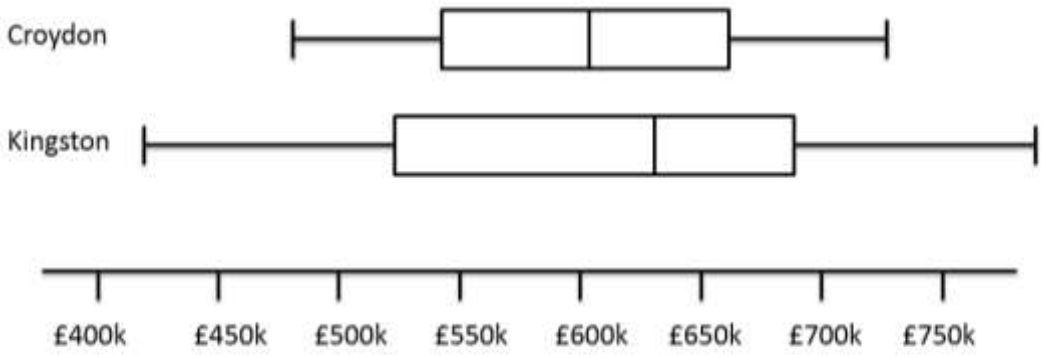


Comparing Box Plots

It is important to be able to compare the data that is shown in 2 or more box plots. You should consider the median and quartiles as well as the spread of the data. Always relate the comparison back to the specific situation being analysed.

Examples

1. **Box Plot comparing house prices of Croydon and Kingston-upon-Thames:**



"Compare the prices of houses in Croydon with those in Kingston". (2 marks)

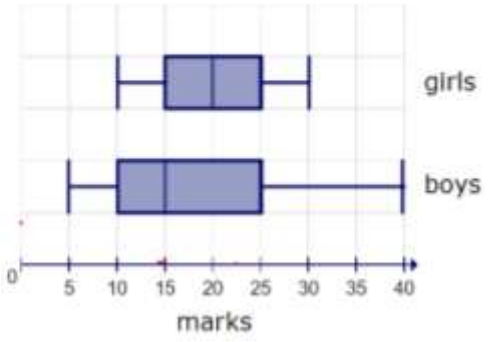
2.

Consider these box plots comparing marks in a maths competition for boys and girls.

Who had the greater median?

boys

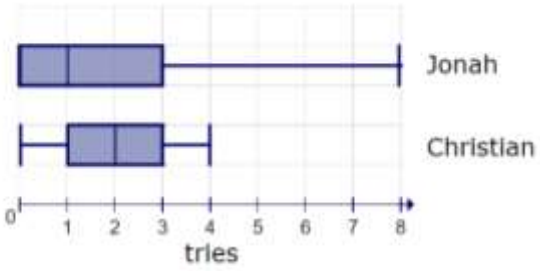
girls



3.

A coach for a rugby club needs to choose between two different wingers for the next game.

The box plots show the number of tries scored by each winger over the last 10 matches.



Which winger should the coach pick?

Jonah | Christian

Ex 3A/3B Pg 42-43, 45

Cumulative Frequency Diagrams

We use cumulative frequency diagrams to consider the running totals of / people/ things up to a given value. They are useful for estimating the median and quartiles.

Example: The table below shows the time taken for a group of runners to run 50m. Draw a Cumulative Frequency curve for the data. Use your graph to estimate the median, LQ, UQ and IQR.

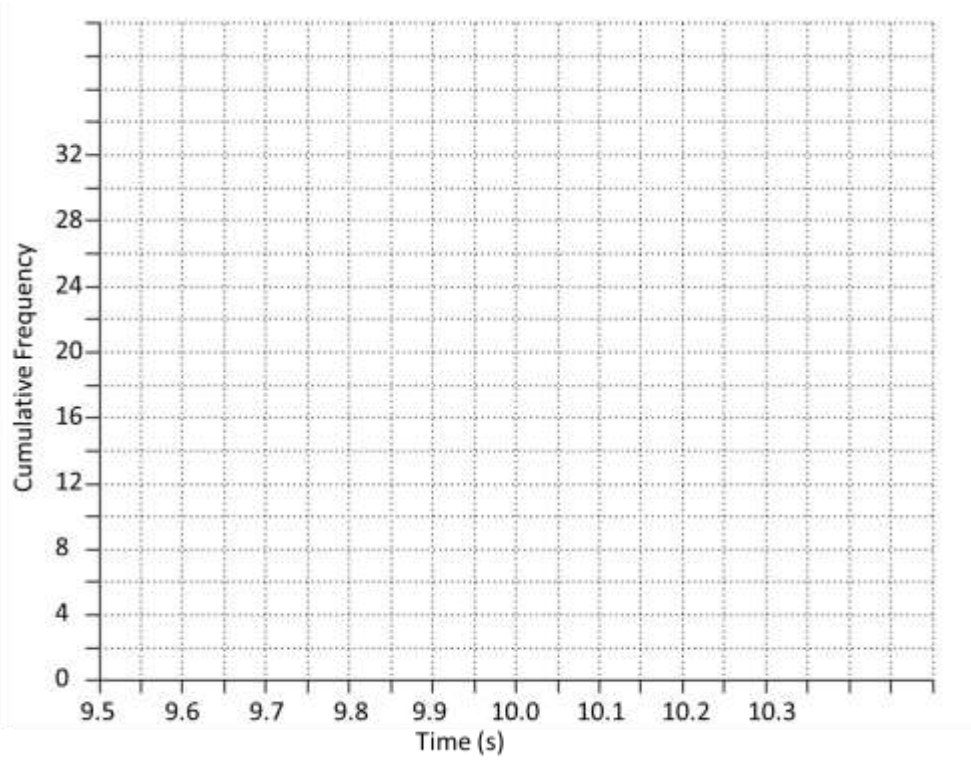
Time (s)	Frequency	C. Freq
$9.6 < t \leq 9.7$	1	1
$9.7 < t \leq 9.9$	4	5
$9.9 < t \leq 10.05$	10	15
$10.05 < t \leq 10.2$	17	32

Median =

LQ =

UQ =

IQR =



Estimate how many runners had a time less than 10.15s.

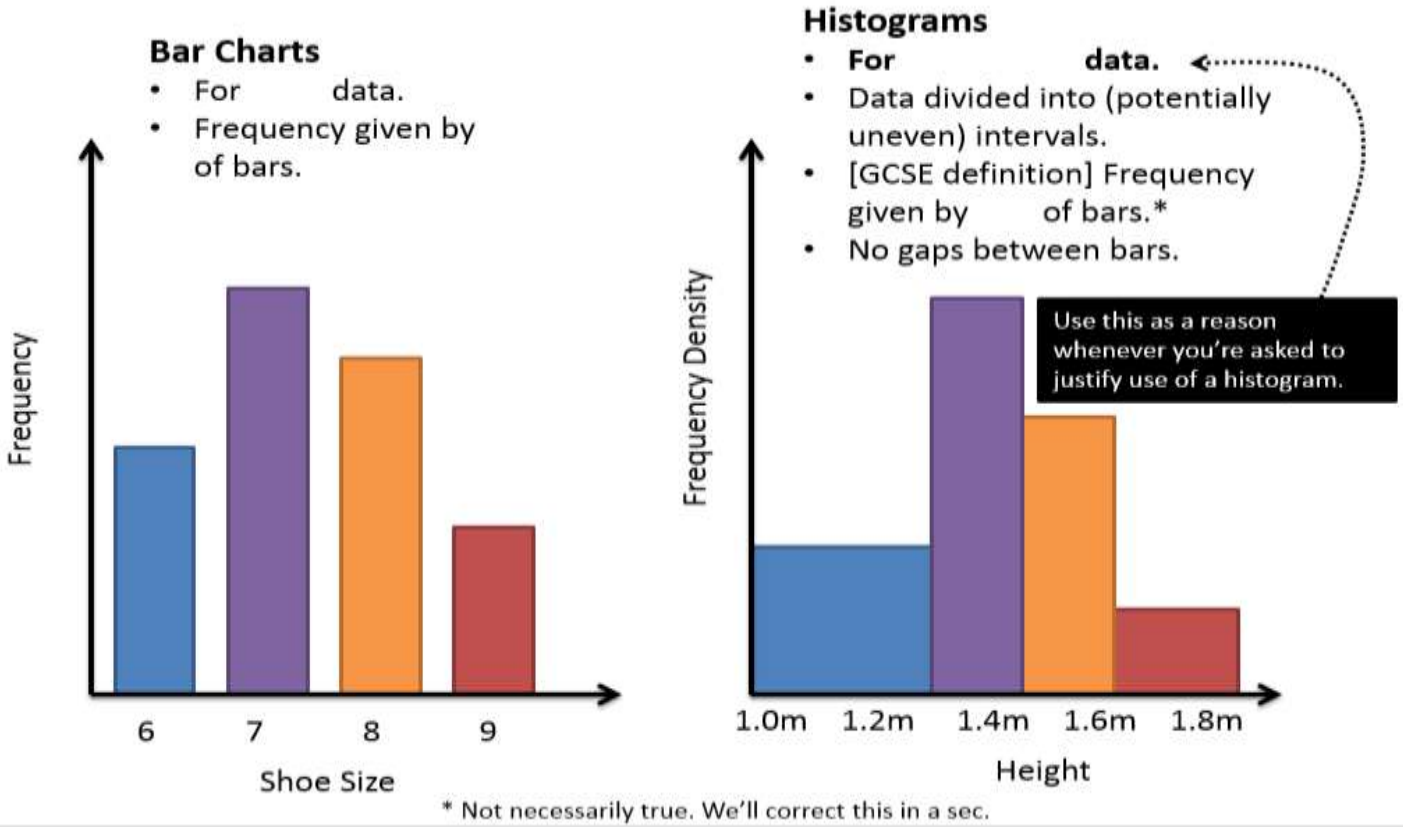
Estimate how many runners had a time more than 9.95

Estimate how many runners had a time between 9.8s and 10s

Ex 3C Pg 47/48

Histograms

You should remember from GCSE that there are some important differences between bar charts and histograms. We will consider 4 important skills.

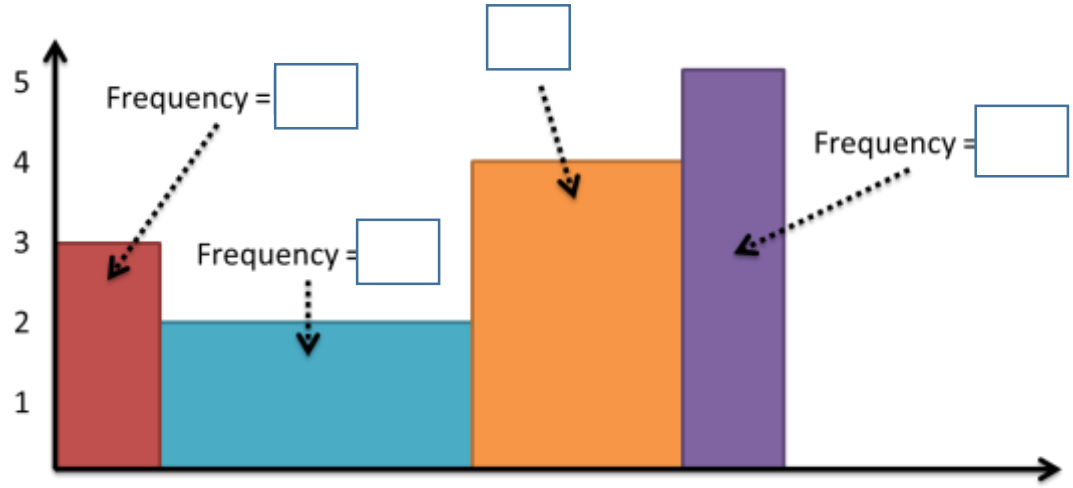


Example

1. Calculate the missing values in the table below

Weight (w kg)	**Frequency**	**Frequency Density**
$0 < w \leq 10$	40	
$10 < w \leq 15$	6	
$15 < w \leq 35$		2.6
$35 < w \leq 45$		1

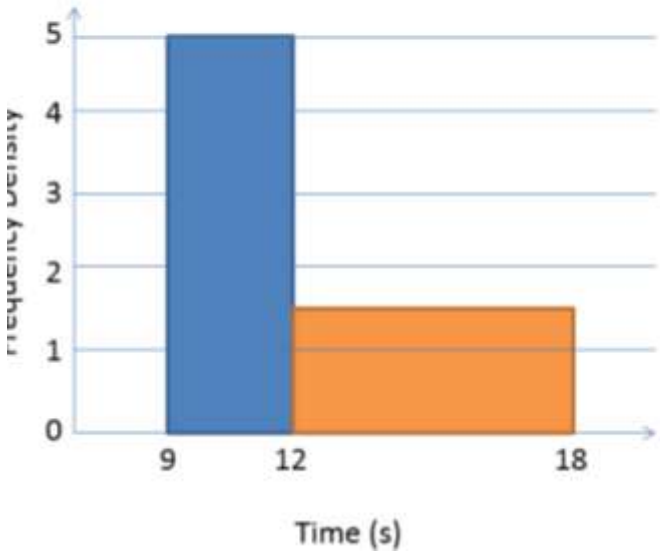
2. Calculate the frequencies



1. **Let's consider the area of the bars:**

Example

There were 60 runners in a 100m race. The following histogram represents their times. Determine the number of runners with times above 14s.

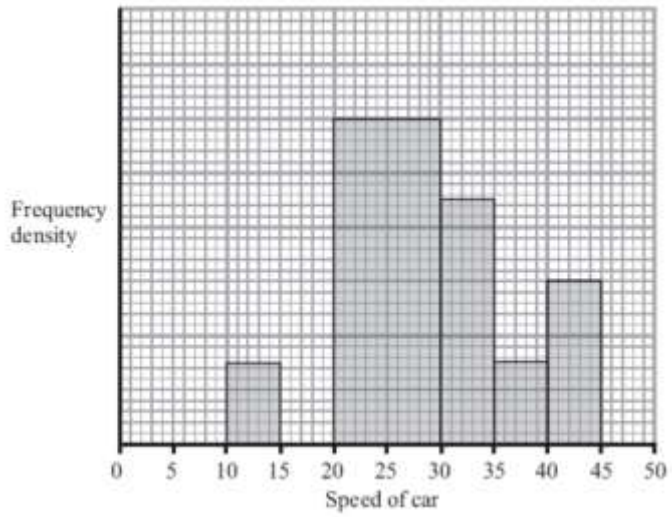


Total frequency is known; therefore find total area and hence the 'scaling'.

Then use this scaling along with the desired area.

Test Your Understanding

A policeman records the speed of the traffic on a busy road with a 30 mph speed limit. He records the speeds of a sample of 450 cars. The histogram in Figure 2 represents the results.



(*a*) Calculate the number of cars that were exceeding the speed limit by at least 5 mph in the sample. *(4 marks)*

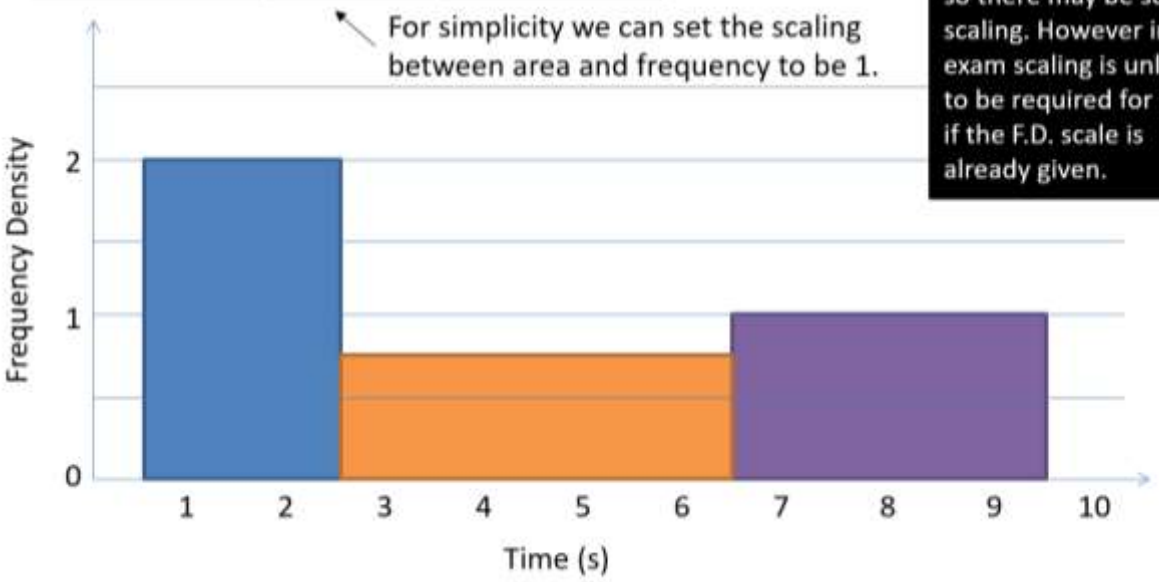
(b) Estimate the value of the mean speed of the cars in the sample. *(3 marks)*

(*c*) Estimate, to 1 decimal place, the value of the median speed of the cars in the sample.(2)

2. Let's Consider the gaps between the classes:

Example

Weight (to nearest kg)	Frequency	F.D.
1-2	4	
3-6	3	
7-9		



Note the gaps affects class width!
Remember the frequency density axis is only correct to scale, so there may be some scaling. However in an exam scaling is unlikely to be required for F.D. if the F.D. scale is already given.

3. Let's consider the width and height on the diagram

An exam favourite is to ask what width and height we'd draw a bar in a drawn histogram.

Example:

The frequency table shows some running times. On a histogram the bar for 0-4 seconds is drawn with width 6cm and height 8cm. Find the width and height of the bar for 4-6 seconds.

Time (seconds)	Frequency
$0 \leq t < 4$	8
$4 \leq t < 6$	9

Tip:

<u>0 -4 class</u>

Class width = Drawn width = Scaling =

Frequency Density (height) = Drawn height = Scaling =

<u>4-6 class:</u>

Test Your Understanding

[May 2009 Q3] The variable x was measured to the nearest whole number. Forty observations are given in the table below.

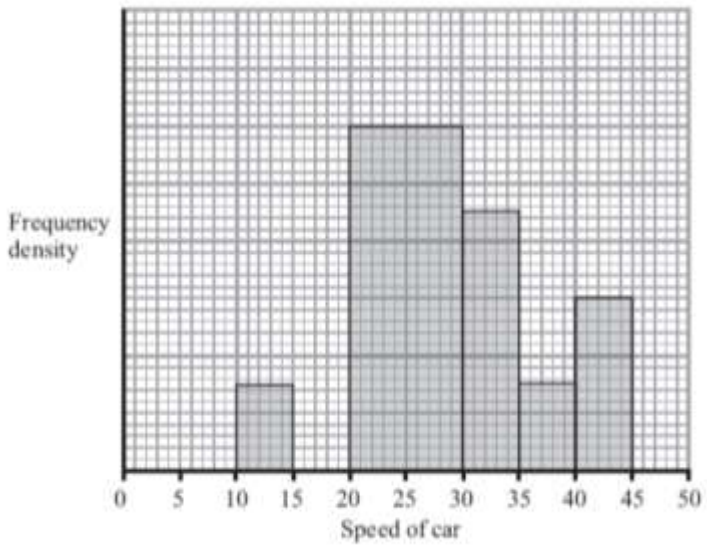
x	10 – 15	16 – 18	19 –
Frequency	15	9	16

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

(*a*) the width, **(1)**
(*b*) the height **(2)**
of the bar representing this class.

4. Forming a frequency polygon

Recall that a frequency polygon can be drawn by using the midpoint of each interval.
This corresponds to the midpoint of the top of each bar in a histogram.



Exercise 3D Pg 50

Supplementary questions on printed sheet

Exercise 3E Pg 53

The Large Data Set

Locations

5 UK weather stations

3 overseas

Time Periods

May – October 1987 (6 months)
May – October 2015 (6 months)

Seasons

May/June are the end of spring
July-Sept is summer
October is autumn

Perth (Australia) is in the southern hemisphere, so July-Sept is winter

UK Great Storm

The night of 15-16th October 1987
Gusts up to 100 knots recorded

Florida hurricanes

12 October 1987 Hurricane Floyd
1-2 October 2015 Hurricane Joaquin

Variables Recorded

Daily Maximum Temperature
°C

Daily Total Rainfall
mm

Daily Total Sunshine
hours

Daily Maximum Relative Humidity
%; mist and fog if > 95%

Daily Mean Windspeed;
Daily Maximum Gust
knots (1kn = 1.15mph)
and Beaufort scale

Daily Mean Wind Direction;
Daily Maximum Gust Direction
bearing (°)
and cardinal direction

Cloud Cover
oktas (eights); 0 – 8

Visibility
Dm (decametres)
1 Dm =10m

Pressure
hPa (hectoPascal)

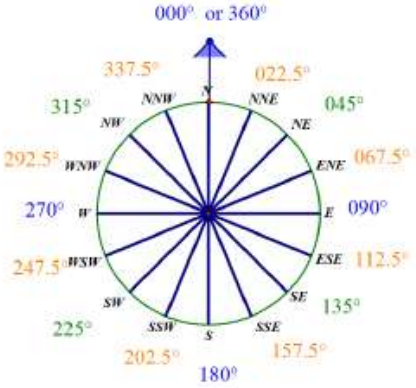
n/a
reading not available

tr (trace)
rainfall < 0.05mm

Beaufort Scale

Discrete, scale of 13 values:
0 (calm, < 1kn)
12 (hurricane, 64kn+)

Cardinal Directions



Oktas

Eighths of the sky covered by cloud
Discrete, scale of 9 values:
0 (clear sky)
8 (completely overcast)

Sources
Maps: *Pearson*
Compass: *mathsmutt.co.uk*