CHAPTER 11

Statistics

11.1 Revision EMBJZ

Measures of central tendency EMBK2

The mean and median of a data set both give an indication where the centre of the data distribution is located. The **mean**, or average, is calculated as

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where the x_i are the data and n is the number of data. We read $\overline{x}$ as "x bar".

The **median** is the middle value of an ordered data set. To find the median, we first sort the data and then pick out the value in the middle of the sorted list. If the middle is in between two values, the median is the average of those two values.

See video: 23C9 at www.everythingmaths.co.za

Worked example 1: Computing measures of central tendency

QUESTION

Compute the mean and median of the following data set:

72,5 ; 92,6 ; 15,6 ; 53,0 ; 86,4 ; 89,9 ; 90,9 ; 21,7 ; 46,0 ; 4,1 ; 51,7 ; 2,2

SOLUTION

Step 1: Compute the mean

Using the formula for the mean, we first compute the sum of the values and then divide by the number of values.

$$\begin{aligned}\overline{x} &= \frac{626,6}{12} \\ &\approx 52,22\end{aligned}$$

Step 2: Compute the median

To find the median, we first have to sort the data:

2,2 ; 4,1 ; 15,6 ; 21,7 ; 46,0 ; 51,7 ; 53,0 ; 72,5 ; 86,4 ; 89,9 ; 90,9 ; 92,6

Since there are an even number of values, the median will lie between two values. In this case, the two values in the middle are 51,7 and 53,0. Therefore the median is 52,35.

Measures of dispersion tell us how spread out a data set is. If a measure of dispersion is small, the data are clustered in a small region. If a measure of dispersion is large, the data are spread out over a large region.

The **range** is the difference between the maximum and minimum values in the data set.

The **inter-quartile range** is the difference between the first and third quartiles of the data set. The quartiles are computed in a similar way to the median. The median is halfway into the ordered data set and is sometimes also called the second quartile. The first quartile is one quarter of the way into the ordered data set; whereas the third quartile is three quarters of the way into the ordered data set.

See video: 23CB at www.everythingmaths.co.za

Worked example 2: Range and inter-quartile range

QUESTION

Determine the range and the inter-quartile range of the following data set.

$$14\ ;\ 17\ ;\ 45\ ;\ 20\ ;\ 19\ ;\ 36\ ;\ 7\ ;\ 30\ ;\ 8$$

SOLUTION

Step 1: Sort the values in the data set

To determine the range we need to find the minimum and maximum values in the data set. To determine the inter-quartile range we need to compute the first and third quartiles of the data set. For both of these requirements, it is easier to order the data set first.

The sorted data set is

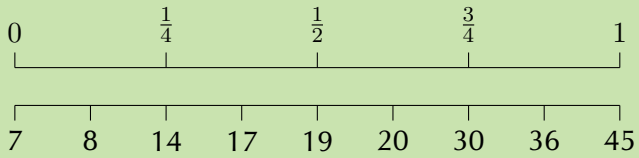
$$7\ ;\ 8\ ;\ 14\ ;\ 17\ ;\ 19\ ;\ 20\ ;\ 30\ ;\ 36\ ;\ 45$$

Step 2: Find the minimum, maximum and range

The minimum value is the first value in the ordered data set, namely 7. The maximum is the last value in the ordered data set, namely 45. The range is the difference between the minimum and maximum: $45 - 7 = 38$.

Step 3: Find the quartiles and inter-quartile range

The diagram below shows how we find the quartiles one quarter, one half and three quarters of the way into the ordered list of values.



From this diagram we can see that the first quartile is at a value of 14, the second quartile (median) is at a value of 19 and the third quartile is at a value of 30.

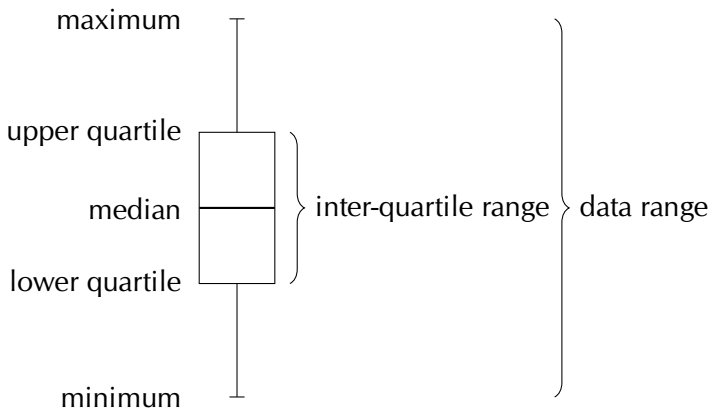
The inter-quartile range is the difference between the first and third quartiles. The first quartile is 14 and the third quartile is 30. Therefore the inter-quartile range is $30 - 14 = 16$.

Five number summary EMBK4

The **five number summary** combines a measure of central tendency, namely the median, with measures of dispersion, namely the range and the inter-quartile range. This gives a good overview of the overall data distribution. More precisely, the five number summary is written in the following order:

- minimum;
- first quartile;
- median;
- third quartile;
- maximum.

The five number summary is often presented visually using a **box and whisker diagram**. A box and whisker diagram is shown below, with the positions of the five relevant numbers labelled. Note that this diagram is drawn vertically, but that it may also be drawn horizontally.



See video: 23CC at www.everythingmaths.co.za

Worked example 3: Five number summary

QUESTION

Draw a box and whisker diagram for the following data set:

$$1{,}25\ ;\ 1{,}5\ ;\ 2{,}5\ ;\ 2{,}5\ ;\ 3{,}1\ ;\ 3{,}2\ ;\ 4{,}1\ ;\ 4{,}25\ ;\ 4{,}75\ ;\ 4{,}8\ ;\ 4{,}95\ ;\ 5{,}1$$

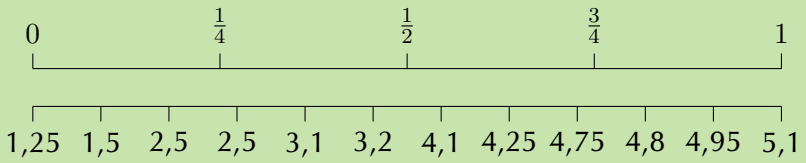
SOLUTION

Step 1: Determine the minimum and maximum

Since the data set is already ordered, we can read off the minimum as the first value (1,25) and the maximum as the last value (5,1).

Step 2: Determine the quartiles

There are 12 values in the data set.



Using the figure above we can see that the median is between the sixth and seventh values, making it.

$$\frac{3{,}2 + 4{,}1}{2} = 3{,}65$$

The first quartile lies between the third and fourth values, making it

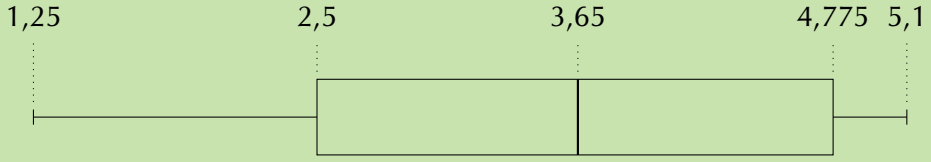
$$Q_1 = \frac{2{,}5 + 2{,}5}{2} = 2{,}5$$

The third quartile lies between the ninth and tenth values, making it

$$Q_3 = \frac{4{,}75 + 4{,}8}{2} = 4{,}775$$

Step 3: Draw the box and whisker diagram

We now have the five number summary as (1,25; 2,5; 3,65; 4,775; 5,1). The box and whisker diagram representing the five number summary is given below.



Exercise 11 – 1: Revision

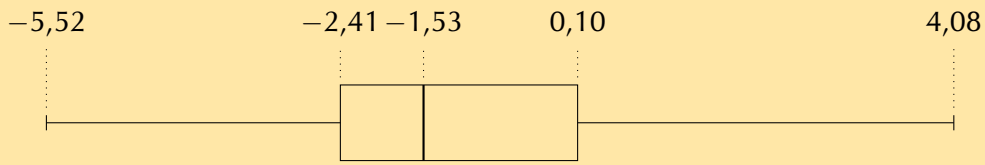
1. For each of the following data sets, compute the mean and all the quartiles. Round your answers to one decimal place.

   a) $-3{,}4\ ;\ -3{,}1\ ;\ -6{,}1\ ;\ -1{,}5\ ;\ -7{,}8\ ;\ -3{,}4\ ;\ -2{,}7\ ;\ -6{,}2$

   b) $-6\ ;\ -99\ ;\ 90\ ;\ 81\ ;\ 13\ ;\ -85\ ;\ -60\ ;\ 65\ ;\ -49$

   c) $7\ ;\ 45\ ;\ 11\ ;\ 3\ ;\ 9\ ;\ 35\ ;\ 31\ ;\ 7\ ;\ 16\ ;\ 40\ ;\ 12\ ;\ 6$

2. Use the following box and whisker diagram to determine the range and inter-quartile range of the data.



3. Draw the box and whisker diagram for the following data.

   $0{,}2\ ;\ -0{,}2\ ;\ -2{,}7\ ;\ 2{,}9\ ;\ -0{,}2\ ;\ -4{,}2\ ;\ -1{,}8\ ;\ 0{,}4\ ;\ -1{,}7\ ;\ -2{,}5\ ;\ 2{,}7\ ;\ 0{,}8\ ;\ -0{,}5$

Think you got it? Get this answer and more practice on our Intelligent Practice Service

1a. 23CD 1b. 23CF 1c. 23CG 2. 23CH 3. 23CJ

 **www.everythingmaths.co.za**

 **m.everythingmaths.co.za**

11.2 Histograms EMBK5

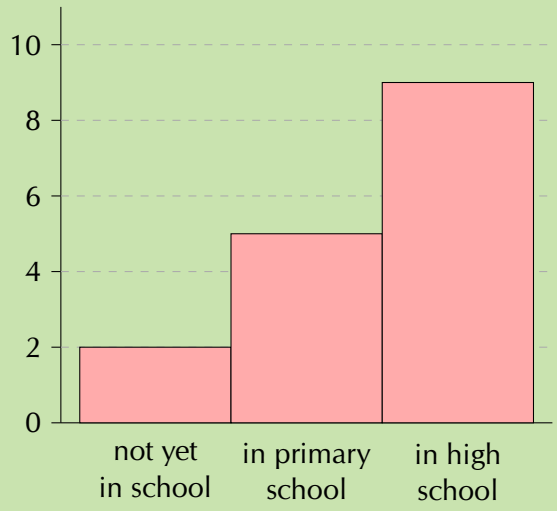
A histogram is a graphical representation of how many times different, mutually exclusive events are observed in an experiment. To interpret a histogram, we find the events on the x-axis and the counts on the y-axis. Each event has a rectangle that shows what its count (or frequency) is.

See video: 23CK at www.everythingmaths.co.za

Worked example 4: Reading histograms

QUESTION

Use the following histogram to determine the events that were recorded and the relative frequency of each event. Summarise your answer in a table.



SOLUTION

Step 1: Determine the events

The events are shown on the x-axis. In this example we have "not yet in school", "in primary school" and "in high school".

Step 2: Read off the count for each event

The counts are shown on the y-axis and the height of each rectangle shows the frequency for each event.

- not yet in school: 2
- in primary school: 5
- in high school: 9

Step 3: Calculate relative frequency

The relative frequency of an event in an experiment is the number of times that the event occurred divided by the total number of times that the experiment was completed. In this example we add up the frequencies for all the events to get a total frequency of 16. Therefore the relative frequencies are:

- not yet in school: $\frac{2}{16} = \frac{1}{8}$
- in primary school: $\frac{5}{16}$
- in high school: $\frac{9}{16}$

Step 4: Summarise

Event	Count	Relative frequency
not yet in school	2	$\frac{1}{8}$
in primary school	5	$\frac{5}{16}$
in high school	9	$\frac{9}{16}$

To draw a histogram of a data set containing numbers, the numbers first have to be grouped. Each group is defined by an interval. We then count how many times numbers from each group appear in the data set and draw a histogram using the counts.

Worked example 5: Draw a histogram

QUESTION

The following data represent the heights of 16 adults in centimetres.

$$162\ ;\ 168\ ;\ 177\ ;\ 147\ ;\ 189\ ;\ 171\ ;\ 173\ ;\ 168$$
$$178\ ;\ 184\ ;\ 165\ ;\ 173\ ;\ 179\ ;\ 166\ ;\ 168\ ;\ 165$$

Divide the data into 5 equal length intervals between 140 cm and 190 cm and draw a histogram.

SOLUTION

Step 1: Determine intervals

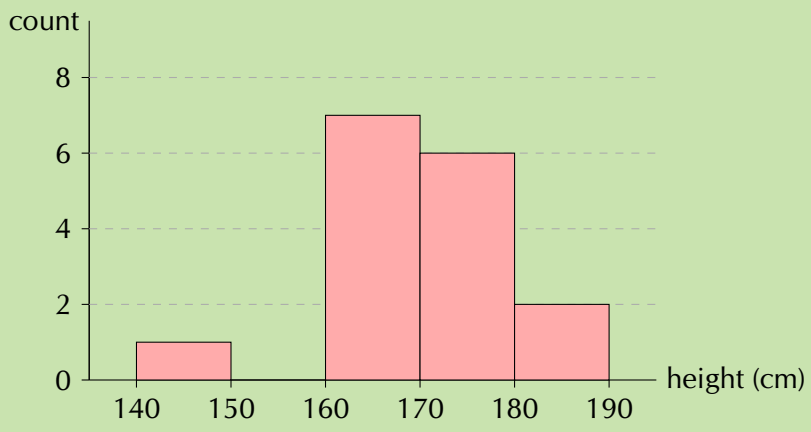
To have 5 intervals of the same length between 140 and 190, we need and interval length of 10. Therefore the intervals are (140; 150]; (150; 160]; (160; 170]; (170; 180]; and (180; 190].

Step 2: Count data

The following table summarises the number of data values in each of the intervals.

Interval	(140; 150]	(150; 160]	(160; 170]	(170; 180]	(180; 190]
Count	1	0	7	6	2

Step 3: Draw the histogram



A frequency polygon is sometimes used to represent the same information as in a histogram. A frequency polygon is drawn by using line segments to connect the middle of the top of each bar in the histogram. This means that the frequency polygon connects the coordinates at the centre of each interval and the count in each interval.

Worked example 6: Drawing a frequency polygon

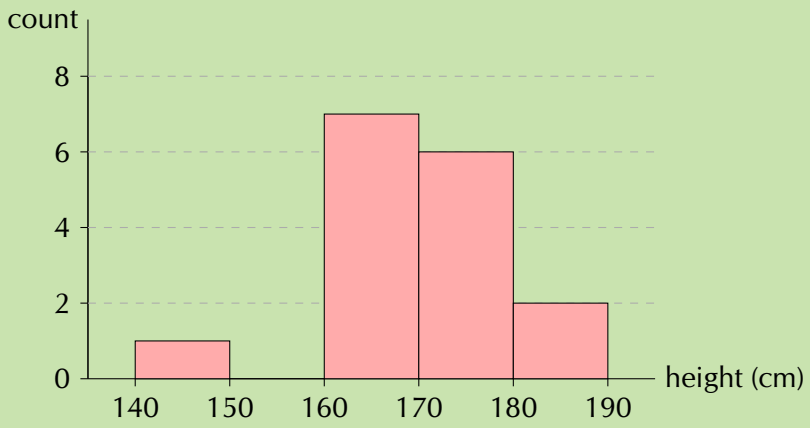
QUESTION

Use the histogram from the previous example to draw a frequency polygon of the same data.

SOLUTION

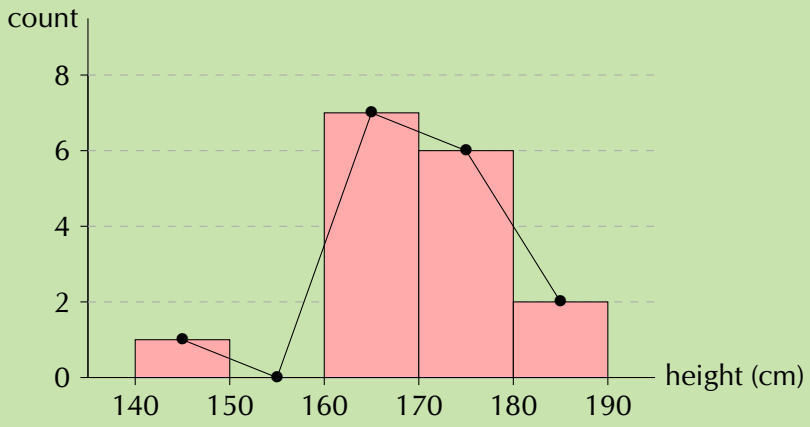
Step 1: Draw the histogram

We already know that the histogram looks like this:



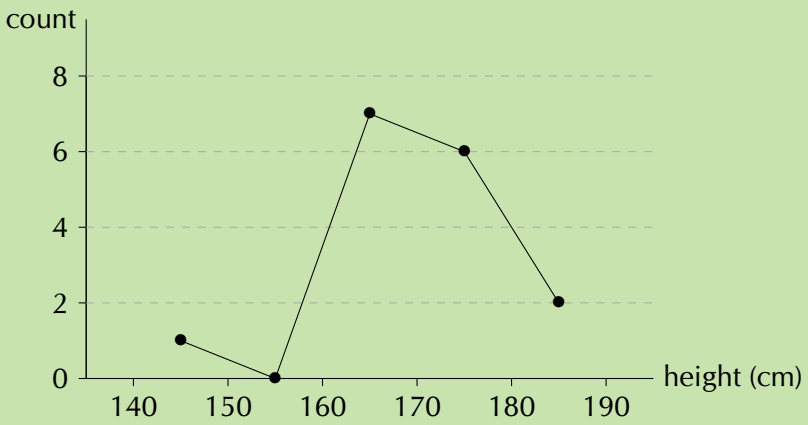
Step 2: Connect the tops of the rectangles

When we draw line segments between the tops of the rectangles in the histogram, we get the following picture:



Step 3: Draw final frequency polygon

Finally, we remove the histogram to show only the frequency polygon.



Frequency polygons are particularly useful for comparing two data sets. Comparing two histograms would be more difficult since we would have to draw the rectangles of the two data sets on top of each other. Because frequency polygons are just lines, they do not pose the same problem.

Worked example 7: Drawing frequency polygons

QUESTION

Here is another data set of heights, this time of Grade 11 learners.

132 ; 132 ; 156 ; 147 ; 162 ; 168 ; 152 ; 174
141 ; 136 ; 161 ; 148 ; 140 ; 174 ; 174 ; 162

Draw the frequency polygon for this data set using the same interval length as in the previous example. Then compare the two frequency polygons on one graph to see the differences between the distributions.

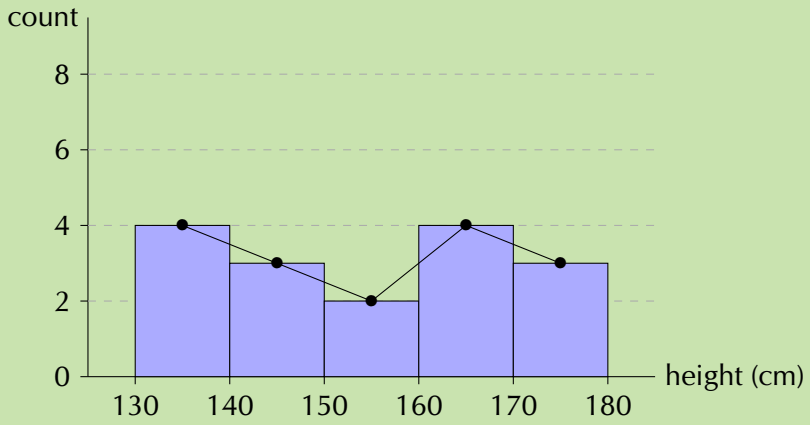
SOLUTION

Step 1: Frequency table

We first create the table of counts for the new data set.

Interval	(130; 140]	(140; 150]	(150; 160]	(160; 170]	(170; 180]
Count	4	3	2	4	3

Step 2: Draw histogram and frequency polygon



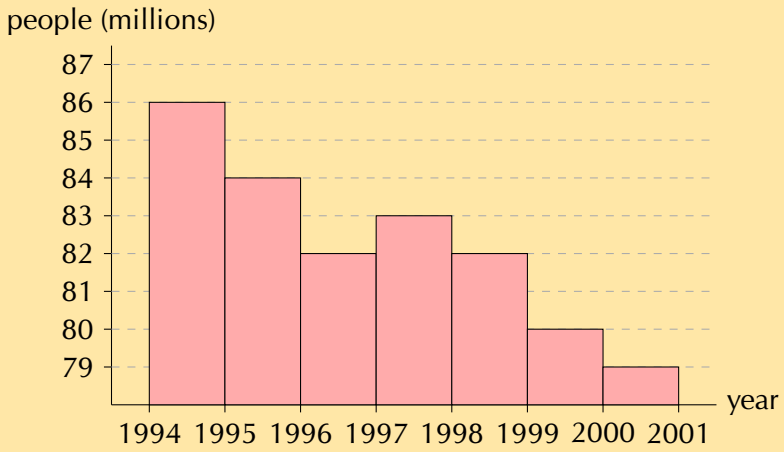
Step 3: Compare frequency polygons

We draw the two frequency polygons on the same axes. The red line indicates the distribution over heights for adults and the blue line, for Grade 11 learners.

From this plot we can easily see that the heights for Grade 11 learners are distributed more towards the left (shorter) than adults. The learner heights also seem to be more evenly distributed between 130 and 180 cm, whereas the adult heights are mostly between 160 and 180 cm.

Exercise 11 – 2: Histograms

1. Use the histogram below to answer the following questions. The histogram shows the number of people born around the world each year. The ticks on the x-axis are located at the start of each year.



   a) How many people were born between the beginning of 1994 and the beginning of 1996?

   b) Is the number people in the world population increasing or decreasing? (Ignore the rate at which people are dying for this question.)

   c) How many more people were born in 1994 than in 1997?

2. In a traffic survey, a random sample of 50 motorists were asked the distance (d) they drove to work daily. The results of the survey are shown in the table below. Draw a histogram to represent the data.

d	$0 < d \leq 10$	$10 < d \leq 20$	$20 < d \leq 30$	$30 < d \leq 40$	$40 < d \leq 50$
f	9	19	15	5	4

3. Below is data for the prevalence of HIV in South Africa. HIV prevalence refers to the percentage of people between the ages of 15 and 49 who are infected with HIV.

year	2002	2003	2004	2005	2006	2007	2008	2009
prevalence (%)	17,7	18,0	18,1	18,1	18,1	18,0	17,9	17,9

   Draw a frequency polygon of this data set.

Think you got it? Get this answer and more practice on our Intelligent Practice Service

1. 23CM 2. 23CN 3. 23CP

 www.everythingmaths.co.za  m.everythingmaths.co.za

Cumulative histograms, also known as ogives, are graphs that can be used to determine how many data values lie above or below a particular value in a data set. The cumulative frequency is calculated from a frequency table, by adding each frequency to the total of the frequencies of all data values before it in the data set. The last value for the cumulative frequency will always be equal to the total number of data values, since all frequencies will already have been added to the previous total.

An ogive is drawn by

- plotting the beginning of the first interval at a y-value of zero;
- plotting the end of every interval at the y-value equal to the cumulative count for that interval; and
- connecting the points on the plot with straight lines.

In this way, the end of the final interval will always be at the total number of data since we will have added up across all intervals.

Worked example 8: Cumulative frequencies and ogives

QUESTION

Determine the cumulative frequencies of the following grouped data and complete the table below. Use the table to draw an ogive of the data.

Interval	Frequency	Cumulative frequency
$10 < n \leq 20$	5	
$20 < n \leq 30$	7	
$30 < n \leq 40$	12	
$40 < n \leq 50$	10	
$50 < n \leq 60$	6	

SOLUTION

Step 1: Compute cumulative frequencies

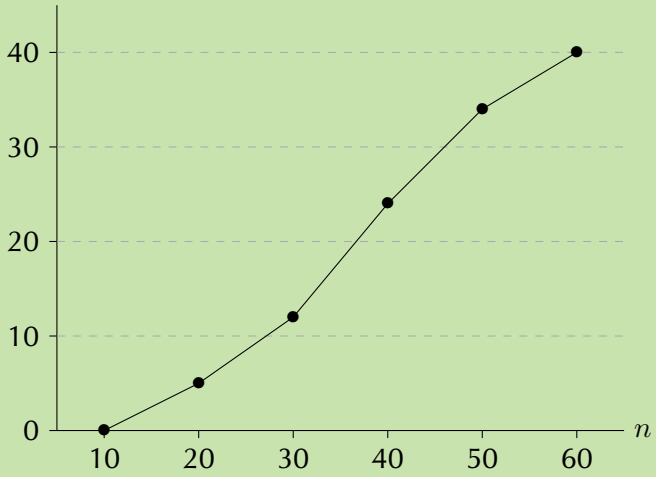
To determine the cumulative frequency, we add up the frequencies going down the table. The first cumulative frequency is just the same as the frequency, because we are adding it to zero. The final cumulative frequency is always equal to the sum of all the frequencies. This gives the following table:

Interval	Frequency	Cumulative frequency
$10 < n \leq 20$	5	5
$20 < n \leq 30$	7	12
$30 < n \leq 40$	12	24
$40 < n \leq 50$	10	34
$50 < n \leq 60$	6	40

Step 2: Plot the ogive

The first coordinate in the plot always starts at a y-value of 0 because we always start from a count of zero. So, the first coordinate is at $(10; 0)$ — at the beginning of the first interval. The second coordinate is at the end of the first interval (which is also the beginning of the second interval) and at the first cumulative count, so $(20; 5)$. The third coordinate is at the end of the second interval and at the second cumulative count, namely $(30; 12)$, and so on.

Computing all the coordinates and connecting them with straight lines gives the following ogive.



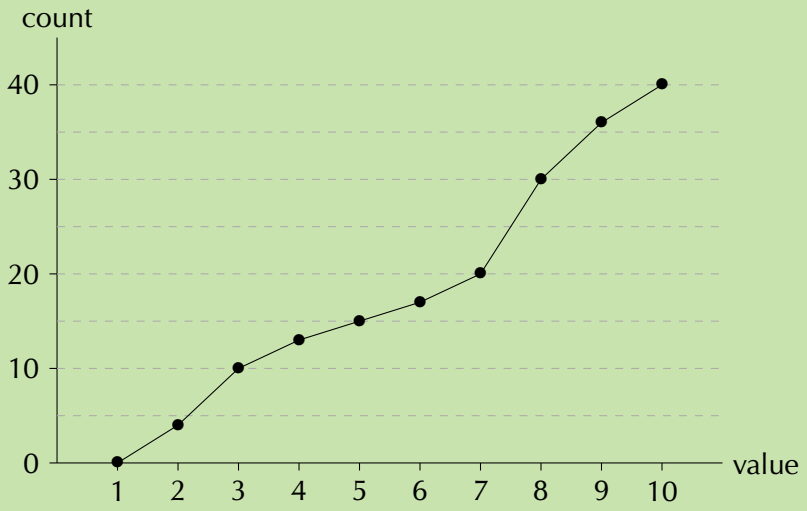
Ogives do look similar to frequency polygons, which we saw earlier. The most important difference between them is that an ogive is a plot of **cumulative** values, whereas a frequency polygon is a plot of the values themselves. So, to get from a frequency polygon to an ogive, we would add up the counts as we move from left to right in the graph.

Ogives are useful for determining the median, percentiles and five number summary of data. Remember that the median is simply the value in the middle when we order the data. A quartile is simply a quarter of the way from the beginning or the end of an ordered data set. With an ogive we already know how many data values are above or below a certain point, so it is easy to find the middle or a quarter of the data set.

Worked example 9: Ogives and the five number summary

QUESTION

Use the following ogive to compute the five number summary of the data. Remember that the five number summary consists of the minimum, all the quartiles (including the median) and the maximum.



SOLUTION

Step 1: Find the minimum and maximum

The minimum value in the data set is 1 since this is where the ogive starts on the horizontal axis. The maximum value in the data set is 10 since this is where the ogive stops on the horizontal axis.

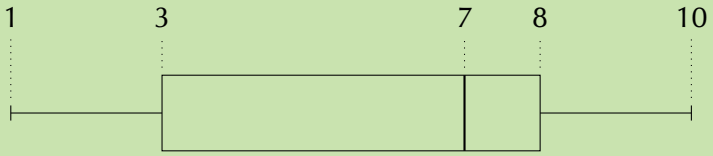
Step 2: Find the quartiles

The quartiles are the values that are $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{3}{4}$ of the way into the ordered data set. Here the counts go up to 40, so we can find the quartiles by looking at the values corresponding to counts of 10, 20 and 30. On the ogive a count of

- 10 corresponds to a value of 3 (first quartile);
- 20 corresponds to a value of 7 (second quartile); and
- 30 corresponds to a value of 8 (third quartile).

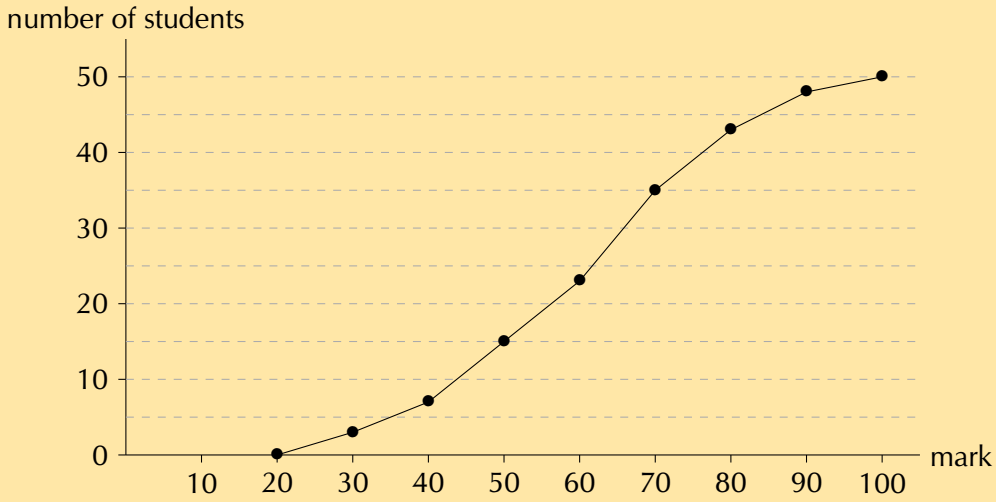
Step 3: Write down the five number summary

The five number summary is $(1; 3; 7; 8; 10)$. The box-and-whisker plot of this data set is given below.



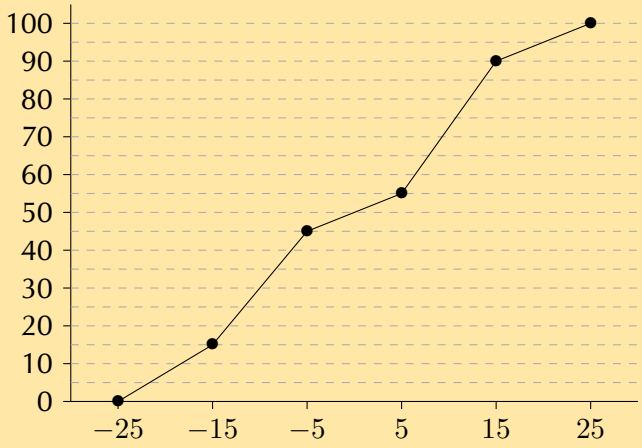
Exercise 11 – 3: Ogives

1. Use the ogive to answer the questions below. Note that marks are given as a percentage.



   a) How many students got between 50% and 70%?
   b) How many students got at least 70%?
   c) Compute the average mark for this class, rounded to the nearest integer.

2. Draw the histogram corresponding to this ogive.



3. The following data set lists the ages of 24 people.

   2; 5; 1; 76; 34; 23; 65; 22; 63; 45; 53; 38

   4; 28; 5; 73; 79; 17; 15; 5; 34; 37; 45; 56

   Use the data to answer the following questions.

   a) Using an interval width of 8 construct a cumulative frequency plot.
   b) How many are below 30?
   c) How many are below 60?
   d) Giving an explanation state below what value the bottom 50% of the ages fall.
   e) Below what value do the bottom 40% fall?
   f) Construct a frequency polygon.

4. The weights of bags of sand in grams is given below (rounded to the nearest tenth):

   50,1; 40,4; 48,5; 29,4; 50,2; 55,3; 58,1; 35,3; 54,2; 43,5

   60,1; 43,9; 45,3; 49,2; 36,6; 31,5; 63,1; 49,3; 43,4; 54,1

   a) Decide on an interval width and state what you observe about your choice.
   b) Give your lowest interval.
   c) Give your highest interval.
   d) Construct a cumulative frequency graph and a frequency polygon.
   e) Below what value do 53% of the cases fall?
   f) Below what value of 60% of the cases fall?

Think you got it? Get this answer and more practice on our Intelligent Practice Service

1. 23CQ 2. 23CR 3. 23CS 4. 23CT

 **www.everythingmaths.co.za**  **m.everythingmaths.co.za**

11.4 Variance and standard deviation EMBK8

Measures of central tendency (mean, median and mode) provide information on the data values at the centre of the data set. Measures of dispersion (quartiles, percentiles, ranges) provide information on the spread of the data around the centre. In this section we will look at two more measures of dispersion called the **variance** and the **standard deviation**.

See video: 23CV at www.everythingmaths.co.za

Variance EMBK9

DEFINITION: *Variance*

Let a population consist of n elements, $\{x_1; x_2; \ldots; x_n\}$. Write the mean of the data as $\overline{x}$.
The variance of the data is the average squared distance between the mean and each data value.

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}$$

NOTE:
The variance is written as σ^2. It might seem strange that it is written in squared form, but you will see why soon when we discuss the standard deviation.

The variance has the following properties.

- It is never negative since every term in the variance sum is squared and therefore either positive or zero.
- It has squared units. For example, the variance of a set of heights measured in centimetres will be given in centimeters squared. Since the population variance is squared, it is not directly comparable with the mean or the data themselves. In the next section we will describe a different measure of dispersion, the standard deviation, which has the same units as the data.

Worked example 10: Variance

QUESTION

You flip a coin 100 times and it lands on heads 44 times. You then use the same coin and do another 100 flips. This time in lands on heads 49 times. You repeat this experiment a total of 10 times and get the following results for the number of heads.

$$\{44; 49; 52; 62; 53; 48; 54; 49; 46; 51\}$$

Compute the mean and variance of this data set.

SOLUTION

Step 1: Compute the mean

The formula for the mean is

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

In this case, we sum the data and divide by 10 to get $\overline{x} = 50{,}8$.

Step 2: Compute the variance

The formula for the variance is

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}$$

We first subtract the mean from each datum and then square the result.

x_i	44	49	52	62	53	48	54	49	46	51
$x_i - \overline{x}$	$-6{,}8$	$-1{,}8$	1,2	11,2	2,2	$-2{,}8$	3,2	$-1{,}8$	$-4{,}8$	0,2
$(x_i - \overline{x})^2$	46,24	3,24	1,44	125,44	4,84	7,84	10,24	3,24	23,04	0,04

The variance is the sum of the last row in this table divided by 10, so $\sigma^2 = 22{,}56$.

Since the variance is a squared quantity, it cannot be directly compared to the data values or the mean value of a data set. It is therefore more useful to have a quantity which is the square root of the variance. This quantity is known as the standard deviation.

DEFINITION: *Standard deviation*

Let a population consist of n elements, $\{x_1; x_2; \ldots; x_n\}$, with a mean of $\overline{x}$. The standard deviation of the data is

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}}$$

In statistics, the standard deviation is a very common measure of dispersion. Standard deviation measures how spread out the values in a data set are around the mean. More precisely, it is a measure of the average distance between the values of the data in the set and the mean. If the data values are all similar, then the standard deviation will be low (closer to zero). If the data values are highly variable, then the standard variation is high (further from zero).

The standard deviation is always a positive number and is always measured in the same units as the original data. For example, if the data are distance measurements in kilogrammes, the standard deviation will also be measured in kilogrammes.

The mean and the standard deviation of a set of data are usually reported together. In a certain sense, the standard deviation is a natural measure of dispersion if the centre of the data is taken as the mean.

Investigation: Tabulating results

It is often useful to set your data out in a table so that you can apply the formulae easily. Complete the table below to calculate the standard deviation of $\{57; 53; 58; 65; 48; 50; 66; 51\}$.

- Firstly, remember to calculate the mean, $\overline{x}$.
- Complete the following table.

index: i	datum: x_i	deviation: $x_i - \overline{x}$	deviation squared: $(x_i - \overline{x})^2$
1	57		
2	53		
3	58		
4	65		
5	48		
6	50		
7	66		
8	51		
	$\sum x_i = \ldots$	$\sum (x_i - \overline{x}) = \ldots$	$\sum (x_i - \overline{x})^2 = \ldots$

- The sum of the deviations is always zero. Why is this? Find out.
- Calculate the variance using the completed table.
- Then calculate the standard deviation.

Worked example 11: Variance and standard deviation

QUESTION

What is the variance and standard deviation of the possibilities associated with rolling a fair die?

SOLUTION

Step 1: Determine all the possible outcomes

When rolling a fair die, the sample space consists of 6 outcomes. The data set is therefore $x = \{1; 2; 3; 4; 5; 6\}$ and $n = 6$.

Step 2: Calculate the mean

The mean is:

$$\begin{aligned}\overline{x} &= \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) \\ &= 3{,}5\end{aligned}$$

Step 3: Calculate the variance

The variance is:

$$\begin{aligned}\sigma^2 &= \frac{\sum (x - \overline{x})^2}{n} \\ &= \frac{1}{6}(6{,}25 + 2{,}25 + 0{,}25 + 0{,}25 + 2{,}25 + 6{,}25) \\ &= 2{,}917\end{aligned}$$

Step 4: Calculate the standard deviation

The standard deviation is:

$$\begin{aligned}\sigma &= \sqrt{2{,}917} \\ &= 1{,}708\end{aligned}$$

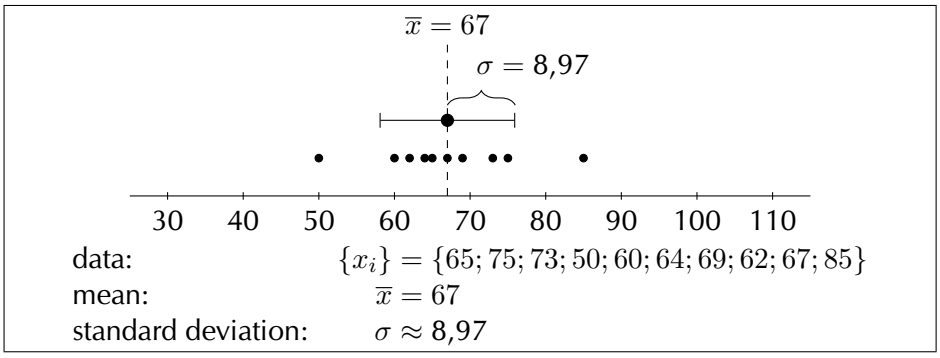
See video: 23CW at www.everythingmaths.co.za

A large standard deviation indicates that the data values are far from the mean and a small standard deviation indicates that they are clustered closely around the mean.

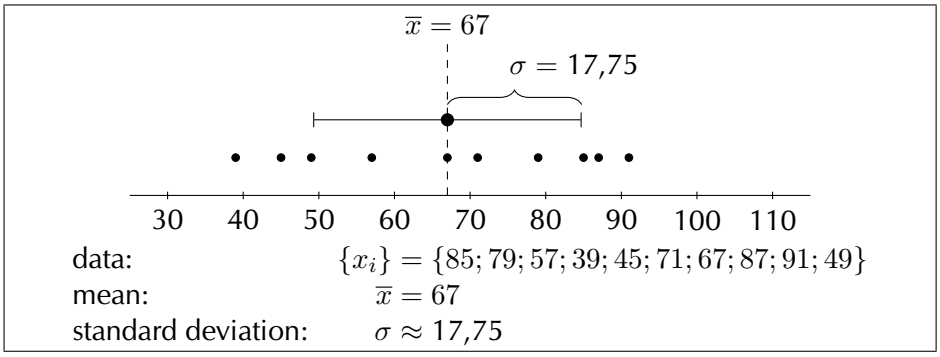
For example, consider the following three data sets:

$$\{65; 75; 73; 50; 60; 64; 69; 62; 67; 85\}$$
$$\{85; 79; 57; 39; 45; 71; 67; 87; 91; 49\}$$
$$\{43; 51; 53; 110; 50; 48; 87; 69; 68; 91\}$$

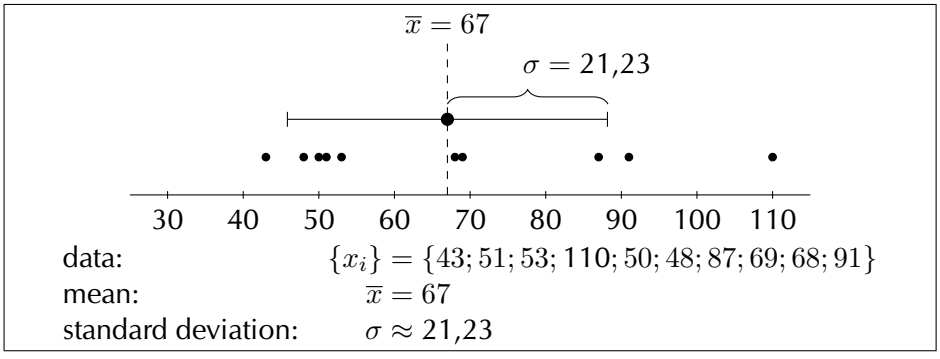
Each of these data sets has the same mean, namely 67. However, they have different standard deviations, namely 8,97, 17,75 and 21,23. The following figures show plots of the data sets with the mean and standard deviation indicated on each. You can see how the standard deviation is larger when the data are more spread out.



data: $\{x_i\} = \{65; 75; 73; 50; 60; 64; 69; 62; 67; 85\}$
mean: $\overline{x} = 67$
standard deviation: $\sigma \approx 8{,}97$



data: $\{x_i\} = \{85; 79; 57; 39; 45; 71; 67; 87; 91; 49\}$
mean: $\overline{x} = 67$
standard deviation: $\sigma \approx 17{,}75$



data: $\{x_i\} = \{43; 51; 53; 110; 50; 48; 87; 69; 68; 91\}$
mean: $\overline{x} = 67$
standard deviation: $\sigma \approx 21{,}23$

The standard deviation may also be thought of as a measure of uncertainty. In the physical sciences, for example, the reported standard deviation of a group of repeated measurements represents the precision of those measurements. When deciding whether

measurements agree with a theoretical prediction, the standard deviation of those measurements is very important: if the mean of the measurements is too far away from the prediction (with the distance measured in standard deviations), then we consider the measurements as contradicting the prediction. This makes sense since they fall outside the range of values that could reasonably be expected to occur if the prediction were correct.

Exercise 11 – 4: Variance and standard deviation

1. Bridget surveyed the price of petrol at petrol stations in Cape Town and Durban. The data, in rands per litre, are given below.

Cape Town	3,96	3,76	4,00	3,91	3,69	3,72
Durban	3,97	3,81	3,52	4,08	3,88	3,68

   a) Find the mean price in each city and then state which city has the lowest mean.

   b) Find the standard deviation of each city's prices.

   c) Which city has the more consistently priced petrol? Give reasons for your answer.

2. Compute the mean and variance of the following set of values.

   150 ; 300 ; 250 ; 270 ; 130 ; 80 ; 700 ; 500 ; 200 ; 220 ; 110 ; 320 ; 420 ; 140

3. Compute the mean and variance of the following set of values.

   $-6{,}9$; $-17{,}3$; 18,1 ; 1,5 ; 8,1 ; 9,6 ; $-13{,}1$; $-14{,}0$; 10,5 ; $-14{,}8$; $-6{,}5$; 1,4

4. The times for 8 athletes who ran a 100 m sprint on the same track are shown below. All times are in seconds.

   10,2 ; 10,8 ; 10,9 ; 10,3 ; 10,2 ; 10,4 ; 10,1 ; 10,4

   a) Calculate the mean time.

   b) Calculate the standard deviation for the data.

   c) How many of the athletes' times are more than one standard deviation away from the mean?

5. The following data set has a mean of 14,7 and a variance of 10,01.

   $$18\ ;\ 11\ ;\ 12\ ;\ a\ ;\ 16\ ;\ 11\ ;\ 19\ ;\ 14\ ;\ b\ ;\ 13$$

   Compute the values of a and b.

Think you got it? Get this answer and more practice on our Intelligent Practice Service

1. 23CX 2. 23CY 3. 23CZ 4. 23D2 5. 23D3

 www.everythingmaths.co.za  m.everythingmaths.co.za

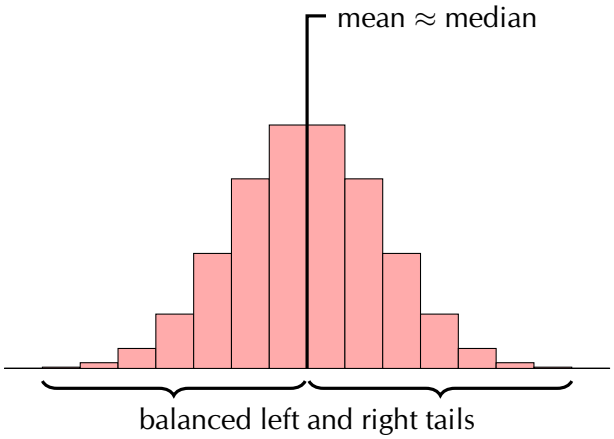
11.5 Symmetric and skewed data EMBKD

We are now going to classify data sets into 3 categories that describe the shape of the data distribution: symmetric, left skewed, right skewed. We can use this classification for any data set, but here we will look only at distributions with one peak. Most of the data distributions that you have seen so far have only one peak, so the plots in this section should look familiar.

Distributions with one peak are called **unimodal distributions**. Unimodal literally means having one mode. (Remember that a mode is a maximum in the distribution.)

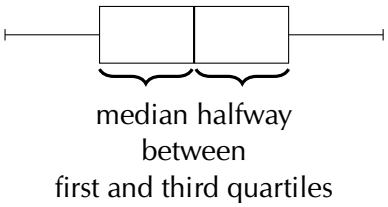
Symmetric distributions EMBKF

A symmetric distribution is one where the left and right hand sides of the distribution are roughly equally balanced around the mean. The histogram below shows a typical symmetric distribution.



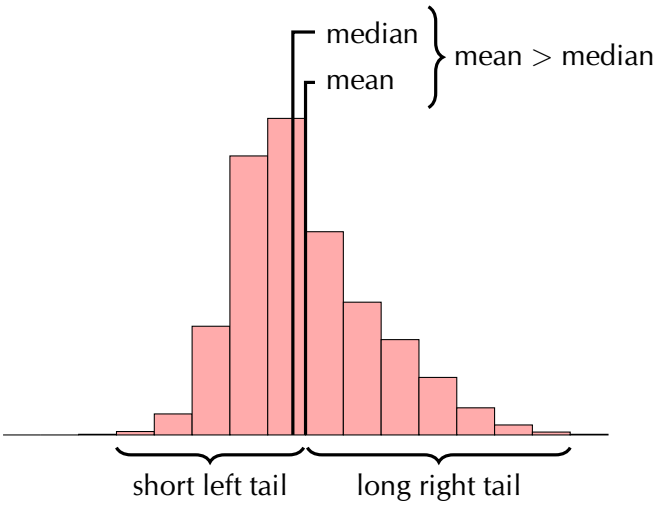
For symmetric distributions, the mean is approximately equal to the median. The **tails** of the distribution are the parts to the left and to the right, away from the mean. The tail is the part where the counts in the histogram become smaller. For a symmetric distribution, the left and right tails are equally balanced, meaning that they have about the same length.

The figure below shows the box and whisker diagram for a typical symmetric data set.

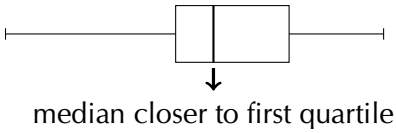


Another property of a symmetric distribution is that its median (second quartile) lies in the middle of its first and third quartiles. Note that the whiskers of the plot (the minimum and maximum) do not have to be equally far away from the median. In the next section on outliers, you will see that the minimum and maximum values do not necessarily match the rest of the data distribution well.

A distribution that is **skewed right** (also known as **positively skewed**) is shown below.



Now the picture is not symmetric around the mean anymore. For a right skewed distribution, the mean is typically greater than the median. Also notice that the tail of the distribution on the right hand (positive) side is longer than on the left hand side.

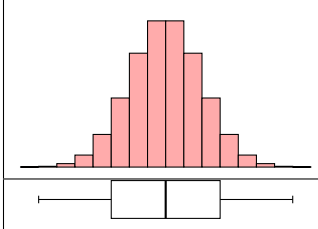
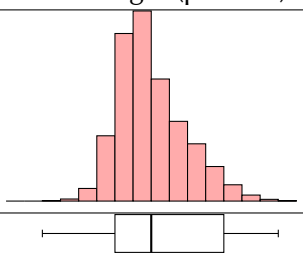
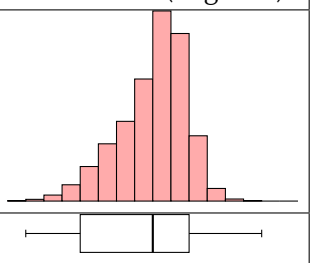


From the box and whisker diagram we can also see that the median is closer to the first quartile than the third quartile. The fact that the right hand side tail of the distribution is longer than the left can also be seen.

A distribution that is skewed left has exactly the opposite characteristics of one that is skewed right:

- the mean is typically less than the median;
- the tail of the distribution is longer on the left hand side than on the right hand side; and
- the median is closer to the third quartile than to the first quartile.

The table below summarises the different categories visually.

Symmetric	Skewed right (positive)	Skewed left (negative)
		

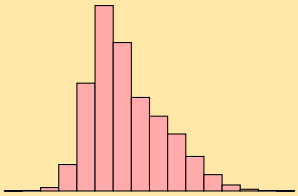
Exercise 11 – 5: Symmetric and skewed data

1. Is the following data set symmetric, skewed right or skewed left? Motivate your answer.

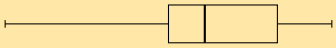
   27 ; 28 ; 30 ; 32 ; 34 ; 38 ; 41 ; 42 ; 43 ; 44 ; 46 ; 53 ; 56 ; 62

2. State whether each of the following data sets are symmetric, skewed right or skewed left.

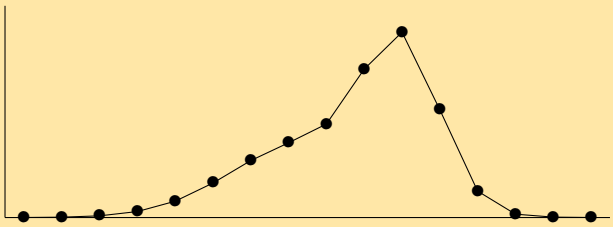
   a) A data set with this histogram:

   b) A data set with this box and whisker plot:

   c) A data set with this frequency polygon:

   d) The following data set:

   11,2 ; 5 ; 9,4 ; 14,9 ; 4,4 ; 18,8 ; −0,4 ; 10,5 ; 8,3 ; 17,8

3. Two data sets have the same range and interquartile range, but one is skewed right and the other is skewed left. Sketch the box and whisker plot for each of these data sets. Then, invent data (6 points in each data set) that matches the descriptions of the two data sets.

Think you got it? Get this answer and more practice on our Intelligent Practice Service

1. 23D4 2a. 23D5 2b. 23D6 2c. 23D7 2d. 23D8 3. 23D9

 **www.everythingmaths.co.za**  **m.everythingmaths.co.za**

An outlier in a data set is a value that is far away from the rest of the values in the data set. In a box and whisker diagram, outliers are usually close to the whiskers of the diagram. This is because the centre of the diagram represents the data between the first and third quartiles, which is where 50% of the data lie, while the whiskers represent the extremes — the minimum and maximum — of the data.

Worked example 12: Identifying outliers

QUESTION

Find the outliers in the following data set by drawing a box and whisker diagram and locating the data values on the diagram.

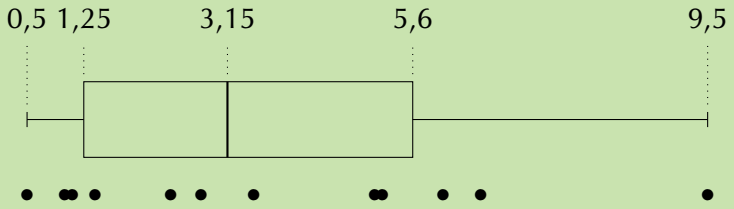
0,5 ; 1 ; 1,1 ; 1,4 ; 2,4 ; 2,8 ; 3,5 ; 5,1 ; 5,2 ; 6 ; 6,5 ; 9,5

SOLUTION

Step 1: Determine the five number summary

The minimum of the data set is 0,5. The maximum of the data set is 9,5. Since there are 12 values in the data set, the median lies between the sixth and seventh values, making it equal to $\frac{2,8+3,5}{2} = 3,15$. The first quartile lies between the third and fourth values, making it equal to $\frac{1,1+1,4}{2} = 1,25$. The third quartile lies between the ninth and tenth values, making it equal to $\frac{5,2+6}{2} = 5,6$.

Step 2: Draw the box and whisker diagram



In the figure above, each value in the data set is shown with a black dot.

Step 3: Find the outliers

From the diagram we can see that most of the values are between 1 and 6. The only value that is very far away from this range is the maximum at 9,5. Therefore 9,5 is the only outlier in the data set.

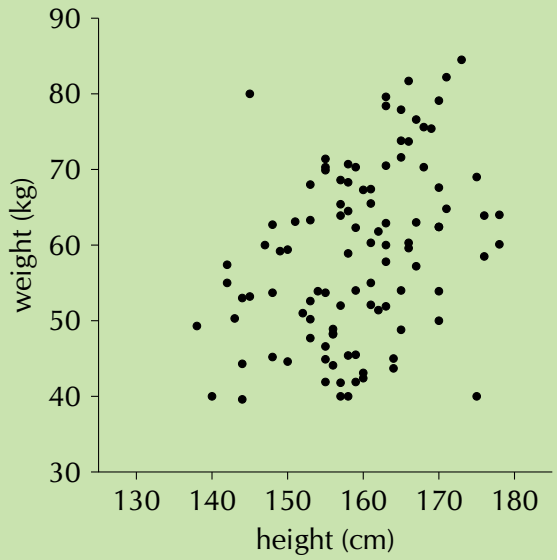
You should also be able to identify outliers in plots of two variables. A **scatter plot** is a graph that shows the relationship between two random variables. We call these data **bivariate** (literally meaning two variables) and we plot the data for two different variables on one set of axes. The following example shows what a typical scatter plot looks like. For Grade 11 you do not need to learn how to draw these 2-dimensional

scatter plots, but you should be able to identify outliers on them. As before, an outlier is a value that is far removed from the main distribution of data.

Worked example 13: Scatter plot

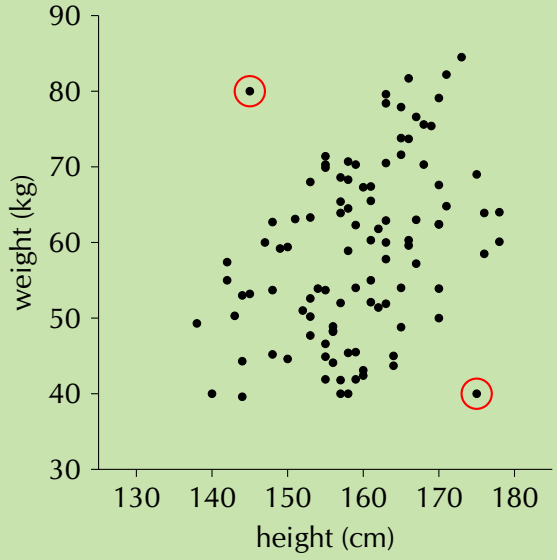
QUESTION

We have a data set that relates the heights and weights of a number of people. The height is the first variable and its value is plotted along the horizontal axis. The weight is the second variable and its value is plotted along the vertical axis. The data values are shown on the plot below. Identify any outliers on the scatter plot.



SOLUTION

We inspect the plot visually and notice that there are two points that lie far away from the main data distribution. These two points are circled in the plot below.



Exercise 11 – 6: Outliers

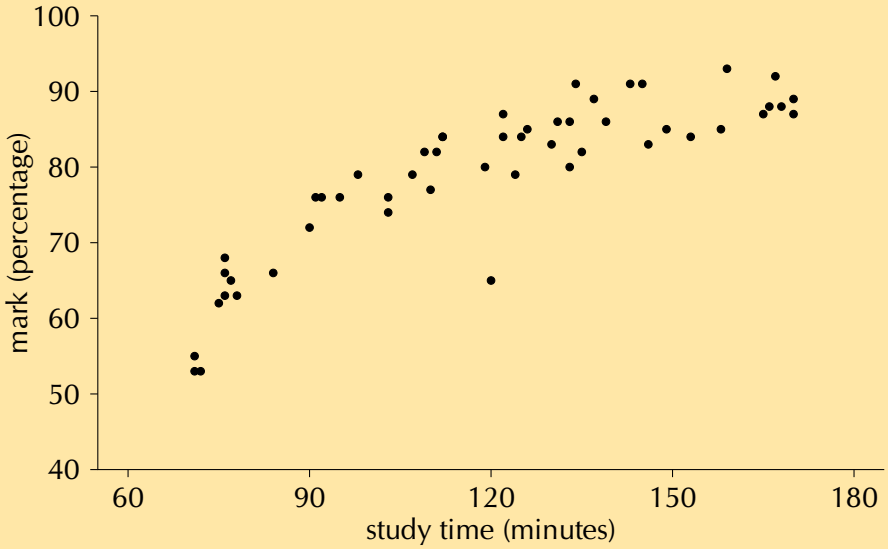
1. For each of the following data sets, draw a box and whisker diagram and determine whether there are any outliers in the data.

   a) 30 ; 21,4 ; 39,4 ; 33,4 ; 21,1 ; 29,3 ; 32,8 ; 31,6 ; 36 ; 27,9 ; 27,3 ; 29,4 ; 29,1 ; 38,6 ; 33,8 ; 29,1 ; 37,1

   b) 198 ; 166 ; 175 ; 147 ; 125 ; 194 ; 119 ; 170 ; 142 ; 148

   c) 7,1 ; 9,6 ; 6,3 ; −5,9 ; 0,7 ; −0,1 ; 4,4 ; −11,7 ; 10 ; 2,3 ; −3,7 ; 5,8 ; −1,4 ; 1,7 ; −0,7

2. A class's results for a test were recorded along with the amount of time spent studying for it. The results are given below. Identify any outliers in the data.



Think you got it? Get this answer and more practice on our Intelligent Practice Service

1a. 23DB 1b. 23DC 1c. 23DD 2. 23DF

www.everythingmaths.co.za m.everythingmaths.co.za

11.7 Summary EMBKJ

See presentation: 23DG at www.everythingmaths.co.za

- Histograms visualise how many times different events occurred. Each rectangle in a histogram represents one event and the height of the rectangle is relative to the number of times that the event occurred.
- Frequency polygons represent the same information as histograms, but using lines and points rather than rectangles. A frequency polygon connects the middle of the top edge of each rectangle in a histogram.
- Ogives (also known as cumulative histograms) show the total number of times that a value or anything less than that value appears in the data set. To draw an ogive you need to add up all the counts in a histogram from left to right.
  - The first count in an ogive is always zero.
  - The last count in an ogive is always the sum of all the counts in the data set.
- The variance and standard deviation are measures of dispersion.
  - The standard deviation is the square root of the variance.
  - Variance: $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$
  - Standard deviation: $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$
  - The standard deviation is measured in the same units as the mean and the data, but the variance is not. The variance is measured in the square of the data units.
- In a symmetric distribution
  - the mean is approximately equal to the median; and
  - the tails of the distribution are balanced.
- In a right (positively) skewed distribution
  - the mean is greater than the median;
  - the tail on the right hand side is longer than the tail on the left hand side; and
  - the median is closer to the first quartile than the third quartile.
- In a left (negatively) skewed distribution
  - the mean is less than the median;
  - the tail on the left hand side is longer than the tail on the right hand side; and
  - the median is closer to the third quartile than the first quartile.
- An outlier is a value that is far away from the rest of the data.

Exercise 11 – 7: End of chapter exercises

1. Draw a histogram, frequency polygon and ogive of the following data set. To count the data, use intervals with a width of 1, starting from 0.

   0,4 ; 3,1 ; 1,1 ; 2,8 ; 1,5 ; 1,3 ; 2,8 ; 3,1 ; 1,8 ; 1,3 ;

   2,6 ; 3,7 ; 3,3 ; 5,7 ; 3,7 ; 7,4 ; 4,6 ; 2,4 ; 3,5 ; 5,3

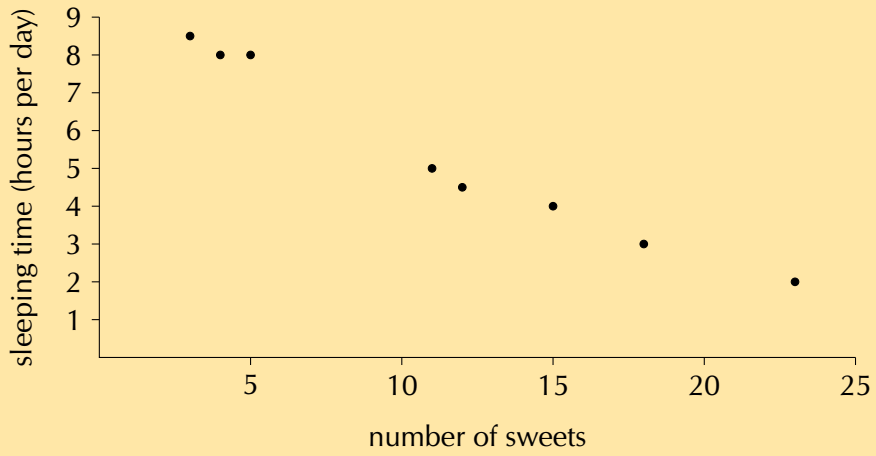
2. Draw a box and whisker diagram of the following data set and explain whether it is symmetric, skewed right or skewed left.

   $-4{,}1$; $-1{,}1$; -1 ; $-1{,}2$; $-1{,}5$; $-3{,}2$; -4 ; $-1{,}9$; -4 ;

   $-0{,}8$; $-3{,}3$; $-4{,}5$; $-2{,}5$; $-4{,}4$; $-4{,}6$; $-4{,}4$; $-3{,}3$

3. Eight children's sweet consumption and sleeping habits were recorded. The data are given in the following table and scatter plot.

Number of sweets per week	15	12	5	3	18	23	11	4
Average sleeping time (hours per day)	4	4,5	8	8,5	3	2	5	8



   a) What is the mean and standard deviation of the number of sweets eaten per day?

   b) What is the mean and standard deviation of the number of hours slept per day?

   c) Make a list of all the outliers in the data set.

4. The monthly incomes of eight teachers are as follows:

   R 10 050; R 14 300; R 9800; R 15 000; R 12 140; R 13 800; R 11 990; R 12 900.

   a) What is the mean and standard deviation of their incomes?

   b) How many of the salaries are less than one standard deviation away from the mean?

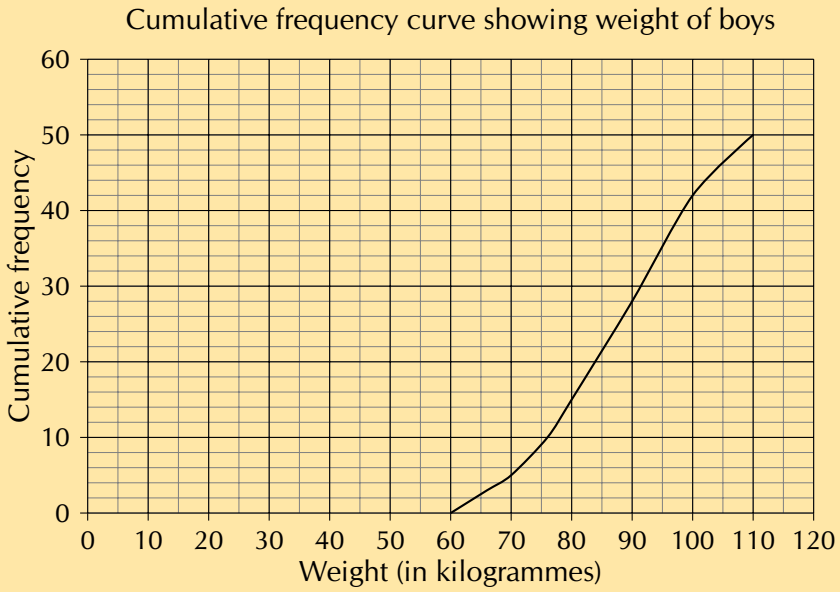
   c) If each teacher gets a bonus of R 500 added to their pay what is the new mean and standard deviation?

   d) If each teacher gets a bonus of 10% on their salary what is the new mean and standard deviation?

   e) Determine for both of the above, how many salaries are less than one standard deviation away from the mean.

f) Using the above information work out which bonus is more beneficial financially for the teachers.

5. The weights of a random sample of boys in Grade 11 were recorded. The cumulative frequency graph (ogive) below represents the recorded weights.



a) How many of the boys weighed between 90 and 100 kilogrammes?
b) Estimate the median weight of the boys.
c) If there were 250 boys in Grade 11, estimate how many of them would weigh less than 80 kilogrammes?

6. Three sets of 12 learners each had their test scores recorded. The test was out of 50. Use the given data to answer the following questions.

Set A	Set B	Set C
25	32	43
47	34	47
15	35	16
17	32	43
16	25	38
26	16	44
24	38	42
27	47	50
22	43	50
24	29	44
12	18	43
31	25	42

a) For each of the sets calculate the mean and the five number summary.
b) Make box and whisker plots of the three data sets on the same set of axes.
c) State, with reasons, whether each of the three data sets are symmetric or skewed (either right or left).

Think you got it? Get this answer and more practice on our Intelligent Practice Service
1. 23DH 2. 23DJ 3. 23DK 4. 23DM 5. 23DN 6. 23DP

 www.everythingmaths.co.za  m.everythingmaths.co.za