Census@school

Data Handling & Probability

Grades 10, 11 and 12

Statistics South Africa

your leading partner in quality statistics

2

Grade 11 Data Handling

In this chapter you:

- Draw histograms
- Draw frequency polygons
- Draw ogives (cumulative frequency curves)
- Calculate variance and standard deviation of ungrouped data;
- Determine whether data is symmetric or skewed
- Identify the values of the outliers.

WHAT YOU LEARNED ABOUT DATA HANDLING IN GRADE 10

In Grade 10 you covered the following data handling concepts:

- Measures of central tendency of lists of data, of data in frequency tables and of data in grouped frequency tables.
- The range, percentiles, quartiles, interquartile and semi-interquartile range
- The five number summary and box-and-whisker diagram
- Using statistical summaries (measures of central tendency and dispersion) to analyse and make meaningful comments on the context associated with the given data.

STATISTICAL GRAPHS

✓ Organised data can often be presented in graphical form.

- Statistical graphs are used to describe data or to analyse it.
- The purpose of graphs in statistics is to communicate the data to the viewers in pictorial form. It is easier for most people to understand data when it is presented as a graph than when it is presented numerically in tables.

✓ In earlier grades you dealt with the following graphs

- Bar graphs and double bar graphs
- Histograms
- Pie charts
- Broken-line graphs.

✓ In Grade 11 you study three statistical graphs often used in research: the ***histogram***, the ***frequency polygon*** and the ***cumulative frequency graph or ogive***.

HISTOGRAMS

✓ A ***histogram*** gives us a visual interpretation of data. It looks very similar to a ***bar graph***, but there are definite differences between them.

HISTOGRAM	BAR GRAPH
• It is a representation of grouped data • There is no gap between the bars ***For example**, you draw a HISTOGRAM to show the number of people whose heights (h) lie in the following intervals (measured in cm): $150 \leq h < 160$; $160 \leq h < 170$; etc*	• It is a representation of ungrouped data that does not have to be numerical • There is generally a gap between the bars ***For example** you draw a BAR GRAPH to show the number of learners in a class who wear glasses and the number who do not wear glasses.*

EXAMPLE 1

The following table lists the marks (given as percentage) obtained by the Grade 11 learners of Musi High School in their mathematics test:

24	70	50	22	63	45	48	52	56	38
65	68	65	17	32	60	62	53	63	45
49	44	56	12	55	83	54	22	67	54
34	77	46	50	58	80	81	39	84	75
55	76	73	80	66	71	62	40	23	76

a) Organise the data using a grouped frequency table.
b) Draw a histogram to illustrate the data.
c) Calculate the modal interval. What does this measure of central tendency tell you about the learners' marks?
d) Estimate the median. What does this measure of central tendency tell you about the learners' marks?

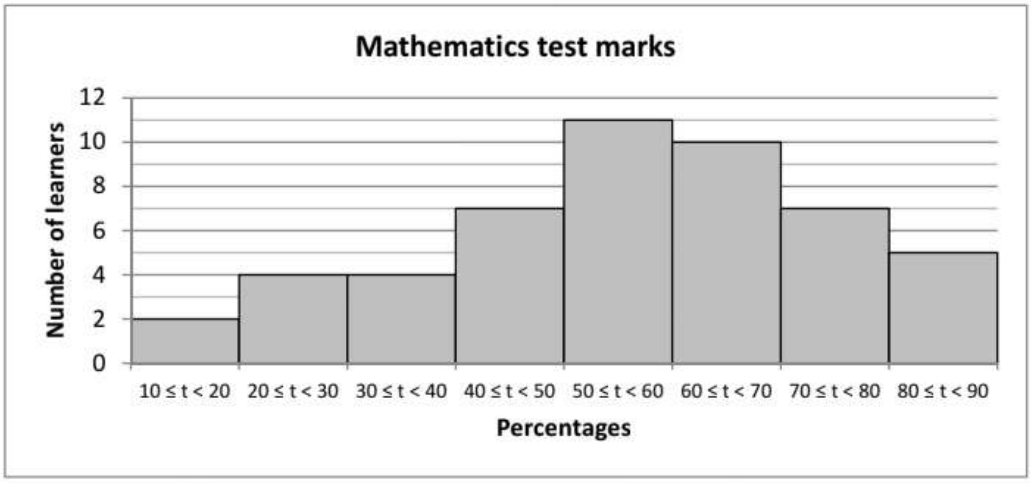
SOLUTION:

a) The lowest mark was 12% and the highest mark was 84%
It is often easiest to use multiples of 10 as the class intervals, so start the first interval at 10% and end the last interval at 90%

EXAMPLE 1 (continued)

Percentages (t)	Frequency (Number of learners)
$10 \leq t < 20$	2
$20 \leq t < 30$	4
$30 \leq t < 40$	4
$40 \leq t < 50$	7
$50 \leq t < 60$	11
$60 \leq t < 70$	10
$70 \leq t < 80$	7
$80 \leq t < 90$	5
TOTAL	50

b) Draw the histogram as follows:
STEP 1: Draw and label the horizontal and vertical axes.
STEP 2: Represent the frequency on the vertical axis and the classes on the horizontal axis.
STEP 3: Using the frequencies (or number of learners) as the heights, draw vertical bars for each class.



c) The modal interval is the interval with the largest frequency or largest number of learners. So the modal interval is $50 \leq t < 60$.
This tells us that more learners got marks in the interval $50 \leq t < 60$ than in any of the other intervals.

d) There are 50 data items (marks/percentages).
The median lies between the 25^{th} and the 26^{th} marks.
Add up the frequencies until you reach 25 (or more than 25):

$$2 + 4 + 4 + 7 + 11 = 28$$

The 28^{th} mark lies in the interval $50 \leq t < 60$
So the median lies in the interval $50 \leq t < 60$
The median $\approx 55\%$ (the midpoint of the interval)
This tells us that 50% of the learners got marks that were less ***than 55%*** and 50% of the learners got marks that were ***more than 55%***

NOTE:

A histogram should have the following:

- A title which describes the information that is contained in the histogram.
- A horizontal axis with a label which shows the scale of values into which the data fit (grouped data intervals)
- A vertical axis with a label which shows the number of times the data within the interval occurred (frequency)
- Adjacent bars (i.e. there are no gaps between the bars).

EXERCISE 2.1

1) The frequency table below represent the distribution of the amount of time (in hours) that 80 high school learners spent in one week watching their favourite sport.

Time in hours	Frequency
$10 < t \leq 15$	8
$15 < t \leq 20$	28
$20 < t \leq 25$	27
$25 < t \leq 30$	12
$30 < t \leq 35$	4
$35 < t \leq 40$	1

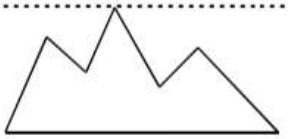
   a) Draw a histogram to represent the data
   b) Calculate
      i) the modal interval
      ii) an estimate of the median
   c) What do these two measures of central tendency tell you about the amount of time the learners devote to watching their favourite sport?

2) In the 2009 Census@School, learners were asked which their favourite subjects at school were. Fifty Grade 11 learners from a certain school in Limpopo chose Science as their favourite subject. The following are their Science marks (as percentages):

31	62	51	44	61	63	59	47	59	67
50	54	61	41	48	74	53	53	53	36
60	42	50	48	42	27	43	42	43	54
49	47	51	28	54	48	83	65	54	35
61	56	57	32	38	32	40	63	56	59

   a) Organise the data in a grouped frequency table.
   b) Draw a histogram to represents the data.
   c) Calculate the modal interval and an estimate of the median and say what these two measures of central tendency tell you about the learners' mark.

FREQUENCY POLYGONS

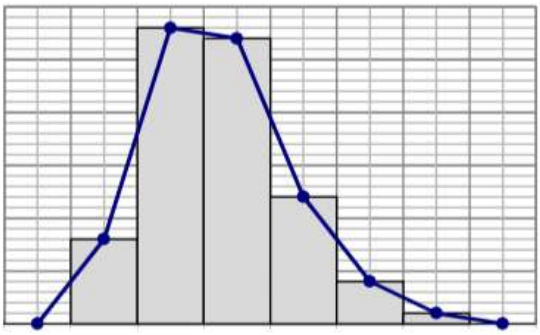


A ***polygon*** is a closed geometric shape made up of line segments.

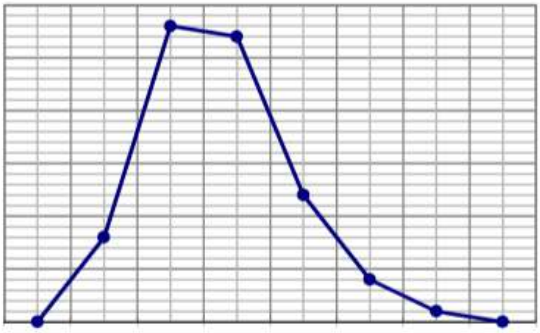
✓ A ***frequency polygon*** can be used instead of a histogram for illustrating *grouped data.*

NOTE: It is called a frequency ***polygon*** because of its shape.

✓ One way of drawing a frequency polygon is to
a) Draw a histogram
b) Join the *midpoints* of the top of the columns of the histogram
c) Extend the line to the midpoint of the *class interval below the lowest value* and to the midpoint of the *class interval above the highest value* so that the line touches the horizontal axis on both sides.



✓ Another way of drawing a frequency polygon is to
a) Calculate the midpoint of each interval and then to plot the ordered pair (midpoint of the interval; frequency)
b) Plot the midpoint of the interval below the lowest interval and the interval above the highest interval and plot the points (midpoint of the interval; 0)
c) Join these points with straight lines.



EXAMPLE 1

Eighty of the learners at Alexandra High School were surveyed to find out how many minutes each week they spent collecting waste material for recycling. The grouped frequency table shows the results of the survey.

a) Use the frequency table to draw a histogram and to then draw a frequency polygon on the histogram.

b)
   i) Find the midpoint of the intervals
   ii) Use the table to draw a frequency polygon on a separate set of axes.

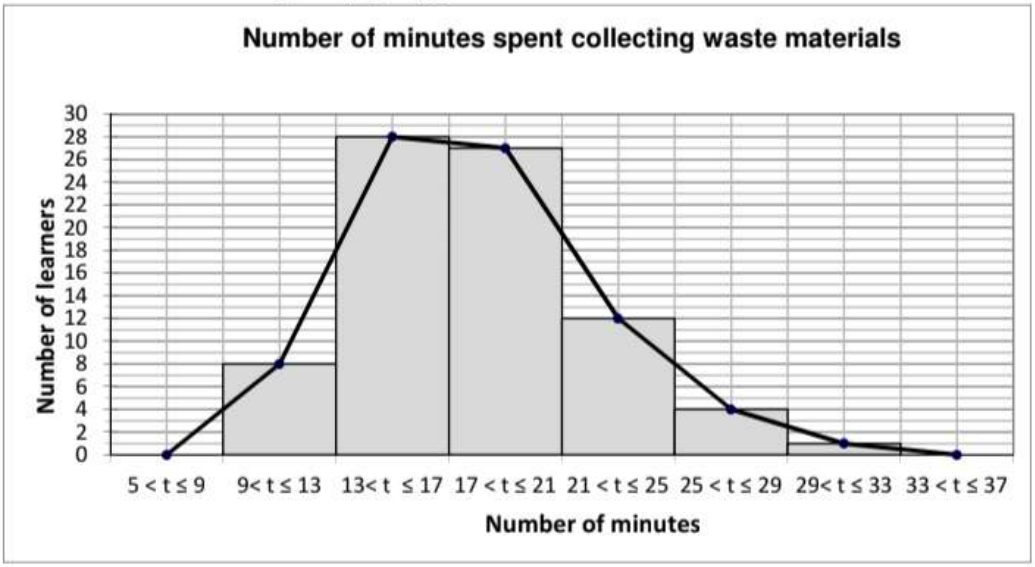
Number of minutes (t)	Number of learners (f)
$9 < t \leq 13$	8
$13 < t \leq 17$	28
$17 < t \leq 21$	27
$21 < t \leq 25$	12
$25 < t \leq 29$	4
$29 < t \leq 33$	1

SOLUTION

a) **Step 1:** Add in two classes with a frequency of zero:

Number of minutes (t)	Number of learners (f)
$5 < t \leq 9$	0
$9 < t \leq 13$	8
$13 < t \leq 17$	28
$17 < t \leq 21$	27
$21 < t \leq 25$	12
$25 < t \leq 29$	4
$29 < t \leq 33$	1
$33 < t \leq 37$	0

Step 2: Draw the histogram and then join the midpoints of the top of the columns to form the frequency polygon.



EXAMPLE 2 (continued)

b)

i) Calculate the midpoint of each interval using the formula:

$$\text{Midpoint} = \frac{lower\ limit\ of\ interval + upper\ limit\ of\ interval}{2}$$

Number of minutes (t)	Mid points	Frequency (f)	Ordered pairs
$5 < t \leq 9$	$\frac{5+9}{2} = \frac{14}{2} = 7$	0	(7; 0)
$9 < t \leq 13$	$\frac{9+13}{2} = \frac{22}{2} = 11$	8	(11; 8)
$13 < t \leq 17$	$\frac{13+17}{2} = \frac{30}{2} = 15$	28	(15; 28)
$17 < t \leq 21$	$\frac{17+21}{2} = \frac{38}{2} = 19$	27	(19; 27)
$21 < t \leq 25$	$\frac{21+25}{2} = \frac{46}{2} = 23$	12	(23; 12)
$25 < t \leq 29$	$\frac{25+29}{2} = \frac{54}{2} = 27$	4	(27; 4)
$29 < t \leq 33$	$\frac{29+33}{2} = \frac{62}{2} = 31$	1	(31; 1)
$33 < t \leq 37$	$\frac{33+37}{2} = \frac{70}{2} = 35$	0	(35; 0)

ii) Plot the ordered pairs (midpoint; frequency) and join them with straight lines. Make sure that the graph touches the horizontal axis on both sides.

NOTE:

The ***main advantage*** of using a frequency polygon instead of a histogram is that you can easily draw *two or more* frequency polygons on the same set of axes and make comparisons between the sets of data.

EXAMPLE 3

The Grade 10 and Grade 11 learners were surveyed to find out the approximate number of hours every week they spend doing their Mathematics and Science homework. The results are summarised in the following grouped frequency table:

Number of hours spent on Mathematics and Science homework each week (t)	**Number of Grade 10 learners**	**Number of Grade 11 learners**
$5 \leq t < 10$	3	12
$10 \leq t < 15$	4	22
$15 \leq t < 20$	7	10
$20 \leq t < 25$	19	6
$25 \leq t < 30$	16	8
$30 \leq t < 35$	1	2
TOTAL	50	60

a) Draw two frequency polygons on the same set of axes to illustrate this data.

b) Use the table and the graphs to answer the following:
   i) What is the modal interval for Grade 10 and also for Grade 11?
   ii) Approximately how many more Grade 11 learners than Grade 10 learners spent between 15 and 20 hours doing their homework each week?
   iii) Which grade spent more time doing their homework?

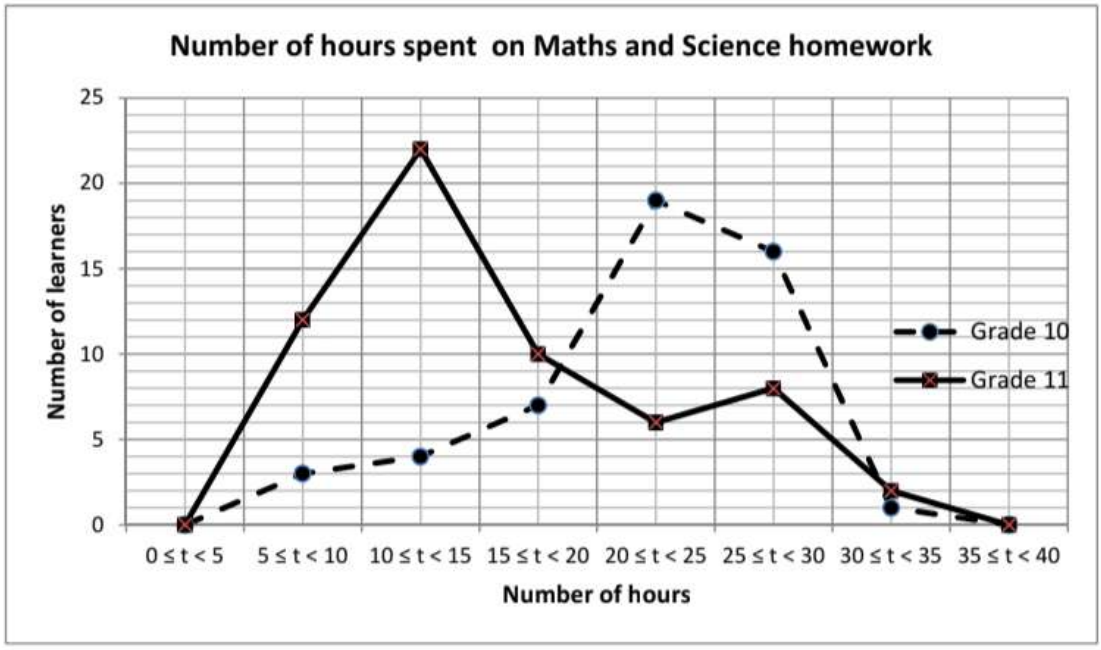
SOLUTION:

a)

Number of hours spent on Mathematics and Science homework each week (t)	**Mid-point of the interval**	**Number of Grade 10 learners**	**Number of Grade 11 learners**
$0 \leq t < 5$	2,5	0	0
$5 \leq t < 10$	7,5	3	12
$10 \leq t < 15$	12,5	4	22
$15 \leq t < 20$	17,5	7	10
$20 \leq t < 25$	22.5	19	6
$25 \leq t < 30$	27,5	16	8
$30 \leq t < 35$	32,5	1	2
$35 \leq t < 40$	37,5	0	0

- *Note that it is not essential to have the same number of data items in the two sets of data.*

EXAMPLE 3 (continued)



b)

i) The modal interval for Grade 10 is $20 < t \leq 25$
The modal interval for Grade 11 is $10 < t \leq 15$

ii) Difference in the number of learners who spent between 15 and 20 hours doing homework each week = Number in Grade 11 – Number in Grade 10
= 10 – 7
= 3
So 3 more Grade 11 learners than Grade 10 learners spent between 15 and 20 hours doing homework each week

iii) According to the table:

36 out of 50 Grade 10 learners (72% of them) spent 20 hours or more doing homework each week.

16 out of 60 Grade 11 learners (27% of them) spent 20 hours or more doing homework each week.

So the Grade 10s spent more time on homework than the Grade 11s.

EXERCISE 2.2

1) The learners at Mjolo High School enjoy taking part in athletics.
Some of the learners took part in the long jump.
The distances they jumped (in metres) are:

> Athletics is a collection of sporting events that involve competitive running, jumping and throwing.

5,46	5,97	6,72	6,26	5,13	6,36
6,11	6,38	6,55	5,84	6,20	6,34
5,80	5,43	5,93	6,64	5,67	6,00
6,05	6,88	5,50	5,51	6,10	5,49

a) Copy and complete the following grouped frequency table:

Distance in metres (m)	Frequency	Midpoints
$5{,}00 < m \leq 5{,}50$		
$5{,}50 < m \leq 6{,}00$		
$6{,}00 < m \leq 6{,}50$		
$6{,}50 < m \leq 7{,}00$		

b) Draw a frequency polygon to illustrate the data.
c) Write down the modal interval.

2) Some of the learners took part in the javelin competition. The best distances (in metres) thrown by each competitor in 2011 and 2012 are shown.

Distance thrown in metres (m)	Number of competitors 2011	Number of competitors 2012
$10 < m \leq 20$	0	1
$20 < m \leq 30$	3	4
$30 < m \leq 40$	14	19
$40 < m \leq 50$	21	13
$50 < m \leq 60$	7	11
$60 < m \leq 70$	0	2
TOTAL	**45**	**50**

a) On the same set of axes, draw frequency polygons to illustrate the 2011 and 2012 results.
b) By referring to the table and the frequency polygons, comment on the performance of the competitors in 2011 and 2012.

OGIVES / CUMULATIVE FREQUENCY CURVES

FREQUENCY

Frequency tells us *how many of each item there are in a data set.*

For example
As part of the Census@School, 170 learners were surveyed to find out the type of dwelling that they lived in.

The following table shows the result of the survey:

Type of house that you live in	Frequency (number of learners)
Traditional dwelling	7
House on separate yard	76
Tent	1
Informal dwelling in an informal settlement	86
	TOTAL = 170

CUMULATIVE FREQUENCY

Cumulative frequency shows the number of results that are *less than* (<) or *less than or equal to* (≤) a stated value in a set of data.

To find the ***cumulative frequency***,

- Add up the frequencies as you go down the frequency table.
- Write each *running total* or *cumulative frequency* in your table.

For example
Using the above information, we can find the cumulative frequency.

Type of house that you live in	Frequency (number of learners)	Cumulative Frequency
Traditional dwelling	7	**7**
House on separate yard	76	7 + 76 = **83**
Tent	1	83 + 1 = **84**
Informal dwelling in an informal settlement	86	84 + 86 = **170**
	TOTAL = 170	

Can you see that the last cumulative frequency is equal to the total frequency? (This is a useful check of your addition.)

You can find ***cumulative frequencies*** of discrete data and continuous data.

- ✓ An ***ogive*** or ***cumulative frequency curve*** is a graph that shows the information in a cumulative frequency table. The graph is useful for estimating the median and inter-quartile range of the grouped data.

- ✓ You can draw an ***ogive*** of ungrouped discrete data, grouped discrete data or grouped continuous data. It can be drawn from a grouped frequency table or an ungrouped frequency table.

EXAMPLE 4

The following frequency table shows the time (in minutes) taken by learners to travel to school.

Time taken to travel to school	Frequency	Cumulative Frequency	Ordered Pairs
$0 < t \leq 10$	4		
$10 < t \leq 20$	12		
$20 < t \leq 30$	28		
$30 < t \leq 40$	32		
$40 < t \leq 50$	29		
$50 < t \leq 60$	15		

a) Complete the table.
b) Draw an ogive to illustrate the information.

SOLUTION:

a) Steps to follow when completing the table:
- Add in an interval with a frequency of 0 before the first interval.
- Find the cumulative frequency by adding the frequencies.
- List the ordered pairs where *the first coordinate = upper limit of the interval* and *the second coordinate = cumulative frequency*.

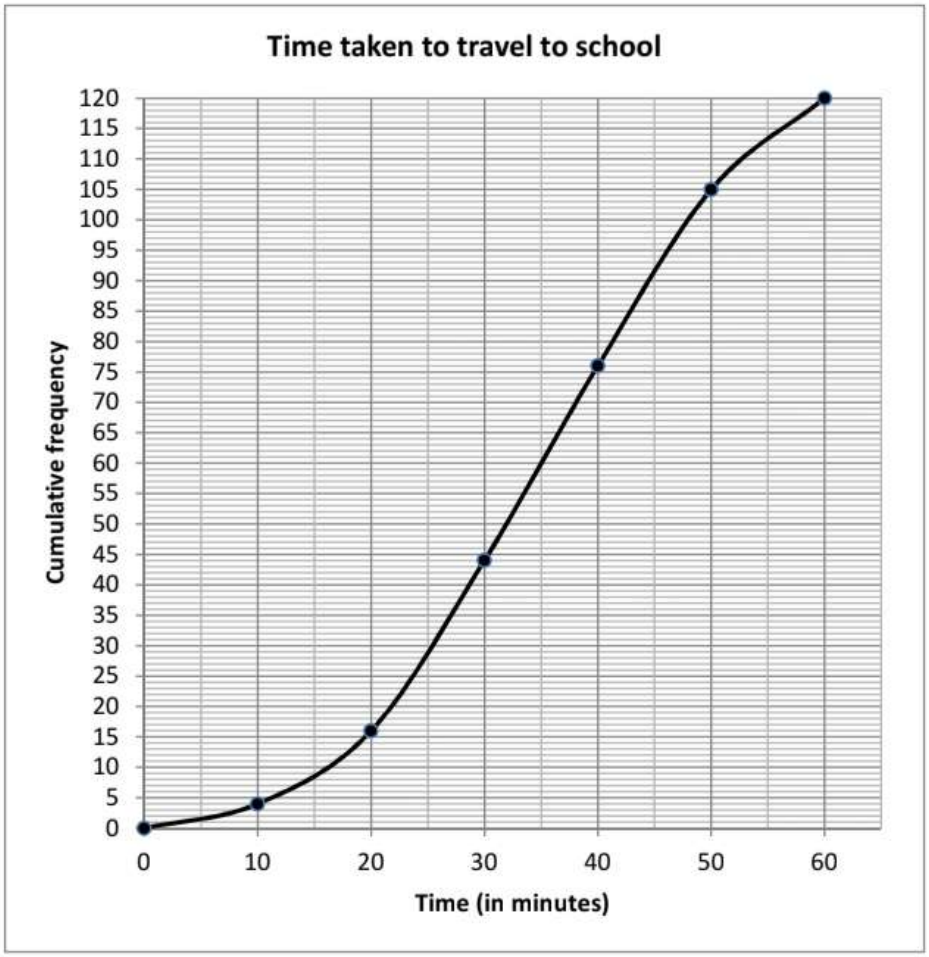
Note: A cumulative frequency of 105 means that 105 learners or less spent 50 minutes or less to walk to school.

Time taken to travel to school	Frequency	Cumulative Frequency	Ordered Pairs
$-10 < t \leq \mathbf{0}$	0	**0**	**(0;0)**
$0 < t \leq \mathbf{10}$	4	**4**	**(10;4)**
$10 < t \leq \mathbf{20}$	12	**4 + 12 = 16**	**(20;16)**
$20 < t \leq \mathbf{30}$	28	**16 + 28 = 44**	**(30;44)**
$30 < t \leq \mathbf{40}$	32	**44 + 32 = 76**	**(40;76)**
$40 < t \leq \mathbf{50}$	29	**76 + 29 = 105**	**(50;105)**
$50 < t \leq \mathbf{60}$	15	**105 + 15 = 120**	**(60;120)**

b) Draw the ogive as follows:
i) Draw the axes and label the variable on the x-axis and the cumulative frequency on the y-axis.
ii) Plot the ordered pairs.
iii) Join the points to form a smooth curve.

EXAMPLE 4 (continued)

The ogive:



✓ Always remember when drawing cumulative frequency curve from a table of grouped data, the ***cumulative frequencies*** are plotted at the ***upper limit*** of the interval**.**

EXAMPLE 5

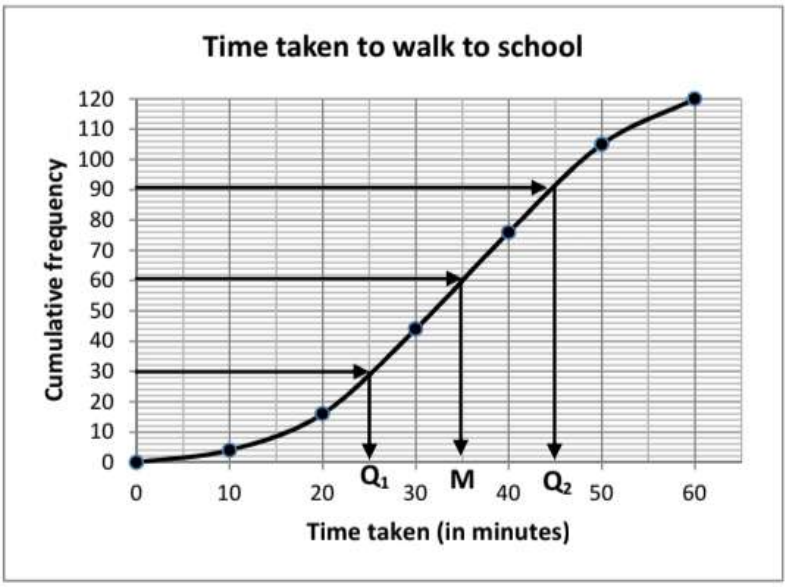
Use the ogive drawn in Example 4 to

a) Determine the approximate values of
   i) the median
   ii) the lower quartile
   iii) the upper quartile of the set of data.

b) What does each of these values tell you about the time taken by the learners?

SOLUTION:

a) This is the ogive drawn in Example 4:



i) To find the approximate value of the ***median (M)***, find the midpoint of the values plotted on the ***cumulative frequency axis.***

- The maximum value is 120, so the median lies between the 60^{th} and 61^{st} term.
- Draw a horizontal line from just above 60 until it touches the ogive.
- From that point draw a vertical line down to the *horizontal* axis.

So the ***median ≈ 35 minutes***.

ii) To find the approximate value of the ***lower quartile (Q_1)***, find the midpoint of the lower half of the values plotted on the ***cumulative frequency axis.***

- There are 60 terms in the lower half of the data, so the lower quartile lies between the 30^{th} and the 31^{st} term.
- Draw a horizontal line from just above 30 until it touches the ogive.
- From that point draw a vertical line down to the *horizontal* axis.

So the ***lower quartile ≈ 25 minutes***.

EXAMPLE 5 (continued)

iii) To find the approximate value of the ***upper quartile*** (Q_3), find the midpoint of the upper half of the values plotted on the ***cumulative frequency axis.***

- There are 60 terms in the upper half of the data, so the upper quartile lies between $60 + 30 = 90^{th}$ and the 91^{st} term.
- Draw a horizontal line from just above 90 until it touches the ogive.
- From that point draw a vertical line down to the *horizontal* axis.

So the ***upper quartile ≈ 45 minutes***.

b)

i) The ***median*** tells us that 50% of the learners took 35 minutes or less or to walk to school.

ii) The ***lower quartile*** tells us that 25% of the learners took 25 minutes or less to walk to school.

iii) The ***upper quartile*** tells us that 75% of the learners took 45 minutes or less to walk to school.

EXERCISE 2.3

1) In the 2009 Census@School learners were asked what their arm span was, correct to the nearest centimetre. The results of two hundred of the Grade 10, 11 and 12 learners who took part were recorded as follows:

Arm span in cm	Frequency	Cumulative Frequency
$130 < h \leq 135$	16	
$135 < h \leq 140$	26	
$140 < h \leq 145$	42	
$145 < h \leq 150$	54	
$150 < h \leq 155$	26	
$155 < h \leq 160$	22	
$160 < h \leq 165$	14	

To find your arm span: Open arms wide, measure the distance across your back from the tip of your right hand middle finger to the tip of your left hand middle finger.

a) Copy and complete the table.

b) Draw an ogive to illustrate the data.

c) Use your ogive to determine approximately how many learners have arm spans that are less than or equal to 152 cm.

d) Use your graph to determine approximately how many learners have arm spans of between 138 cm and 158 cm.

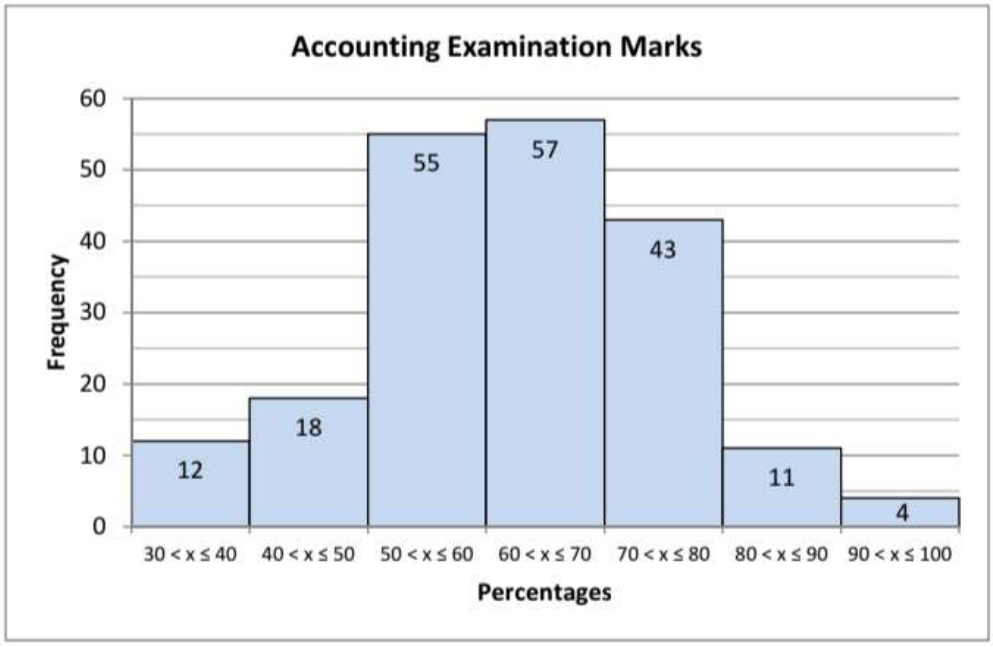
EXERCISE 2.3 (continued)

2) Fifty learners who travel by car to school were asked to record the number of kilometres travelled to and from school in one week. The following table shows the results:

Number of kilometres	Number of learners	Cumulative frequency
$10 < x \leq 20$	2	2
$20 < x \leq 30$		9
$30 < x \leq 40$		13
$40 < x \leq 50$		26
$50 < x \leq 60$		42
$60 < x \leq 70$		50
	TOTAL = 50	

a) Copy the table and then fill in the second column of the table.
b) Draw an ogive to illustrate the data.
c) Use your graph to estimate the median number of kilometres travelled per week.

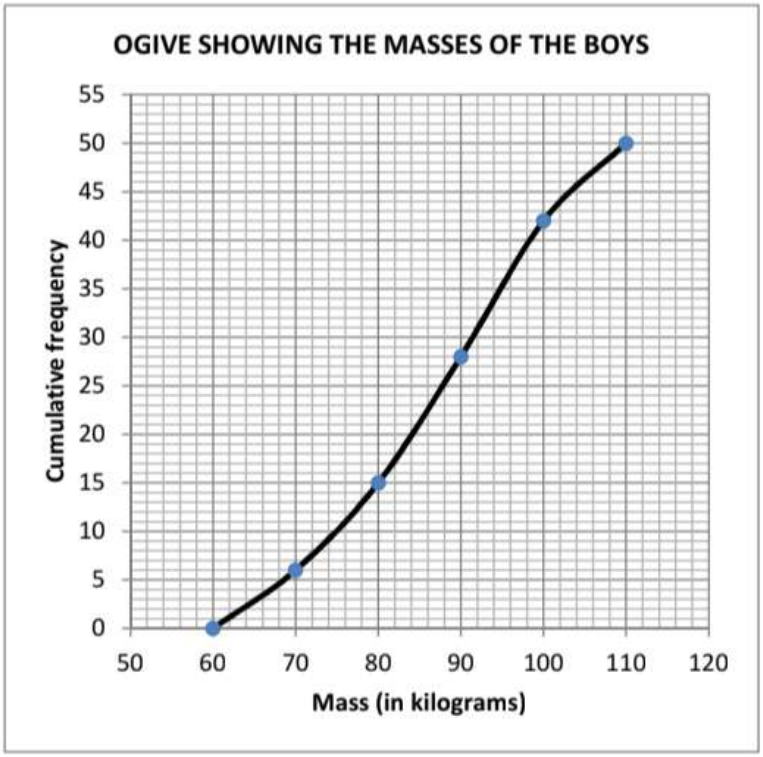
3) The histogram below shows the distribution of the Accounting examination marks for 200 learners.



a) Draw a grouped frequency table to record the data shown on the histogram.
b) Draw an ogive to illustrate the data in the frequency table.
c) Use the ogive to estimate how many learners scored 72% or more for the examination.

EXERCISE 2.3 (continued)

4) The masses of a random sample of 50 boys in Grade 11 were recorded. This cumulative frequency graph (ogive) represents the recorded masses.



a) How many of the boys had a mass between 90 and 100 kilograms?
b) Estimate the median mass of the boys.
c) Estimate how many of boys had mass less than 80 kilograms.

CHOOSING WHICH DISPLAY TO USE

The following table will help you when you have to select the appropriate diagram or graph for your data by identifying the diagrams most commonly associated with different types of data.

DATA TYPE	DESCRIPTION	EXAMPLE	TYPE OF DISPLAY
Qualitative data	Data that can be arranged into categories that are not numerical such as physical traits, gender, and colours.	*To show frequencies* e.g. 10 girls in this class have blonde hair, 18 black hair, 12 brown hair, etc.	Bar graph
		To show proportions e.g. area by province in South Africa: Western Cape 10,6%; Gauteng 1,5%; Free State 10,6%; KwaZulu Natal 7,7%; Limpopo 10,3%; Mpumalanga 6,3%; North West 8,6%; Northern Cape 30,5%.	Pie chart
Discrete Data	Data that has a finite number of different responses such as the number of people in a household.	Few different values 2, 4, 67, 34, 69	Tally table for counting, bar graph for display
		Many different values 4, 24, 25, 26, 45, 37, 38, 48, 53, 120, 75, 67, 100, 89, 47, 58, 87, 55, 45,	Stem-and-leaf diagram or a bar graph or a histogram or a box and whisker diagram or a frequency polygon
Grouped data	Data which have been arranged in groups or classes rather than showing all the original figures.	*Equal intervals* $3 \leq x < 9$ $9 \leq x < 15$ $15 \leq x < 21$	Histogram or frequency polygon
Cumulative data	Data that is increasing by successive additions of the same numbers	*Continuous variable* Time, height, etc.	Ogive
Two samples	Qualitative data	For example comparing the number of learners in the class with brown eyes, blue eyes, and green eyes.	Compound bar graphs or side-by-side pie charts.
	Discrete data	For example comparing the English exam marks for boys and girls in a Grade 11 class.	For ungrouped data use back-to-back stem-and-leaf diagrams or compound bar graphs. For grouped data use frequency polygons
	Continuous data	For example comparing the heights of the girls and the boys in a Grade 11 class.	Ogives or frequency polygons
Two variables	Used to decide whether there is a relationship between the two variables	*Data in pairs* Shoe size and the age of a person.	Scatter plot

VARIANCE AND STANDARD DEVIATION

✓ The ***interquartile range (IQR)*** measures the spread of the *middle half* of the data and is closely linked to the ***median***.

Interquartile range = upper quartile – lower quartile

Or

$$IQR = Q_3 - Q_1$$

✓ We can define ***two more measures of dispersion***, taking into account all of the data, which are linked to the ***mean***. They are the ***variance*** and the ***standard deviation***.

✓ The ***variance*** is the *mean of the sums of the squares of the deviations from the mean*.
We find the variance by:

i) Finding the mean: $\bar{x} = \frac{\sum x}{n}$

ii) Finding the deviation from the mean of each item of the data set:
Deviation = data item – mean $= x - \bar{x}$

iii) Squaring each deviation : $(deviation)^2 = (x - \bar{x})^2$

iv) Finding the sum of the squares of the deviations:
$\sum(deviation)^2 = \sum(x - \bar{x})^2$

v) Finding the mean of the squares of the deviations by dividing by the number of terms in the data set:

$$Variance = \frac{\sum(deviation)^2}{number\ of\ data\ items} = \frac{\sum(x-\bar{x})^2}{n}$$

✓ The ***standard deviation*** is the square root of the variance:

$$Standard\ Deviation = \sqrt{variance} = \sqrt{\frac{\sum(deviation)^2}{number\ of\ data\ items}} = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$$

✓ When data elements are ***tightly clustered together***, the standard deviation and variance are ***small***; when they are ***spread apart***, the standard deviation and the variance are ***relatively large***.

- A data set with ***more data near the mean*** will have ***less spread*** and a ***smaller standard deviation***
- A data set with ***lots of data far from the mean*** which will have a ***greater spread*** and a ***larger standard deviation***.

EXAMPLE 6

a) Calculate the variance and the standard deviation of the following two data sets:

Set A	182	182	184	184	185	185	186
Set B	152	166	176	184	194	200	216

b) Use the two standard deviations to compare the distribution of data in the two sets.

SOLUTION:

a) **Step 1:** Find the mean of each set

$$\text{Mean of Set A} = \frac{182+182+184+184+185+185+186}{7} = \frac{1\,288}{7} = 184$$

$$\text{Mean of Set B} = \frac{152+166+176+184+194+200+216}{7} = \frac{1\,288}{7} = 184$$

Step 2: Find the deviation from the mean of each item in the data set

Step 3: Square each deviation

Step 4: Find the variance

Step 5: Find the standard deviation

GROUP A			GROUP B		
Data item	**Deviation from the mean**	**(Deviation)^2**	**Data item**	**Deviation from the mean**	**(Deviation)^2**
182	$182 - 184 = -2$	$(-2)^2 = 4$	152	$152 - 184 = -32$	$(-32)^2 = 1\,024$
182	$182 - 184 = -2$	$(-2)^2 = 4$	166	$166 - 184 = -18$	$(-18)^2 = 324$
184	$184 - 184 = 0$	$0^2 = 0$	176	$176 - 184 = -8$	$(-8)^2 = 64$
184	$184 - 184 = 0$	$0^2 = 0$	184	$184 - 184 = 0$	$0^2 = 0$
185	$185 - 184 = 1$	$1^2 = 1$	194	$194 - 184 = 10$	$10^2 = 100$
185	$185 - 184 = 1$	$1^2 = 1$	200	$200 - 184 = 16$	$16^2 = 256$
186	$186 - 184 = 2$	$2^2 = 4$	216	$216 - 184 = 32$	$32^2 = 1\,024$
		$\sum(\text{deviations})^2 = 14$			$\sum(\text{deviations})^2 = 2\,792$
$\textit{Variance} = \frac{\sum(\textit{deviation})^2}{\textit{number of data items}} = \frac{14}{7} = 2$			$\textit{Variance} = \frac{\sum(\textit{deviation})^2}{\textit{number of data items}} = \frac{2\,792}{7} = 398,857\ldots$		
$\textit{Standard Deviation} = \sqrt{\textit{variance}} = \sqrt{2} \approx 1{,}414$			$\textit{Standard Deviation} = \sqrt{\textit{variance}} = \sqrt{\frac{2\,792}{7}} \approx 19{,}971$		

b) The larger standard deviation in Group B indicates that the data items are generally much further from the mean than the data items in Group 1.
This means that the data items in Group B are more spread out than the data items in Group A.

EXAMPLE 7

Use a scientific calculator to calculate the standard deviation of 9, 7, 11, 10, 13 and 7.

SOLUTION:

CASIO *fx-82ZA PLUS* calculator	SHARP EL-W535HT calculator
Press the following keys [MODE] [2: STAT] [1: 1 – VAR] 9 (=) 7 (=) 11 (=) 10 (=) 13 (=) 7 (=) [AC] [SHIFT : 1] [STAT] [4: VAR] [3: σx] [=]	Press the following keys: [MODE] [1: STAT] [0:SD] 9 [CHANGE] 7 [CHANGE] 11 [CHANGE] 10 [CHANGE] 13 [CHANGE] 7 [CHANGE] [ALPHA] [6: σx] [=]

So the Standard Deviation ≈ 2,141

EXERCISE 2.4

Where necessary, give decimal answers correct to 1 decimal place

1) The arm spans (in cm) of the eleven players in each of two different soccer teams A and B are recorded.
   a) The arm spans for **TEAM A** are:
      203, 214, 187, 188, 196, 199, 205, 203, 199, 194 and 206
      i) Calculate the mean of the arm spans using the formula: $\bar{x} = \frac{\Sigma x}{n}$.
      ii) Copy and complete the table given.
      iii) Calculate the standard deviation of the arm spans using the formula:
      $\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$.
   b) For **TEAM B**, the variance is 875 cm^2. Calculate the standard deviation of the arm spans of **TEAM B.**
   c) Make a comment about the dispersion of the arm spans of the players in both teams.

x	$x - \bar{x}$	$(x - \bar{x})^2$
203		
214		
187		
188		
196		
199		
205		
203		
199		
194		
206		
$n =$		$\Sigma(x - \bar{x})^2 =$

EXERCISE 2.4 (continued)

2) The time (in minutes) taken by a group of athletes from Lesiba High School to run a 3 km cross country race is: 18 21 16 24 28 20 22 29 19 23
Use your calculator to determine
   a) The mean time taken to complete the race.
   b) The standard deviation of the time taken to complete the race.

3) The following tables show the masses (in kilograms) of the A and B rugby teams at Sir John Adamson High School:

TEAM A

51	82	71	64	81	81	76	77	62	68
70	74	81	61	68	69	67	71	68	74
80	62	70	68	62					

TEAM B

83	79	67	79	87	62	60	83	76	79
94	110	73	97	70	68	103	85	74	55
47	63	62	87	74					

   a) Use your calculator to determine the mean and the standard deviations of each data set.
   b) Is the standard deviation a good measure for determining which team plays better? Give reasons for your answer.

4) As part of the Census@School, learners had to record the length (in centimetres) of their right foot without a shoe. The girls (G) and boys (B) in Grade 11C measured their foot lengths and recorded the results in the following table.

G:	29	22	28	23	23	29	29	25	27	23	27	21	24	21	20	25	22	29
B:	28	30	26	29	25	28	26	25	28	22	30	25	21	27	25	23		

   a) Use your calculator to determine the mean and standard deviation of the foot lengths of
      i) The girls
      ii) The boys.
   b) Use the mean and the standard deviation of the foot lengths to comment on the differences in foot sizes of the two groups.

SYMMETRIC AND SKEWED DATA

✓ A ***measure of shape*** describes the *distribution of the data* within a data set.

✓ A distribution of data values can be ***symmetric*** or ***skewed***.

- In a ***symmetric distribution***, the two sides of the distribution are a mirror image of each other

- In a ***skewed distribution***, the two sides of the distribution are NOT mirror images of each other.

✓ Both ***frequency polygons*** and ***box-and-whisker diagrams*** can be used to illustrate symmetric and skewed data.

KEY FEATURES OF A SYMMETRIC DISTRIBUTION

- ***The shape is symmetrical***
- ***The mode, median and mode have the same value***.
- ***Most of the data are clustered around the centre.***

  In fact, about 68% of the data lie within 1 standard deviation of the mean
  About 95% of the data lie within 2 standard deviations of the mean
  About 99,7% of the data lie within 3 standard deviations of the mean.

KEY FEATURES OF SKEWED DATA

Skewness is the tendency for the values to be more frequently around the high or low ends of the *x*-axis.

- With a ***positively skewed distribution***, the ***tail on the right side is longer*** than the left side

  Most of the values tend to cluster toward the left side of the x-axis (i.e. the smaller values) with increasingly fewer values on the right side of the x-axis (i.e. the larger values).

- With a ***negatively skewed distribution***, the ***tail on the left side is longer*** than the right side.

  Most of the values tend to cluster toward the right side of the x-axis (i.e. the larger values) with increasingly fewer values on the left side of the x-axis (i.e. the smaller values).

EXAMPLE 8

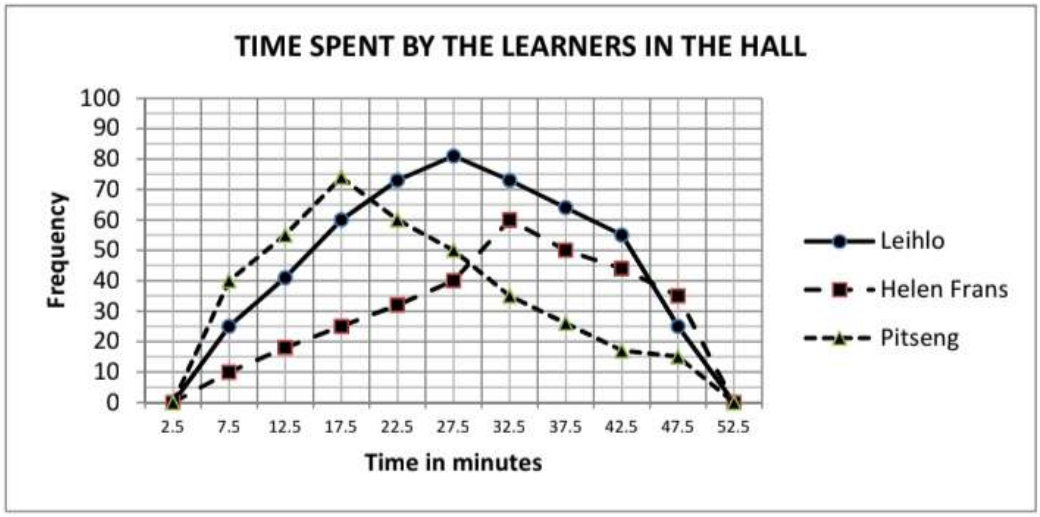
The Grade 10 learners of Leihlo Secondary School, Helen Frans Secondary School and Pitseng Secondary School attended a meeting at a hall in Senwabarwana about the problems they have encountered with the bus company which transports them to school.
The following table shows the time the learners spent in the meeting:

Time spent in the hall (in minutes)	Midpoint of the intervals (in minutes)	Frequency		
		Leihlo Secondary School	Helen Frans Secondary School	Pitseng Secondary School
$0 < t \leq 5$	2,5	0	0	0
$5 < t \leq 10$	7,5	25	10	40
$10 < t \leq 15$	12,5	41	18	55
$15 < t \leq 20$	17,5	60	25	74
$20 < t \leq 25$	22,5	73	32	60
$25 < t \leq 30$	27,5	81	40	50
$30 < t \leq 35$	32,5	73	60	35
$35 < t \leq 40$	37,5	64	50	26
$40 < t \leq 45$	42,5	55	44	17
$45 < t \leq 50$	47,5	25	35	15
$50 < t \leq 55$	52,5	0	0	0

a) Draw frequency polygons to represent the time spent by learners of each school in the hall.
b) Describe the shapes of the polygons.

SOLUTION:

a)



b) The data from Leihlo Secondary is symmetric
The data from Helen Frans is not symmetric. It is more spread out on the left and clustered more closely together on the right. We say that it is ***skewed left.***
The data from Pitseng is also not symmetric. It is more spread out on the right and clustered more closely together on the left. We say that it is ***skewed right***.

- ✓ Note that if the mean and the median of a data set are known, then
  - If ***mean – median ≈ 0,*** then the distribution is ***symmetric***
  - If ***mean – median > 0,*** then the distribution is ***positively skewed***
  - If ***mean – median < 0,*** then the distribution is ***negatively skewed***

- ✓ Note that for a box-and-whisker diagram
  - If the distribution is symmetric, the median is in the middle of the box and the whiskers are equal in length
  - When data is more spread out on the left side and clustered on the right, the distribution is said to be negatively skewed or skewed to the left.
  - When the data is more spread out on the right side clustered on the left, the distribution is said to be positively skewed or skewed to the right.

EXAMPLE 9

Use the data given in Example 8 for the following:

a) Calculate the mean and the five-number-summary for the time spent by learners in each school.
b) Draw box-and-whisker diagrams to represent the data.
c) State whether each data set is symmetric, positively skewed or negatively skewed.

SOLUTION:

Leihlo Secondary School

a) $\bar{x} \approx 27{,}5$ minutes
Minimum value ≈ 7,5 minutes
Lower quartile = $Q_1 \approx 17{,}5$ minutes
Median ≈ 27,5 minutes
Upper quartile = $Q_3 \approx 37{,}5$ minutes
Maximum value ≈ 47,5 minutes

b)

c) Mean – median = 27,5 minutes – 27,5 minutes = 0
This means that the distribution is symmetric.

EXAMPLE 9 (continued)

Helen Frans Secondary School

a) $\bar{x} \approx 31{,}6$ minutes
Minimum value $\approx 7{,}5$ minutes
Lower quartile = $Q_1 \approx 22{,}5$ minutes
Median $\approx 32{,}5$ minutes
Upper quartile = $Q_3 \approx 37{,}5$ minutes
Maximum value $\approx 47{,}5$ minutes

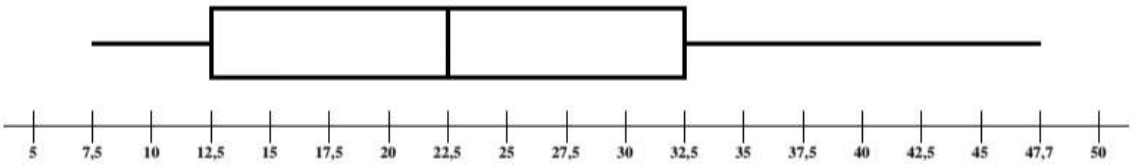
b)

c) Mean – median = 31,6 minutes – 32,7 minutes = – 1,1
This means that the distribution is negatively skewed (or skewed left).

Pitseng Secondary School

a) $\bar{x} \approx 22{,}99$ minutes
Minimum value $\approx 7{,}5$ minutes
Lower quartile = $Q_1 \approx 12{,}5$ minutes
Median $\approx 22{,}5$ minutes
Upper quartile = $Q_3 \approx 32{,}5$ minutes
Maximum value $\approx 47{,}5$ minutes

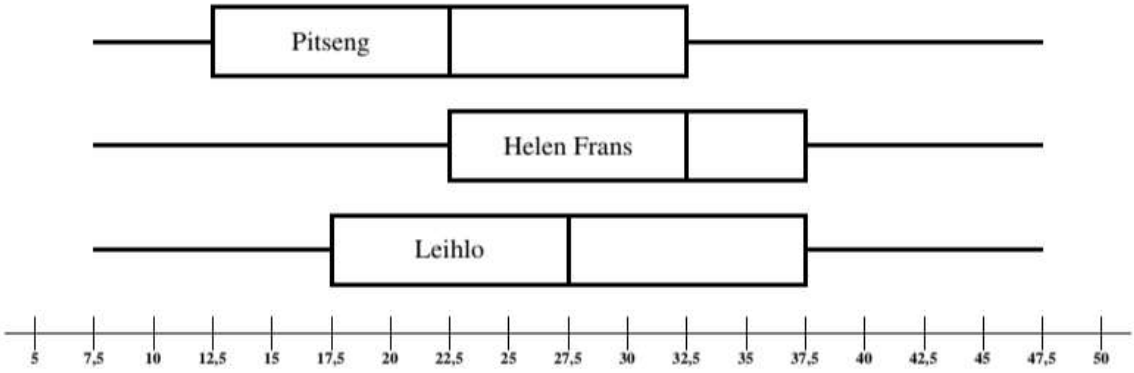
b)



c) Mean – median = 22,99 minutes – 22,5 minutes = 0,49
This means that the distribution is positively skewed (or skewed right).

EXAMPLE 9 (continued)

When we draw all three box and whisker diagrams on the same page, we can immediately see that the Leihlo data is symmetric, the Helen Frans data is negatively skewed, and the Pitseng data is positively skewed.



EXERCISE 2.5

1) The box and whiskers diagrams of two sets A and B are shown below.

a) Write down what is common to both sets of data.
b) Which data set is symmetrical? State the reasons.
c) Is the other data set skewed left or right? State the reasons.

2) For the 2009 Census@School, 47 Grade 11 learners recorded how long (in minutes) it took them to travel to school. The following data was obtained:

Time (in minutes)	Frequency
$5 < t \leq 10$	1
$10 < t \leq 15$	5
$15 < t \leq 20$	9
$20 < t \leq 25$	13
$25 < t \leq 30$	11
$30 < t \leq 35$	8

a) Use the given information to determine the five number summary.
b) Draw a box and whisker diagram to illustrate the five number summary.
c) Comment on the spread of the time taken to complete the task.

EXERCISE 2.5 (continued)

3) Three high schools in Limpopo have a total number of 132 Grade 12 learners. These learners completed the question in the 2009 Census@School where they were asked to record the distance (in kilometres) they travel each day from home to school. The results of the survey are shown in the grouped frequency below.

Distance in kilometres (x)	***Frequency***	***Midpoint of intervals***
$0 < x \leq 5$	12	
$5 < x \leq 10$	29	
$10 < x \leq 15$	13	
$15 < x \leq 20$	63	
$20 < x \leq 25$	12	
$25 < x \leq 30$	3	

a) Copy and complete the table.
b) Draw a frequency polygon to illustrate the data.
c) Determine the median of the data in the table.
d) Use your calculator to determine the mean of the data.
e) Calculate ***mean – median***
f) By referring to the shape of the polygon and the relationship between the mean and the median, state whether the distribution of the data is symmetric, positively skewed or negatively skewed.

OUTLIERS

- ✓ An ***outlier*** is a data entry that is ***far removed from the other entries*** in the data set e.g. a data entry that is much smaller or much larger than the rest of the data values.

- ✓ An outlier has an influence on the ***mean*** and the ***range*** of the data set, but has no influence on the median or lower or upper quartiles.

- ✓ An outlier can affect the ***skewness*** of the data.

- ✓ Any data item that is **less than $Q_1 - 1{,}5 \times IQR$** <u>OR</u> **more than $Q_3 + 1{,}5 \times IQR$** is an outlier.

EXAMPLE 1

Investigate the following data set:

1, 8, 12, 14, 14, 15, 17, 17, 19, 26, 32

a) Calculate (where necessary correct to 1 decimal place)
   i) The mean
   ii) The median
   iii) The interquartile range

b) Are any of the entries in the data set outliers?

<u>SOLUTION</u>:

a)

i) Mean $= \bar{x} = \frac{1+8+12+14+14+15+17+17+19+26+32}{11}$

$= \frac{175}{11}$

$= 15{,}9090...$

$\bar{x} \approx 15{,}9$

ii) There are 11 terms so the median is the 6^{th} term.
Median = 15

iii) There are 5 terms less than the median so Q_1 is the 3^{rd} term. So $Q_1 = 12$.
To find Q_3 we add 3 terms to the position of the median and get the 9^{th} term.
So $Q_3 = 19$

IQR = 19 – 12 = 7

b) Lower outlier $< \mathbf{Q_1 - 1{,}5 \times IQR}$

$< 12 - 1{,}5 \times 7$

$< 1{,}5$

So 1 is an outlier

Upper outlier $> \mathbf{Q_3 + 1{,}5 \times IQR}$

$> 19 + 1{,}5 \times 7$

$> 29{,}5$

And 32 is also an outlier

EXERCISE 2.6

1) Determine the interquartile range and then find outliers (if there are any) for the following set of data:
   10,2 ; 14,1 ; 14,4 ; 14,4 ; 14,5 ; 14,5 ; 14,6 ;
   14,7 ; 14,7 ; 14,9 ; 15,1 ; 15,9 ; 16,4 ; 18,9

2) A class of 20 learners has to submit Mathematics assessment tasks over the course of the year. While some learners were conscientious others were not.

   The following table shows the number of assessment tasks each learner handed in:

9	5	11	8	12	2	6	9	15	10
12	6	9	3	9	13	14	16	4	7

   a) Determine the IQR
   b) Determine the outliers (if any).

3) The following are the ages of boys in one of the Grade 8 class of Dendron Secondary School:
   12 12 13 14 14 13 12 15 15 14 12 19 14 12 9
   a) Determine the five number summary.
   b) Determine the outliers, if any.

REFERENCES

Bowie L. et al. (2007). *Focus on Mathematical Literacy Grade 12*. Maskew Miller Longman.
Freund J. E. (1999). *Statistics A First Course*. Prentice Hall, New Jersey.
Larson R. and Farber B. (2006). *Elementary Statistics Picturing the World*. Third Edition. Pearson, Prentice Hall.
Upton G and Cook I (2001) *Introducing Statistics 2nd edition*. Oxford
Statistics South Africa (2010) *Census At School Results (2009).*
The Answer Series *Grade 12 Mathematics Paper 3, notes, questions and answers.*