Edexcel International AS Maths: Statistics 1

Your notes

Working with Data

Contents

Your notes

Data Presentation

What graphs and diagrams should I be familiar with?

- You will be expected to be able to use a variety of graphs such as:
  - **Stem-and-leaf diagrams**
    - Can be used with ungrouped data of a single variable
    - Shows all the data and the shape of its distribution
  - **Box plots**
    - Can be used with ungrouped data of a single variable
    - Shows the range, interquartile range and quartiles clearly
    - Very useful for comparing data patterns quickly
  - **Histograms**
    - Can be used with continuous grouped data of a single variable
    - Can be used with varying group sizes
    - Shows the frequencies of the group, represented by the area of each bar
  - **Frequency polygon**
- Displays the shape of continuous grouped data
- Created by joining the midpoints at the top of each bar of a histogram
- **Scatter diagrams**
  - Can be used with ungrouped **bivariate** data (data for two variables)
  - Shows the pattern of the relationship between the variables
- You will not be expected to draw the graphs from scratch but you may be asked to add to a graph
  - You should make sure you know how to draw them anyway as it will help strengthen your understanding of the graphs

What should I look out for when interpreting graphs?

- Look carefully at the context of the information given in the graph
- Check the scales on both axes carefully, including units
  - Sometimes the numbers will be abbreviated to fit on the scale, for example if a population is given in millions then the number 60 will represent 60 000 000

- Look carefully at the labels and units to determine how a value should be read

- If there is more than one graph represented on the same set of axes take extra care to ensure you are reading from the correct one
- Beware of misleading graphs, the scales on the axes, units and representation can be manipulated to make a graph look more/less convincing

Your notes

Worked Example

A student is collecting information on his friends' interests and believes that his friends who only have dogs spend more time outside than his friends who only have cats. He has surveyed 20 friends with only cats and 20 friends with only dogs and has written down the total amount of time, rounded to the nearest hour, each of them spent outside last week. Describe, with a reason, which diagram would be best for the student to use to display the data.

Answer:

- The data is raw and ungrouped
- Time, rounded to the nearest hour, is discrete
- The student wants to compare the distributions

The student should use two box plots to display the data, one for his friends with cats and one for his friends with dogs.

Examiner Tips and Tricks

- Take the time needed when working with diagrams, they are usually 'easy marks' questions but it is common for students to rush them and make silly mistakes.

Your notes

Stem and Leaf Diagrams

What is a stem and leaf diagram?

- A **stem and leaf** diagram shows **ALL RAW** data and groups it into **class intervals**
- Stem and leaf diagrams lend themselves to two-digit data but can be used with three-digit data, rarely more

0	1	3	4	4		(4)
1	0	2	6	7		(4)
2	3	3	3	5	8	(5)
3	1	2				(2)

KEY 2 | 3 REPRESENTS 2.3

- The numbers in **brackets** indicate how many values are in that class interval
    - These are not always included but can be useful when there is a large amount of data to display

How do I draw a stem and leaf diagram?

- Identify the **stems** and the **leaves**
    - **Leaves** would always be **single** digits
        - the number **2** would be represented by **12 | 2**
- If starting from **unordered** data draw **two** diagrams
    - The **first** diagram should get the data into the right **format**
        - i.e. a list of stems with their corresponding leaves
    - The **second** diagram should have stems and leaves in **order**, with a **key**
    - This helps **accuracy** as values are less likely to be missed out

What are stem and leaf diagrams used for?

- The data is arranged into **classes** so at a glance it is possible to see the **modal class interval**
- As the data is in order the **median**, **quartiles**, **maximum** and **minimum** can be identified easily
    - Check you can do this – find the minimum, maximum, median and upper and lower quartiles from the stem and leaf diagram at the start of this revision note

Your notes

- Note that these five values are those needed in order to construct a **box-and-whisker diagram** (box plot)

- **Outliers**, once defined, can be easily identified and removed

What about back-to-back stem and leaf diagrams?

- These are used when it is helpful for the data to be split into two comparable categories such as boy/girl, child/adult, UK/non-UK. Etc

	BOYS		GIRLS	
(2)	4 2	0	1 3 4 4	(4)
(4)	9 6 3 1	1	0 2 6	(3)
(4)	6 5 4 0	2	3 3 3 5 8	(5)
(1)	2	3	1	(1)

KEY 6|2|3 REPRESENTS 2.6 FOR THE BOYS AND 2.3 FOR THE GIRLS

- Note that the leaves on the left-hand side of the stems (Boys) increase from the centre outwards

Are there any variations on stem and leaf diagrams?

- There are a few minor variations on stem and leaf diagrams that you may see online or in different textbooks

n = 15 — NUMBER OF DATA VALUES

2 | 3 REPRESENTS 2.3 — KEY (AT TOP)

LOW 0.1 0.3 — LOW VALUED OUTLIERS

0	4 4	(2)
1	0 2 6 7	(4)
2	3 3 3 5 8	(5)
3	1	(1)

HIGH 3.2 — HIGH VALUED OUTLIERS

- Some or all the different/extra features in the diagram above may appear
- These differences can be applied to back-to-back stem and leaf diagrams
- With large amounts of data, the stems may be split into **two rows**
    - Every stem will be listed **twice**
    - The **first** row for a **stem** will contain **leaves 0 – 4**
    - The **second** row will contain **leaves 5 – 9**

What might I be asked to do with a stem and leaf diagram?

- You may be asked to **draw** or **complete** a stem and leaf diagram

Your notes

- Find **statistical measures** – **median, quartiles** and interquartile range in particular
  - From which you may be required to draw a **box-and-whisker diagram**
- Identify and remove **outliers**
- **Compare** data shown by stem and leaf diagrams (either separate or back-to-back); comment on two things and each should be in both terms of the maths and the context of the question
  - a comment about **average** (use median)

e.g. the girls' **median** of 88% was higher than the boys' median of 65% so on average the girls performed better on the test

- a comment about **variation** (**spread**) (use interquartile range)

e.g. the girls' **interquartile range** of 30% was greater than the boys' 15% so the boys had more consistent scores on the test

- **Analyse** what would happen to statistical measures such as the median and quartiles if a **value changed** or a **new value** were to be added to the data

Worked Example

The following stem and leaf diagrams show the times taken by some children and adults to complete a level on a computer game.

				Children			Adults					
(3)			9	8	8	1						(0)
(4)		6	4	3	1	2	3	5	7	8	8	(5)
(5)	7	6	5	4	2	3	0	4	7			(3)
(3)			3	0	0	4	0	1	2			(3)

2 | 3 represents a time of 23 seconds

(a) Compare the times taken to complete the level between the children and the adults.

(b) It is later discovered two of the adults' times had been omitted from the diagram –times of 23 and 42 seconds.
Briefly explain whether adding these times would change the adults' median time.

Answer:

a) To compare, use an average – so medians – and a measure of spread – so interquartile range

	n	Q_1	Q_2	Q_3
Children	15	4th	8th	12th
Adults	11	3rd	6th	9th

	Children						Stem	Adults					
(3)			9	8	8		1						(0)
(4)		6	4	3	1		2	3	5	7	8	8	(5)
(5)	7	6	5	4	2		3	0	4	7			(3)
(3)			3	0	0		4	0	1	2			(3)

	Q_1	Q_2	Q_3	IQR
Children	21	32	37	37 − 21 = 16
Adults	27	30	40	40 − 27 = 13

The children had a slightly higher median time of 32 seconds compared to the adults' 30 seconds
This means, on average, the adults completed the level on the computer game quicker than the children
The children also had a greater interquartile range of 16 seconds compared to the adults' 13 seconds
This means the adults' times to complete the level on the computer game were more consistent (less varied)

b) 23 seconds is less than the adults median of 30 seconds
42 seconds is greater
Therefore there will be no change to the adults' median time

Examiner Tips and Tricks

- **Accuracy** is important
  - (Lightly) tick off values as you add them to a stem and leaf diagram
  - Check you have the right number of data values in total on your diagram
    - Other checks can include ensuring the median has the same number of values either side of it

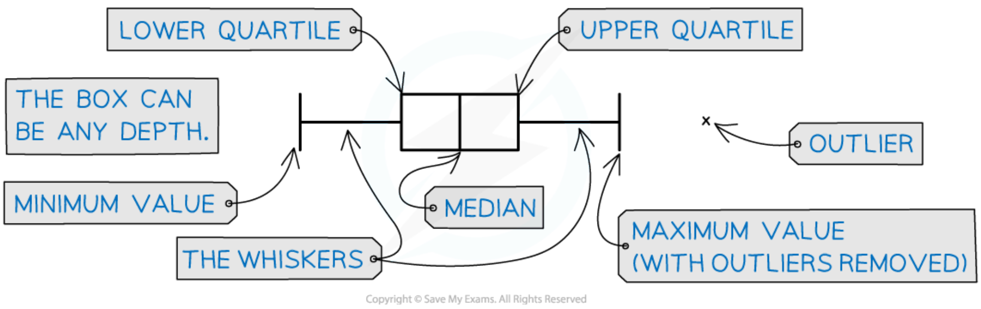
Your notes

Your notes

Box plots

What is a box plot?

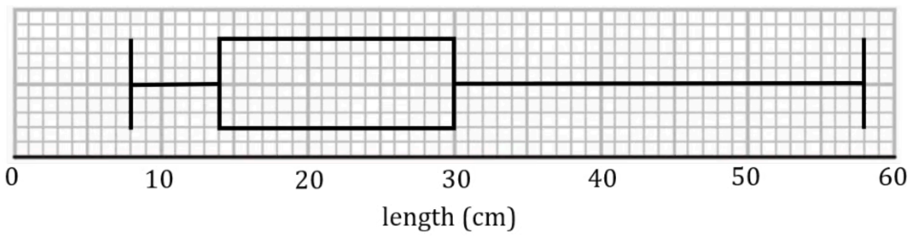
- A box plot is a graph that clearly shows key statistics from a data set
  - It shows the **median, quartiles, minimum** and **maximum values** and **outliers**
  - It does not show any other individual data items
- The middle 50% of the data will be represented by the box section of the graph and the lower and upper 25% of the data will be represented by each of the whiskers
- Any outliers are represented with a cross on the outside of the whiskers
  - If there is an outlier then the whisker will end at the value before the outlier
- Only one axis is used when graphing a box plot
  - It is still important to make sure the axis has a clear, even scale and is labelled with units
- Box plots are often used for comparing two sets of data
  - Both box plots will be drawn one above the other on the same scale on the x-axis
  - They are useful for **comparing data** because it is easy to see the main shape of the distribution of the data from a box plot



The key features of a box plot

Worked Example

The incomplete box plot below shows the tail lengths in cm of some students' pets.

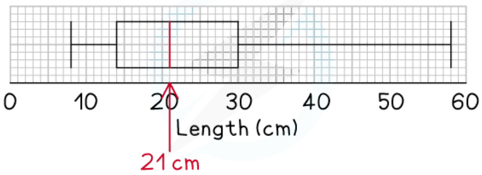


(a) Given that the median tail length was 21 cm, complete the box plot.

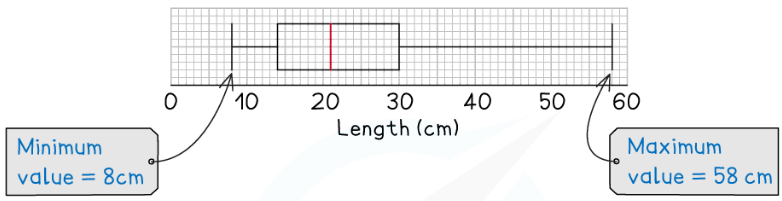
(b) Find the range and interquartile range of the tail lengths.

Answer:

(a) Check the scale carefully before plotting the median

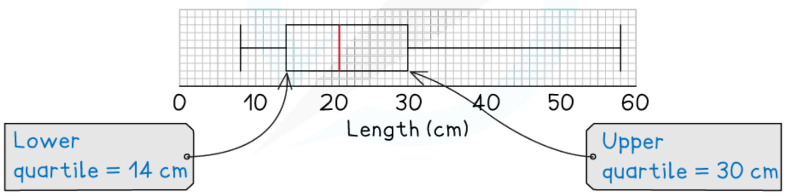


(b) Range = maximum value − minimum value



Range = 58 − 8 = 50 cm

Interquartile range = Upper quartile − Lower quartile



Interquartile range = 30 − 14 = 16 cm

Range = 50 cm, IQR = 16 cm

Examiner Tips and Tricks

- Remember a box plot is a graph and should be treated like one, even though there is only one axis. It should have a title, a clear, even scale that is labelled with

Your notes

units if there are any. If drawing two box plots on the same axis label each one clearly.

Your notes

Cumulative frequency

What is a cumulative frequency graph?

- A cumulative frequency graph is used with data that has been organised into a **grouped frequency** table
- The cumulative frequency graph can be used to find estimates of percentiles and quartiles
  - As the data is grouped, it is not possible to find the actual values of these statistics

What are the main features of a cumulative frequency graph?

- A cumulative frequency graph considers how much data there is up to a certain value, including the data in that group and the one below
- Cumulative frequency will always be plotted on the y – axis
  - Consider the scale carefully because this will usually be a large number
  - You may be asked to add to one or both axes, remember to label both axes clearly and include units on the x – axis if they are needed
- The cumulative frequency is calculated by adding the frequency in each group, or class, to the frequency in the ones before
  - This is essentially accumulating the frequencies as you go
- The cumulative frequency for each class must be plotted against the **upper boundary** of each corresponding class
  - The cumulative frequency that corresponds with each upper boundary will not only consider the frequency of the data in that class, but all of the data in the groups below it too
- When the points have been plotted they should be joined up with a smooth curve
  - However, some may be joined with straight lines from point to point

How do we read statistics from a cumulative frequency graph?

- Quartiles and percentiles can be read from a cumulative frequency graph
  - The median, Q_2 is read from the y – axis scale at the $\frac{n}{2}$th value

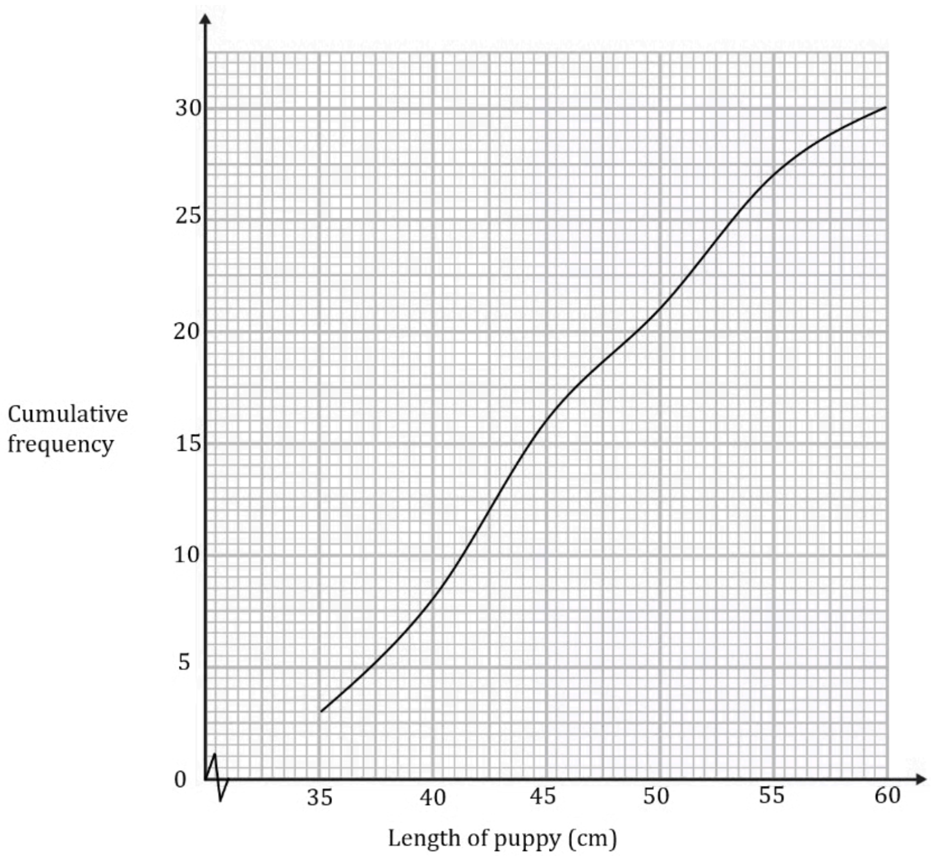
- The lower quartile, Q_1 ,is read from the $\frac{n}{4}$ th value and the upper quartile, Q_3 is read from the $\frac{3n}{4}$ th value
- Any percentile can be read from the graph by finding the percent of the total frequency and reading from the value on the y - axis

- To read the corresponding data value once the position on the y – axis is known, use a ruler to draw a line from the y – axis to the graph and then down to the x – axis
- Sometimes the frequency of values greater than or less than a particular data value will need to be found, this time you will have to read from the x – axis to the y – axis
  - Take particular care if the question asks for a frequency greater than a particular data value, the value found from the y – axis will need to be subtracted from the total frequency

Your notes

Worked Example

The cumulative frequency graph below shows the lengths in cm, l , of a group of puppies in a training group.



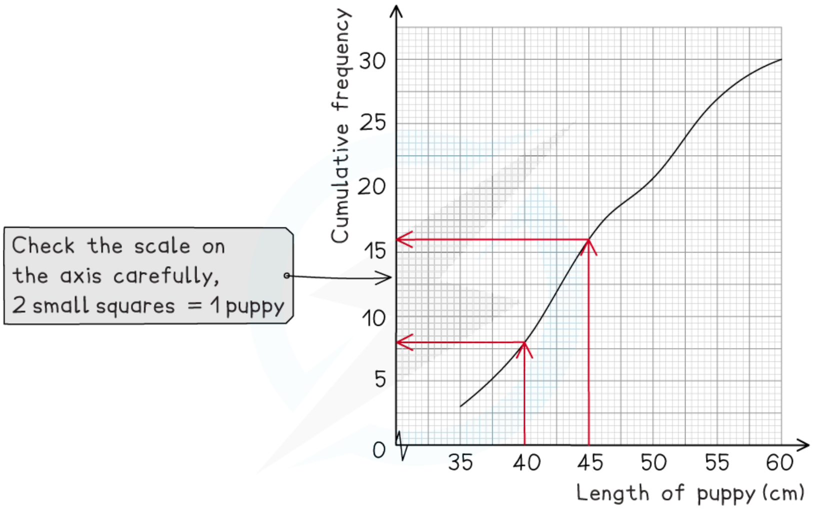
(i) Given that the group $40 < l \leq 45$ was one of the groups used in the data collection, find the number of puppies that were in this group.

(ii) Use the graph to find an estimate for the interquartile range of the puppies.

(iii) X % of the puppies are greater than 53.5 cm long, use your graph to find an estimate for the value of X.

Answer:

(i) Find the frequency at 40 cm and 45 cm.



8 puppies are ⩽ 40 cm

16 puppies are ⩽ 45 cm

∴ $40 < \ell \leqslant 45$ has 16 − 8 puppies

There are 8 puppies in the group $40 < \ell \leqslant 45$

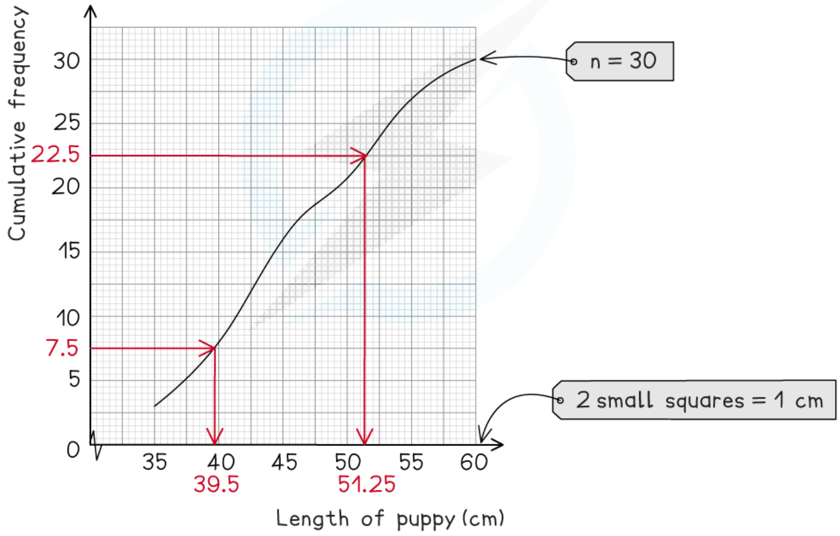
Your notes

(ii) Interquartile range = upper quartile (Q_3) − Lower quartile (Q_1)

$$n = 30 \Rightarrow Q_1 = \frac{30}{4}^{th} \text{ value} = 7.5^{th} \text{ value}$$

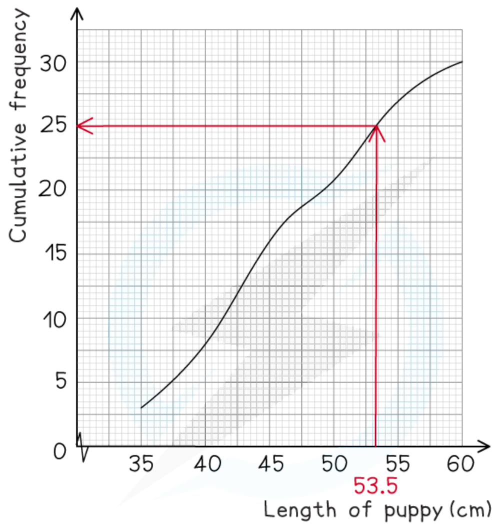
$$Q_3 = 3\left(\frac{30}{4}\right)^{th} \text{ value} = 22.5^{th} \text{ value}$$

Can be seen from the top of the graph



IQR = 51.25 − 39.5 = 11.75 cm

(iii) Read from the x−axis at 53.5 cm.



Approximately 25 puppies are less than 53.5, so 5 must be greater than 53.5 cm.

$$\frac{5}{30} = 16.67... \%$$

$$x \approx 16.7 \% \text{ (3.s.f.)}$$

Your notes

Examiner Tips and Tricks

- If you are asked to read values from your graph make sure you use a ruler and mark the lines on clearly to show where you took your readings from. Remember that the graph shows the accumulated frequencies so if you need only the frequency you may need to subtract the previous value.

Your notes

Histograms

What is a histogram?

- A histogram is similar to a bar chart but with some **key differences**
  - A histogram is for displaying **grouped continuous data** whereas a bar chart is for **discrete** or **qualitative data**
  - There will never be any gaps between the bars of adjacent groups in a histogram
  - Whilst in a bar chart the frequency is read from the **height** of the bar, in a histogram the height of the bar is the **frequency density**
- On a histogram **frequency density** is plotted on the y – axis
  - This allows a histogram to be plotted for **unequal class intervals**
  - It is particularly useful if data is spread out at either or both ends
- The **area** of each bar on a histogram will be **proportional** to the frequency in that class

What are the key features of a histogram?

- You will not be asked to draw a histogram but you may have to add information to one so you should make sure you are familiar with the process for drawing one
- **Step 1.** Always check that there are no gaps between the upper boundary of a class and the lower boundary of the next class
  - If there are gaps you will need to close them by changing the boundaries before carrying out any calculations
    - Consider whether the values are rounded or truncated before closing the gaps
- **Step 2.** Find the **class width** of each group by subtracting the lower boundary from the upper boundary
- **Step 3**. Calculate the **frequency density** for each group using the formula:

$$\text{frequency density} = k \times \frac{\text{frequency}}{\text{class width}}$$

- **Step 4.** The histogram will be drawn with the data values on the x – axis and frequency density on the y – axis
  - Remember that the scale on both axes must be even, although the class widths may be uneven
    - Both axes should be clearly labelled and units included on the x – axis
  - Most often, the bars will have different widths

- Occasionally you will be asked to add a **frequency polygon** to the histogram
  - This is done by joining up the midpoints at the top of each bar
  - You should not join up the first or last midpoint to the x – axis (it is not really a polygon!)

How do I interpret a histogram?

- It is important to remember that the y – axis does not tell us the frequency of each bar in the histogram
- The **area** of the bar gives information about the frequency
  - Most of the time, the frequency will be the area of the bar and is found by multiplying the class width by the frequency density
  - Occasionally, the frequency will be proportional to the area of the bar
    - Frequency = $k \times \text{area}$
    - In these cases more information will be given to help you find the value of *k*
- You may be asked to find the frequency of part of a bar within a histogram
  - Find the area of that section of the bar using any information you have already found out
  - You will need to have found the value of *k* first

What are frequency polygons?

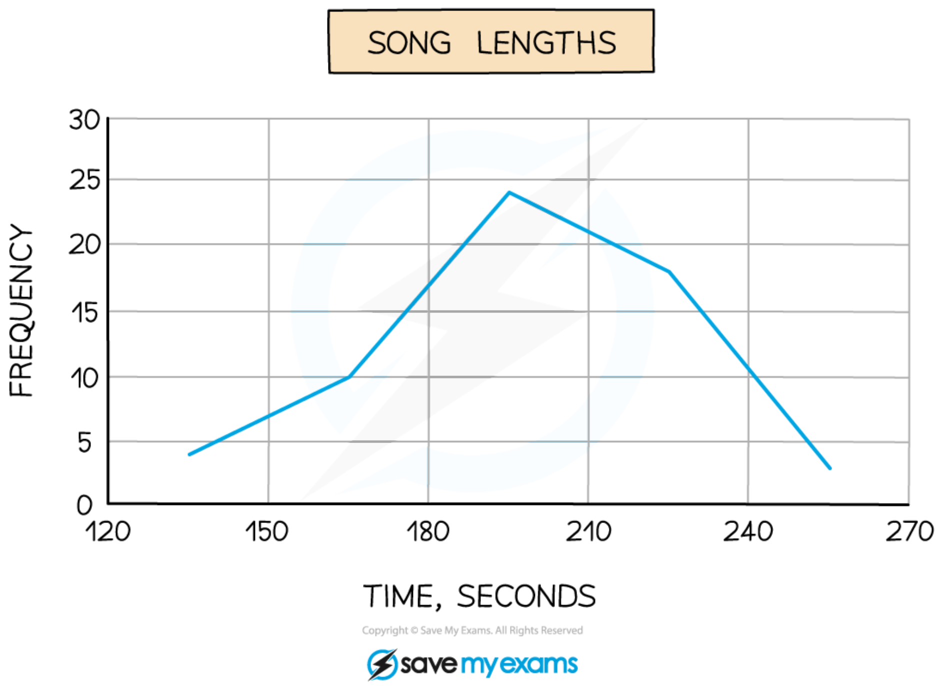
- **Frequency polygons** are a very simple way of showing frequencies/frequency densities for **continuous**, **grouped** data and give a quick guide to how frequencies change from one class to the next

What do I need to know about frequency polygons?

- Apart from plotting and joining up points with straight lines there are 2 rules for frequency polygons:
  - Plot points at the **MIDPOINT** of class intervals
  - Unless one of the frequencies/densities is 0 do not join the frequency polygon to the x-axis, and do not join the first point to the last one
- The result is not actually a polygon but more of an open one that 'floats' in mid-air!
- You may be asked to draw a frequency polygon and/or use it to make comments and compare data

Your notes

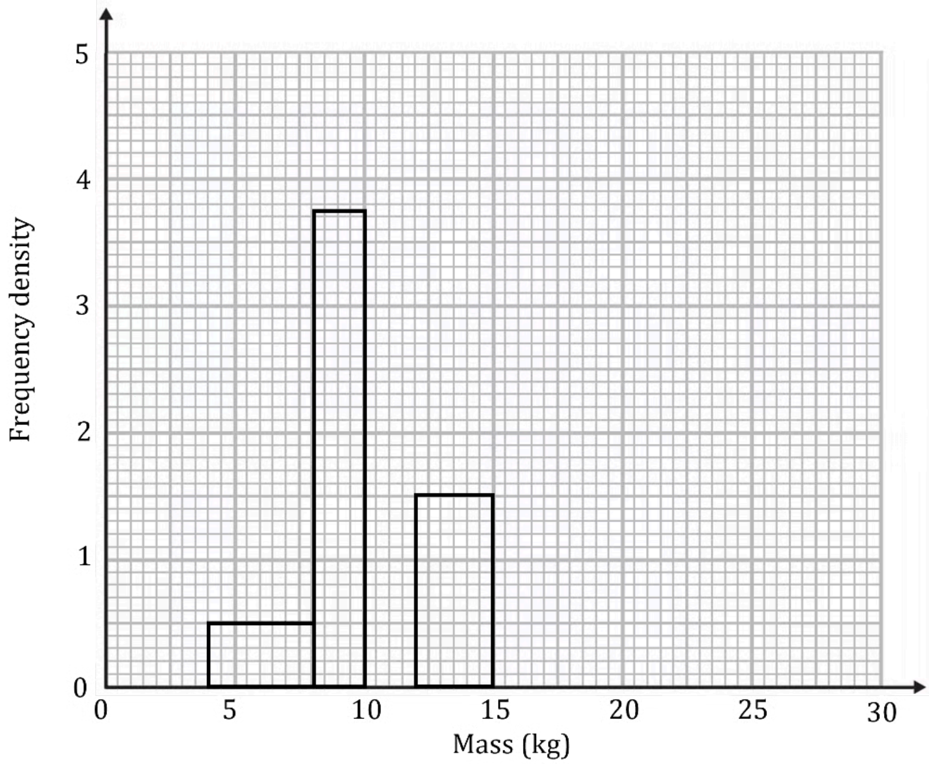
Your notes



Worked Example

The table below and its corresponding histogram show the mass, in kg, of some new born bottlenose dolphins.

Mass, m kg	Frequency
$4 \le m < 8$	4
$8 \le m < 10$	15
$10 \le m < 12$	19
$12 \le m < 15$	
$15 \le m < 30$	6



(i) Use the histogram to find the value of k in the formula

$$\text{frequency density} = k \times \frac{\text{frequency}}{\text{class width}}$$

(ii) Estimate the number of dolphins whose weight is greater than 13 kg.

Answer:

(i) Start by finding the frequency density in terms of k for at least the first group:

$$\text{Frequency density} = k \times \frac{\text{frequency}}{\text{class width}}$$

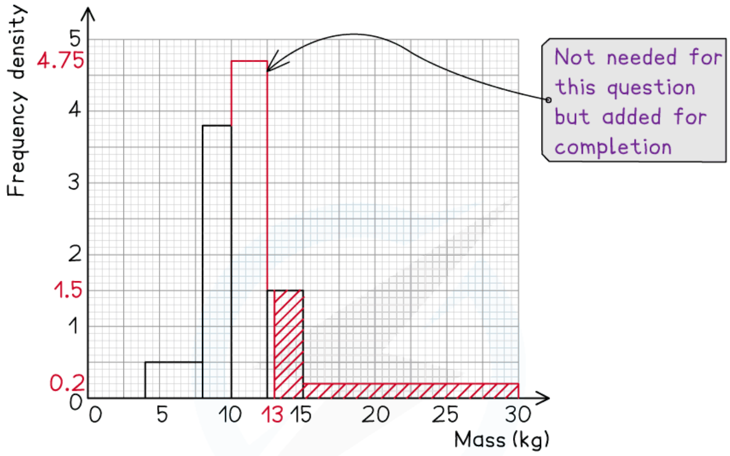
Mass, m kg	Frequency	Class width	Frequency density
$4 \leqslant m < 8$	4	4	$k\left(\frac{4}{4}\right) = k$
$8 \leqslant m < 10$	15	2	$k\left(\frac{15}{2}\right) = 7.5k$
$10 \leqslant m < 12$	19	2	$k\left(\frac{19}{2}\right) = 9.5k$
$12 \leqslant m < 15$		3	
$15 \leqslant m < 30$	6	15	$k\left(\frac{6}{15}\right) = 0.4k$

From the first bar we can see that the height gives the frequency density for the first group as 0.5

$k = 0.5$

(ii) Draw the bar for the class $15 \leqslant m < 30$

$0.4\,k = 0.4\,(0.5) = 0.2$



Use the histogram to estimate the number of dolphins in the group $13 < m \leqslant 15$

$$\text{Frequency density} = k \times \frac{\text{frequency}}{\text{class width}}$$

$$1.5 = 0.5 \times \frac{f}{2}$$ 15 − 13

There are an estimated 6 dolphins in the group $13 < m \leqslant 15$
From the table we can see there are 6 dolphins in the group $15 < m \leqslant 30$

An estimated 12 dolphins are greater than 13 kg

Examiner Tips and Tricks

- Remember that the area of a bar in a histogram is not always the frequency itself but could be proportional to the frequency. Look carefully at the scales on the axes, it will rarely be a simple 1 unit to 1 square.

Your notes

Outliers

What are outliers?

- Outliers are extreme data values that do not fit with the general pattern of the data
- They can come from one or two extreme events or from mistakes in the data collection
- Outliers will affect some statistics that are calculated from the data
  - They can have a big effect on the mean, but not on the median or usually the mode
  - The range will be completely changed by a single outlier, but the interquartile range will not be affected
  - When calculating the mean or the range it is important to decide whether the outlier(s) should be included in the calculations
    - The question will tell you whether to include the outliers or not
    - You may have to decide which value is the outlier to be removed
    - In general outliers are included if they are a valid piece of data and excluded if it is likely that they are erroneous

How are outliers calculated?

- Most of the time within this syllabus the outliers will be a particular distance either side of the **interquartile range**
  - The most common way to calculate an outlier will be using the formulae:
    - A value that is less than $Q_1 - k$(interquartile range)
    - A value that is greater than $Q_3 + k$(interquartile range)
    - k is a constant that will be given to you in the exam, commonly k=1.5
- Outliers could also be situated a number of **standard deviations** away from the **mean**
  - The most common way to calculate an outlier will be using the formulae
    - A value that is less than $\bar{x} - k\sigma$
    - A value that is greater than $\bar{x} + k\sigma$
    - k is a constant that will be given to you in the exam, commonly $k = 2$

How are outliers represented on box plots?

- On a box plot an outlier is represented as a cross either side of the maximum or minimum value

- If the maximum or minimum value is discovered to be an outlier, the **new** maximum or minimum value will need to be found for the box plot
  - If the data value just above the minimum or just below the maximum is known, this will become the new value
  - If the data value is not known, the new minimum or maximum will become the outlier boundary

Your notes

Worked Example

The ages, in years, of a number of children attending a birthday party are given below:

2, 7, 5, 4, 8, 4, 6, 5, 5, 29, 2, 5, 13,

An outlier is defined as an observation that falls more than $1.5 \times$ the interquartile range above the upper quartile or below the lower quartile

(i) Identify any outliers within the data set.

(ii) Decide which values (if any) should be removed, justify your answer.

Answer:

(i) Start by putting the data in order of size

2, 2, 4, 4, 5, 5, 5, 5, 6, 7, 8, 13, 29

LQ Median UQ

Find the interquartile range:

$Q_1 = 4, \quad Q_3 = 7.5$

$IQR = Q_3 - Q_1 = 7.5 - 4 = 3.5$

Find the upper and lower bound for the outliers:

Lower bound $= Q_1 - 1.5\,(IQR)$

$4 - 1.5\,(3.5) = -1.25$ So there are no outliers on the lower tail.

Upper bound $= Q_3 - 1.5\,(IQR)$

$7.5 + 1.5\,(3.5) = 12.75$ So there are two outliers on the upper tail.

Outliers are 13 and 29

(ii) The number 29 is clearly an error because this is not a child.

The number 13 may be an error but is unlikely. We should investigate further before removing it from the data set.

Remove the number 29 only

Examiner Tips and Tricks

- Read the question carefully to determine which type of outlier you should be finding and to make sure you are using the correct method.

Your notes

Interpreting Data

You may be asked to comment on how statistics could affect data or how removing or adding a new piece of data could change statistics you have calculated. You may also be asked to compare two data sets of a similar context.

Analysing statistics calculated from a data set:

- You could be asked to use some statistics you have calculated, such as the median or interquartile range, to make decisions about the set of data you were analysing
  - Writing about the mean or median gives information about where the data is located
  - Writing about the range, interquartile range, standard deviation or variance gives information about how spread out the data is
    - A lower value means the data set is more consistent
    - A greater value means the data set is more spread out, or varied
- When commenting on a data set you should usually write about both a measure of location and a measure of spread
- You should pair the median with the range or interquartile range and the mean with the standard deviation or variance
  - Use the mean and standard deviation/variance when the data is roughly symmetrical and does not contain outliers
  - Use the median and interquartile range when the data contains outliers
- You should always write your analysis in the context of the data
  - Sometimes a lower mean/median is better: time taken to complete a puzzle
  - Sometimes a higher mean/median is better: score on a test

Changing a data set:

- Sometimes it might be discovered that a value was omitted from a data set and needs to be added in
- Similarly a data value could be found to be an error and will need to be removed (**cleaned**) from the data set
- It is important to be aware of how the **measures of location** and **spread** may change when the data is added or removed
  - Adding or removing a data value to the set could change the mean, median and quartiles
  - How each statistics changes will depend on where the data value lies within the data set

- Adding a data point below the mean, or removing one from above will cause the mean to decrease
- Adding a data point above the mean, or removing one from below will cause the mean to increase
- The median and quartiles may or may not change depending on the data value, you should always check these cases individually

- Adding or removing extreme values will change the value of the mean by a lot but affect the median in the same way as any other value

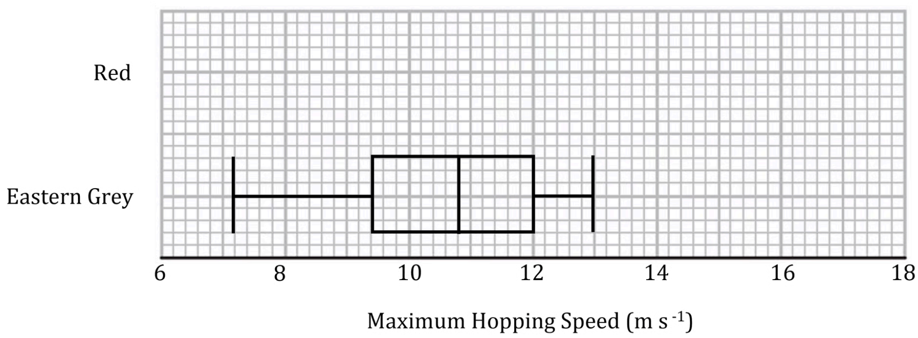
Comparing two data sets

- When comparing two data sets you must comment on both:
  - A **measure of location** such as the mean, median or mode and
  - A **measure of spread** such as the range, interquartile range, variance or standard deviation
    - If you comment on the mean as the measure of location you should use the standard deviation or variance as the measure of spread
    - If you comment on the median as the measure of location you should use the interquartile range as the measure of spread
- You should use information about the data to decide which measure of location and measure of spread is the best to use
  - If data contains extreme values then it is best to use the median and interquartile range to compare the data sets
  - Extreme values can cause the mean to be an unreliable statistic
- It is common to be asked to compare two data sets that are represented as two **box plots** on the same scale
  - You should write about both the median and the interquartile range
- When writing a comparison about the data you should always write in **context**

Worked Example

The diagram below shows a box plot of the average hopping speeds, in m s^{-1}, of some Eastern Grey kangaroos living in an Australian nature reserve.

Your notes



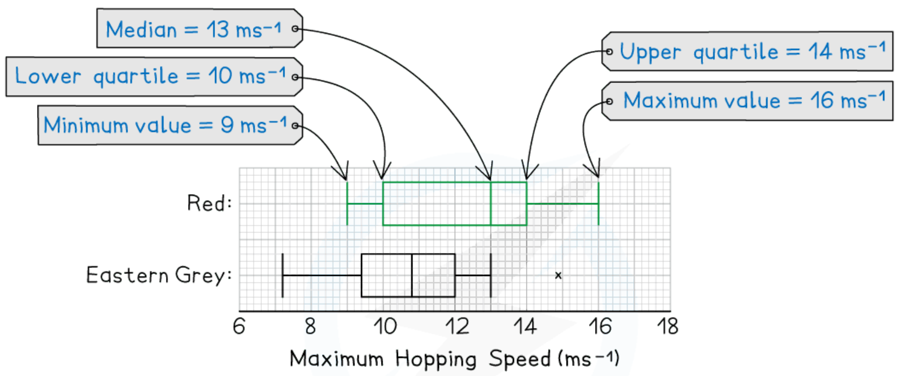
The maximum hopping speeds, in m s^{-1}, of some Red Kangaroos living in the nature reserve are summarised below. There are no outliers in this sample.

Lower quartile: 10 Median: 13 Upper quartile: 14

Minimum value: 9 Maximum value: 16

Draw a box plot on the grid to represent the data for the Red Kangaroos and compare the distribution of the hopping speeds for the Red and Eastern Grey kangaroos.

Answer:



To compare the distributions, comment on the median (to compare location) and the interquartile range (to compare spread).

The median hopping speed for the Red Kangaroos is 2.2 ms^{-1} faster than median hopping speed for the Eastern Grey Kangaroos. The interquartile range for the Red Kangaroos is 1.4 ms^{-1} greater than for the Eastern Grey Kangaroos. This suggests the Red Kangaroos are faster but with more varied hopping speeds than the Eastern Greys.

Examiner Tips and Tricks

- Remember to always write about data sets and their distributions in the context of the question. The number of marks available in comparison question are often an indication of how much you should say in the answer.

Your notes

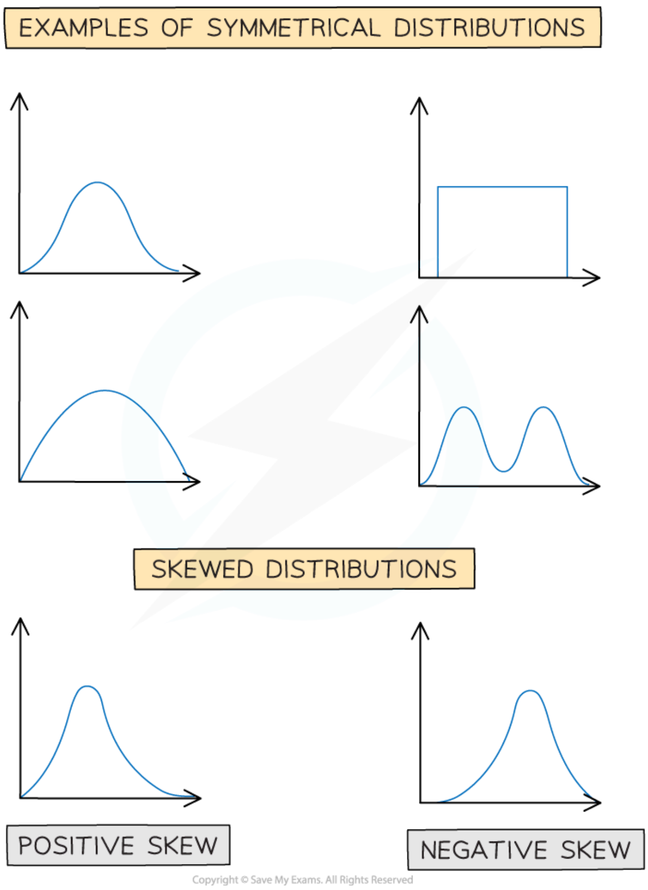
Your notes

Skewness

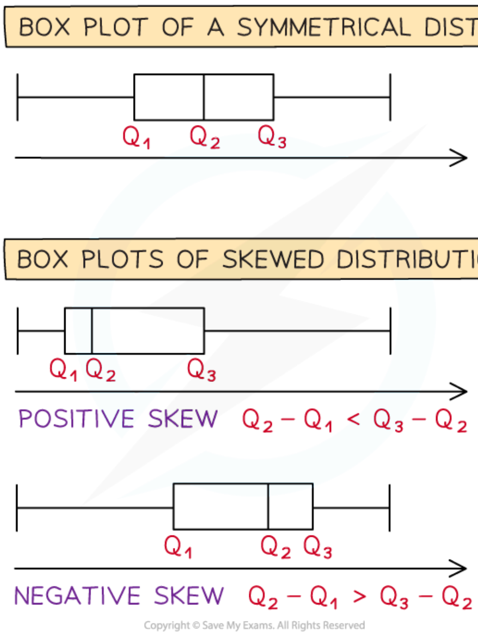
The distribution of a data set is either symmetrical or it has skewness.

What is skewness?

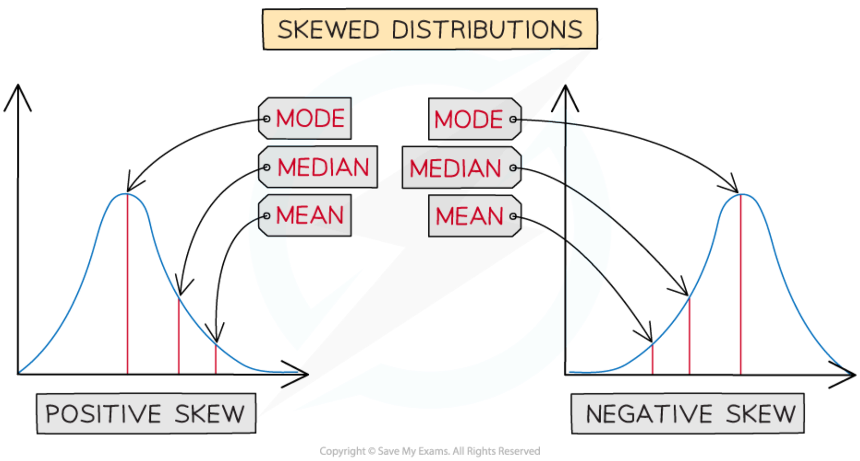
- Skewness describes the way in which data in a non – symmetrical distribution is leaning
- A distribution that has its tail on the right side has **positive skew**
- A distribution that has its tail on the left side has **negative skew**



- If the distribution is shown on a box plot looking at the difference between the quartiles can help decide how it is skewed
  - If the median is closer to the lower quartile then the distribution has **positive skew**
    - $Q_3 - Q_2 > Q_2 - Q_1$
  - If the median is closer to the upper quartile then the distribution has **negative skew**
    - $Q_3 - Q_2 < Q_2 - Q_1$



- Looking at the values of the statistics can help you decide whether distribution is positively skewed or negatively skewed
  - In a positively skewed distribution
    - Mode < median < mean
  - In a negatively skewed distribution
    - Mean < median < mode



Examiner Tips and Tricks

- It can help to comment on skewness when asked to compare distributions.

Your notes