Unit 3: Data representation

Introduction to Unit 3

In this unit you will look at different ways to represent data in tables, charts, graphs and diagrams. The emphasis is not on the techniques to produce these representations, but on the question of whether or not the representation best represents the data.

Purpose of Unit 3

The aim of this unit is to look at a variety of ways to represent data and to compare these for the best representation of the data given. The unit will look at : frequency tables, pictograms, bar charts, line bar charts, histograms, pie charts, line graphs, frequency polygons, stem-leaf plots, scatter plots.

Objectives

At the end of this unit you should be able to:

- organise data
- describe data
- read and interpret displays of data
- construct appropriate displays of data: frequency table, pictogram, bar chart, line bar chart, histogram, pie chart, line graph, frequency polygon, stem-leaf plots, scatter plots
- justify the choice of display used for given data
- critically analyse data displays
- state common pupil errors in data representation
- illustrate methods to misrepresent data
- use appropriate project work in the classroom to assist the pupils in their learning of data representation

Time

To study this unit will take you about 10 hours.

Unit 3: Data representation

Section A: Represent or model data

What is the best way to represent the collected data? Is the data discrete or continuous, is the data qualitative or quantitative, how does one change from one form of representation to another, what is the effect of changing scale? These are questions to be considered. Too frequently the emphasis is on operational understanding, on the techniques of drawing a bar chart, a pie chart, a cumulative frequency curve while questions as to why to use the (most of the time) stated representation in the given circumstances (are they valid? are they appropriate?) are hardly considered. It should be left to the pupils to decide what is the most appropriate way to represent their data (and that is the difficult part—the actual drawing of chart is not the problem in general, and could be done by a computer). One way to do this is by comparing different forms of representing the data.

Graphical displays should:

- show the data
- induce the viewer to think about substance rather than about methodology
- represent large data sets in a relatively small space
- make large data sets coherent
- encourage pupils to make comparisons between different pieces of data
- reveal the data at several levels of detail
- serve a reasonably clear purpose
- be integrated with the statistical and verbal descriptions of the data

Section B: Tables

Data collected is generally first tabulated in frequency distribution tables. These tables might contain data that is grouped or ungrouped. Sometimes two-way tables are used.

These were covered in the previous Unit 2, section D1.

Section C: Nature and format of data

The type of representation that can be used depends on

a) the nature of the data, i.e., discrete or continuous data
b) the format in which the data is given ungrouped or grouped

Discrete data

Discrete data can be displayed in bar charts (categorical data), bar-line graphs (discrete quantitative data) or pie charts (categorical data / discrete quantitative, provided the number of categories or discrete values is not large).

In a **bar graph** or **bar-line graph** the height of the bar or line is proportional to the frequency.

Bars are to be drawn separated equally, with same width. The discrete value or category is placed at the centre of the bar. The frequencies, along the vertical axis, are placed against the lines (NOT the spaces). Bar-line graphs are very appropriate with discrete data (number of children in the family, shoe size of pupils, etc.), bar graphs (also called frequency diagrams) are more appropriate for grouped discrete data or for categorical data.

In a **pie chart** the angle at the centre of each sector is proportional to the frequency. Therefore the radius of the pie chart is not relevant. The number of sectors should, generally, not exceed 6 - 8 to make the presentation meaningful and allow comparison between the various sectors.

Continuous data

Continuous data is best displayed in histograms. In a histogram the frequencies are proportional to the area of the bar. In cases where bars of the same width are considered the histogram becomes a bar graph, but the bars touch each other. Details will be discussed below.

N.B. It is rather common to display certain discrete data (for example, scores on a test, number of children in a family, i.e., numerical data that can be ordered) in a bar graph with the bars touching each other. This is strictly speaking not correct, but you should not try to make the distinction with students of this age.

Independent vs. dependent variables

An **independent variable** is presumed to have an effect on another variable. It is the variable that is manipulated or changed by the researcher to investigate the effect on a **dependent variable**. It is also known as the **manipulated** or **experimental variable** that we have discussed above. The effect of the manipulation is observed on the dependent variable. The independent variable is a variable that by itself does not necessarily give rise to the behaviour of interest except if manipulated.

The dependent (or outcome) variable is that variable which occurrence or frequency of occurrence depends on the conditions and the manipulation of the independent variable. It is called the dependent variable because its value depends on and varies with the value of the independent variable.

The independent variable is commonly plotted along the horizontal axis and the dependent variable along the vertical axis.

Write down the different data representations (charts, graphs, diagrams) you remember.

1. What type of data is most appropriately represented by each of the representations you listed above?
2 What type of data cannot be represented by each of the representations you listed?

Section D: Graphical representations

Data can be represented in various ways, and in the following sections you are going to look at the following representations of data.

- Bar charts (Section D1)
- Line/stick graphs (Section D2)
- Histograms (Section D3)
- Pie charts (Section D4)
- Pictograms (Section D5)
- Line graphs/charts (Section D6)
- Frequency polygons (Section D7)
- Stem-leaf diagrams (Section D8)
- Scatter diagram (Section D9)

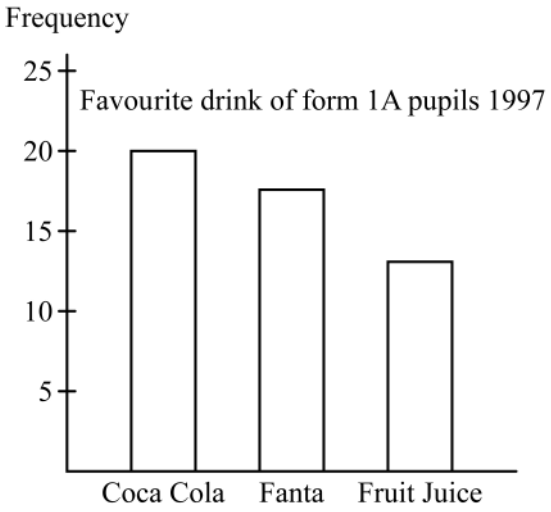
In each section special attention will be given to the type of data that can be represented in that particular way.

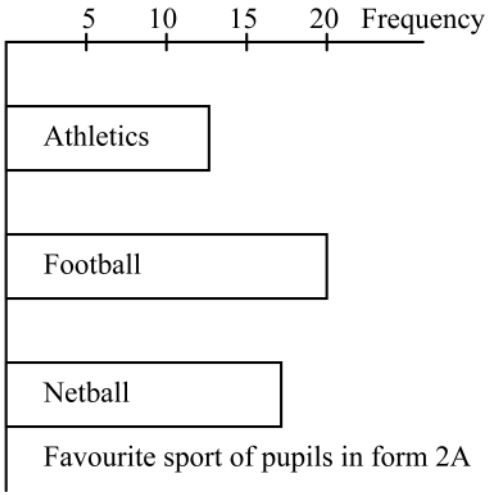
Section D1: Bar charts (bars horizontal or vertical)

Most appropriate use: to compare categories (qualitative data, the independent variable is non-numerical) and grouped discrete quantitative data (scores on a test, amount spend by customers in a shop)

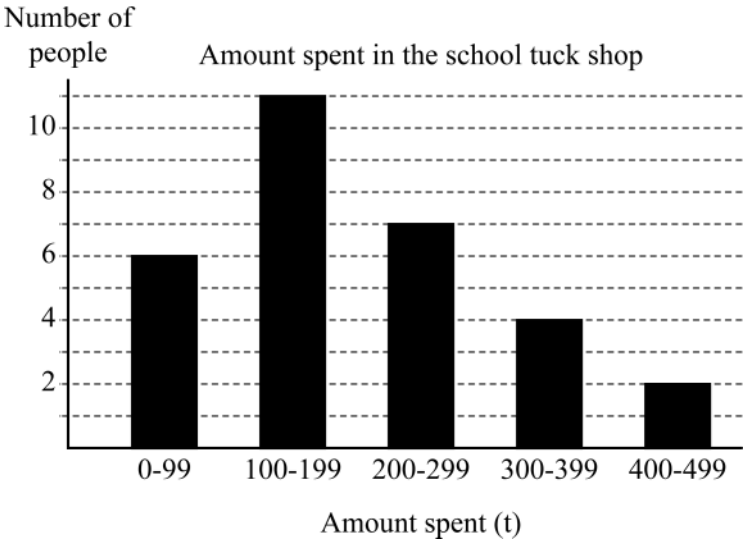
How to draw: Rectangles with equal width are used. The height/length represents the frequency of the category. Do not draw the bar adjacent. Label the diagram as a whole (title), the bars and the frequency axis. Indicate scale on the frequency axis.

Examples of qualitative data display

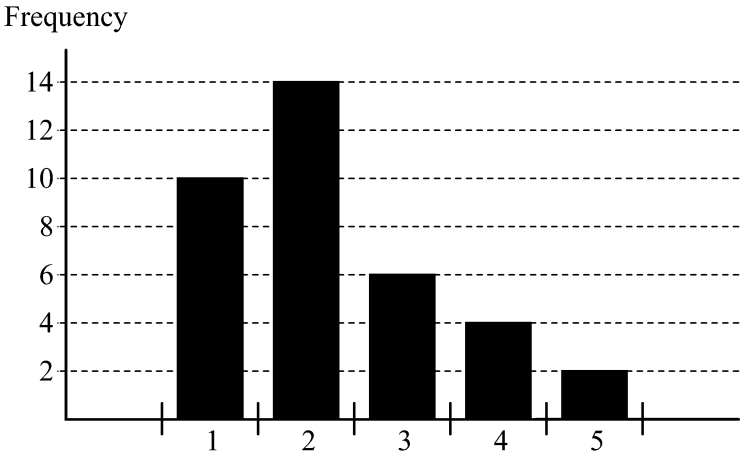




Examples of quantitative discrete (grouped) data display.



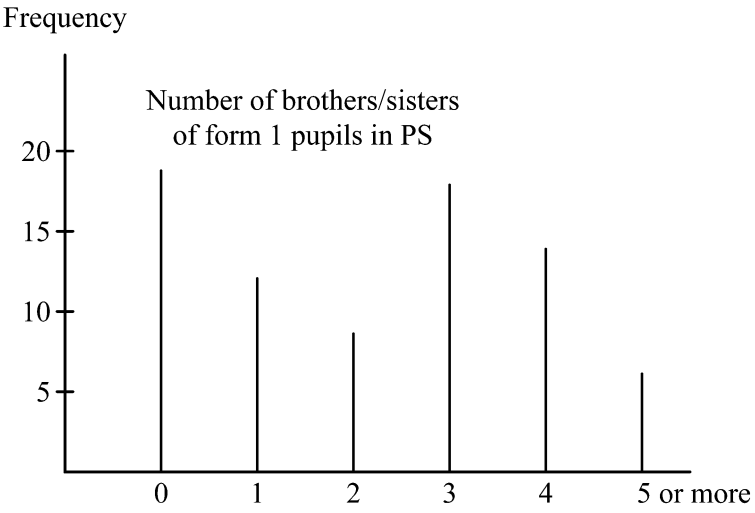
Number of customers in a group



Section D2: Line/stick graphs (can be horizontally or vertically displayed)

Most appropriate use: to compare discrete variables

How to draw: lines/sticks of length proportional to the frequencies.Labelling as with the bar graph.



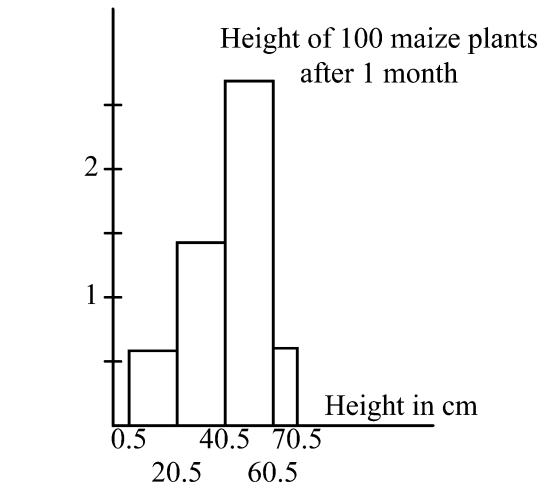
Section D3: Histograms

Most appropriate use: to represent **grouped continuous** variables. Always depicts frequency (or count) versus a continuous or nearly continuous variable.

How to draw: Rectangles whose areas are proportional to the frequencies. The rectangles are adjacent (that is, the rectangles touch each other.) The axes are labelled, the graph has a title.

Example: The height of 100 maize plants was measured, to the nearest cm, one month after planting.

Height of maize plants	Frequency	Frequency density
1 - 20 cm	12	0.6
21 - 40 cm	28	1.4
41 - 60 cm	54	2.7
61 - 70 cm	6	0.6



N.B.

(i) Different notations for the classes are in use, 1 - 20 standing for heights from 1 to 20 both inclusive in the above case. Also the notation [1, 20] or $1 \leq \text{height} \leq 20$ can be used.

Some books use as the first class 0 - 20 to mean $0 \leq \text{height} < 20$ and write the next class as 20 - 40 to imply $20 \leq \text{height} < 40$, etc. The notation to be used is a matter of agreement.

(ii) Attention is to be paid to the upper and lower boundaries. The context dictates how they have to be taken.

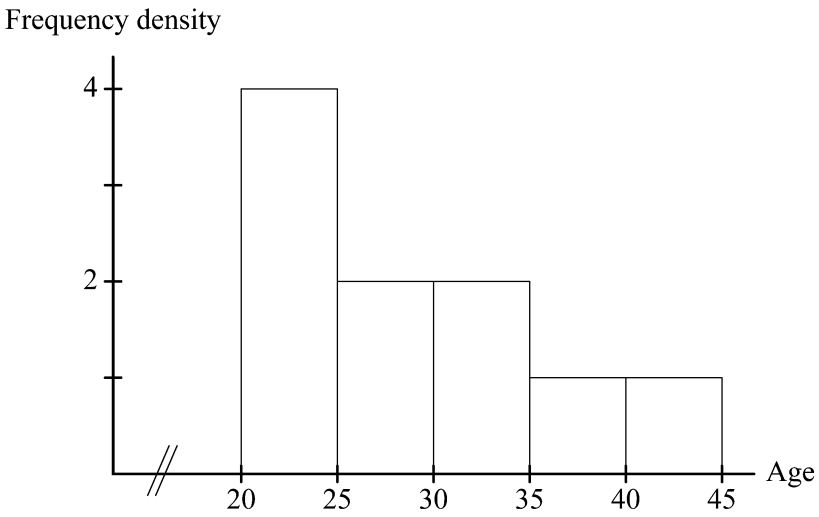
In the above example of measurements to the nearest cm, the boundaries are half way between two classes. The rectangles are to be drawn—in the above example—from 0.5 (the lower boundary) to 20.5 (the upper boundary), from 20.5 to 40.5 and the last one from 60.5 to 70.5.

In case the variable is age (a continuous variable) the situation is different. Ages are given in completed years, not to the nearest year. A person of 20 years and 11 months and 25 days is still considered to be 20. Consider the following example.

The ages of applicants for a teaching post have the following distribution.

Age	Frequency
20 - 24	4
25 - 29	2
30 - 34	2
35 - 44	2

In the class 20 - 24 fall all applicants with age $20 \leq \text{age} < 25$, in the class 25 - 29 fall all applicants with ages $25 \leq \text{age} < 30$, etc. The class boundaries are 20, 25, 30, 35, 45.



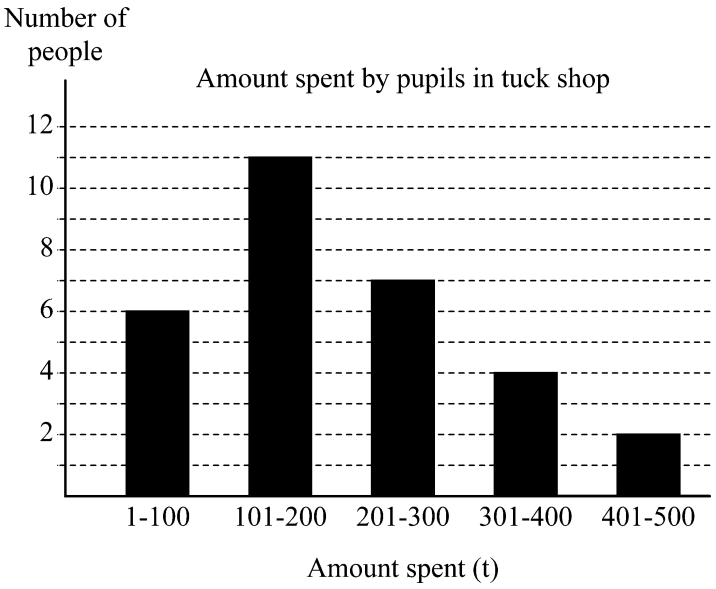
(iii) The frequency density is the frequency divided by the class width (upper boundary – lower boundary of the class). This is a fine point, probably one that you should not teach.

(iv) In cases that classes have all the same width, the frequency density and frequency are directly proportional and some authors will label the axis: "frequencies" in that case.

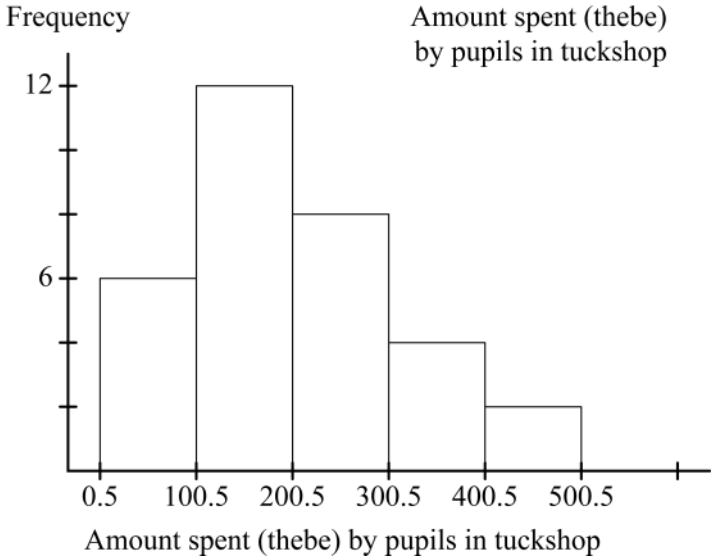
(v) Discrete grouped data (test scores, amount of money spent in a shop) should be displayed in bar graphs. However it is rather common practice to display grouped discrete data as if continuous.

Amount spent by pupils in the tuckshop (thebe)	Frequency
1-100	6
101 - 200	11
201 - 300	7
301- 400	4
401- 500	2

The data is best displayed on a bar chart.



However the data is also displayed in histograms - taking the data as if it is continuous. Class boundaries 0.5 /100.5/200.5, etc. are then correct, but for students of this age, whole-number boundaries on a histogram would be "close enough."



(vi) Grouping data is a means to summarise the raw data. Be aware that by grouping some of the original information is lost.

If, for example, in a test marked out of 10 the scores were:

Mark	0	1	2	3	4	5	6	7	8	9	10
Frequency	3	3	3	2	3	3	4	3	3	2	1

Then by grouping:

Mark	Frequency
0 -1	6
2 - 3	5
4 - 5	6
6 - 7	7
8 and more	6

some of the original information can no longer be found in the grouped frequency table. This has implications for calculation of mean / median / mode. These measures obtained from the raw data will differ from (approximated) values obtained from the grouped frequency table. Changing the class width will again lead to different approximations for the measures of central tendency. Grouping results in what is referred to as "grouping error." The error is reduced by using small class intervals. If the class intervals are increased so does the 'grouping error' in the approximation for the mean and median obtained from the grouped frequency table.

Self mark exercise 1

1. Represent the following data in a bar chart.

   The month in which form 1 pupils in a school were born are tabulated in the frequency table below.

Mon.	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Freq.	3	33	21	6	18	30	24	18	54	21	6	9

2. Represent the following data in a bar chart.

   The amount (in thousands of litres) of petrol sold at a petrol station during a month was

Type of petrol	Leaded	Unleaded	Diesel
Number of litres (x 1000)	45	35	20

3. Represent the following data in a bar chart.

   The percent of pupils obtaining a certain grade in a test are listed:

Grade	A	B	C	D	E	F	G
% of pupils	2	5	36	24	15	9	4

4. Use the following raw data of the length (mm) of nails found in packets of 'assorted nails'.

11	48	53	32	28	15	17	45	37	41
55	31	23	36	42	27	19	16	46	39
41	28	43	36	21	51	37	44	33	40
15	38	54	16	46	47	20	18	48	29
31	41	53	18	24	25	20	44	13	45

   a) Make a grouped frequency table taking class intervals 10 -14, 15 - 19, etc., and draw a histogram.

   b) Make a grouped frequency table taking class intervals 10 - 19, 20 -29, etc., and draw the histogram.

      Compare the two representations of the data.

Suggested answers are at the end of this unit.

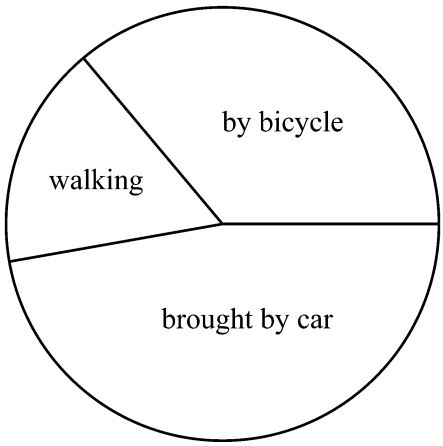
Section D4: Pie charts

Most appropriate use: to represent data as part of a whole, to illustrate differences in categories (qualitative or discrete variables) provided the number of categories is limited (generally between 2 and 8).

How to draw: Measure of the angle at the centre of the circle is proportional to the frequency

(measure of the angle at the centre = $\frac{\text{frequency of the category}}{\text{total of all frequencies}} \times 360^{\circ}$)

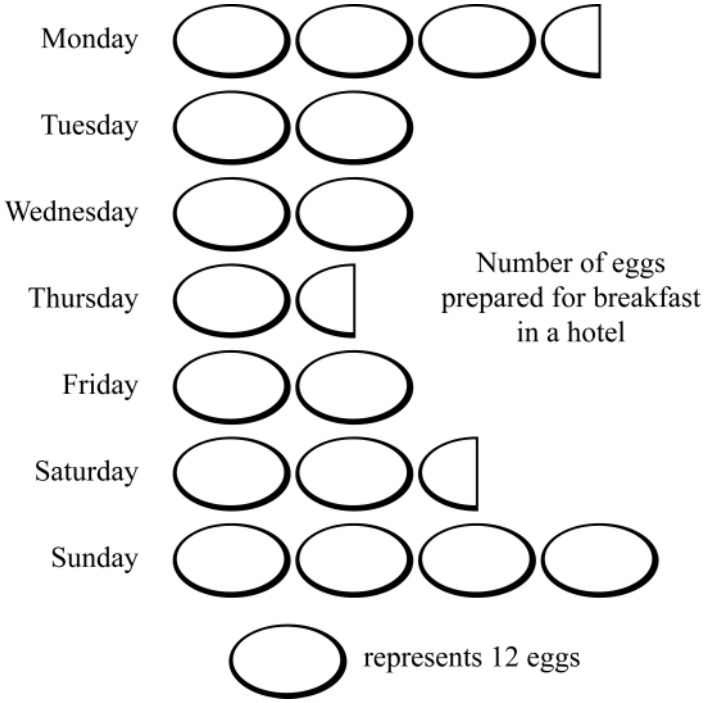
How the form 1 pupils come to Marua Pula



Section D5: Pictograms

Most appropriate use: to illustrate broad differences between categories (qualitative and discrete variables).

How to draw: Draw simple pictures (instead of bars) to represent the frequency. A key is to be added to show what each picture represents.



Self mark exercise 2

1. Display the following data in a pie chart and pictogram.

   The total world wool production was distributed over various countries as follows in 1994:

Country	% of wool world production produced
Australia	30%
USSR	30%
New Zealand	20%
Argentina	10%
SA	10%
Others	10%

2. Display the following data in a pie chart and pictogram.

   The type of vehicles coming to a petrol station during one day are tabulated below

Person cars	26
Lorries	12
Busses	8
Combis	14

3. Display the following data in a pie chart and pictogram.

   The sizes of T-shirts sold during a month in a shop were

Size	S	M	L	XL
Number sold	12	34	56	18

4. Comparing the two representations, pie chart and pictogram, list some advantages and disadvantages of each.

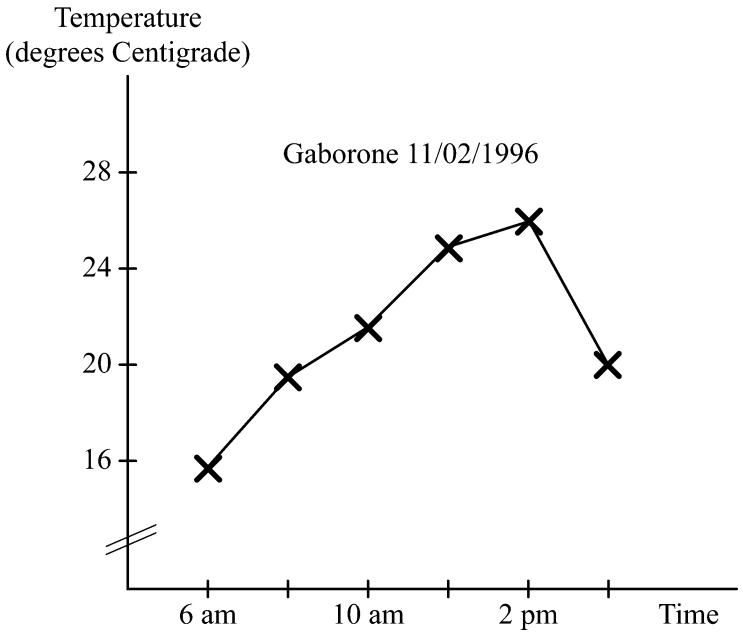
Suggested answers are at the end of this unit.

Section D6: Line graphs/charts

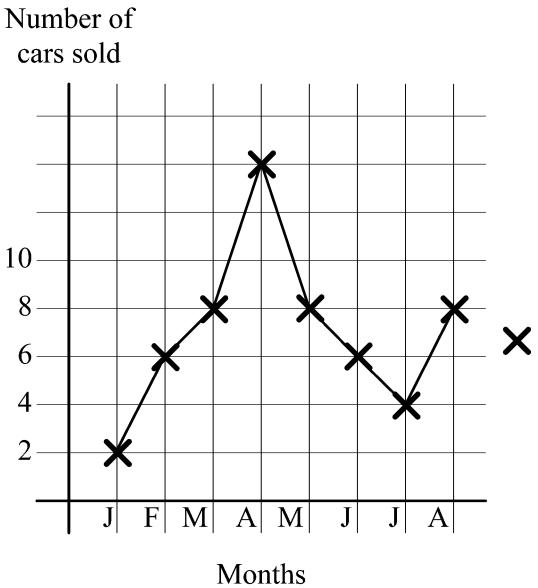
Most appropriate use: to illustrate changes of continuous variables (over time)

How to draw: plot the given corresponding pairs of data as points and join consecutive points by line segments.

Example



If **trends**—changes over time—are looked for, **line graphs** can be used. Line graphs are used for both discrete and continuous data. For example, in the line graph is displayed the number of cars sold in a certain garage over the first 8 months of a year. Although 'in between' values such as $2\frac{1}{2}$, $3\frac{1}{4}$, etc., do not exist the points are joined with straight lines to show the trend. For trend lines not all 'in between' values have to make sense. Trend lines should not be confused with linear graphs. In linear graphs the 'in between' values must exist, otherwise it would be inappropriate to draw the line.



Self mark exercise 3

1 Represent the following data in a line graph and comment on the trend.

Number of pupils in a primary school 1993 - 1999

Year	1993	1994	1995	1996	1997	1998	1999
Number	450	435	465	478	490	510	524

2 a) Represent the following data in a line graph and comment on the trend.

Infant mortality rate per 1000 live births

Year	1971	1981	1991
Infant mortality rate	100	71	45

b) If the trend is continuous what do you expect the infant mortality rate to be in 2001?

3 The number of teacher trained for the Senior Secondary School in Botswana are tabulated:

Year	1996	1997	1998	1999
Number trained	81	105	115	184

a) Represent this data in a line graph and comment on the trend.

b) If the trend is continuous what number of Senior Secondary school teachers do you expect to be trained in the year 2000?

Suggested answers are at the end of this unit.

Section D7: Frequency polygons

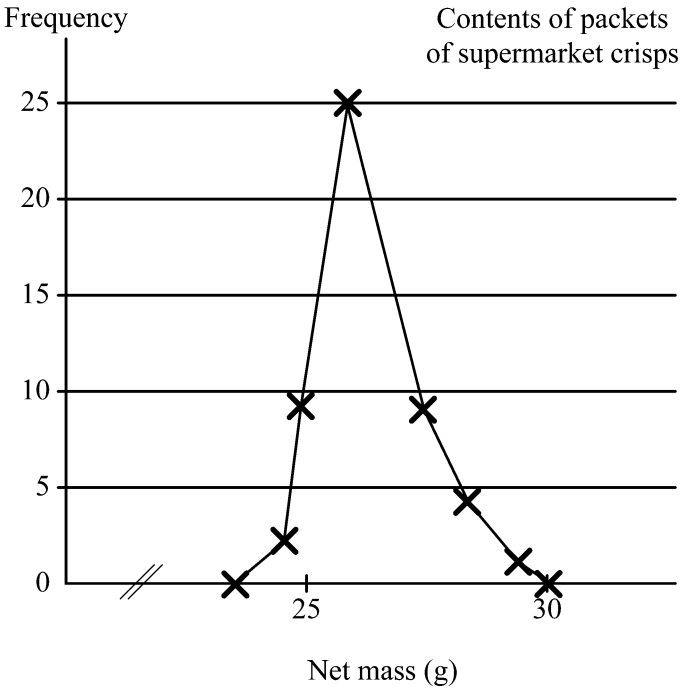
Most appropriate use: to *compare* **grouped continuous** variables, for example, distribution of the height of girls (in grouped form) and distribution of the height of boys (in grouped form). Drawing the frequency polygons on the same axes allows easy comparison. This is far superior to drawing two histograms on one grid!

How to draw: Plotting the points with co-ordinates (mid-interval value, frequency) and connecting these with straight line segments. Adding points on the horizontal axis at both sides.

Example: The following grouped frequency table summarises the net content of 50 packets of a supermarket's brand of potato crisps labelled 25 grams.

Net mass ($\times$ g)	Mid-point	Frequency
$23.5 \leq \times < 24.5$	24.0	2
$24.5 \leq \times < 25.5$	25.0	9
$25.5 \leq \times < 26.5$	26.0	25
$26.5 \leq \times < 27.5$	27.0	9
$27.5 \leq \times < 28.5$	28.0	4
$28.5 \leq \times < 29.5$	29.0	1
	TOTAL	50

A frequency polygon is drawn by plotting the frequency for each interval against the mid-point of the interval and joining the points with straight line segments. The points (24.0, 2), (25.0, 9), ... (29.0, 1) are plotted and joined. The polygon is continued to the axis so the points (23.0, 0) and (30.0, 0) are included.



Self mark exercise 4

1. A factory producing hats measured the circumference of the heads of 100 people in centimetres to the nearest centimetre. The following results were obtained.

Head circumference c (cm)	Frequency
$50 \le c < 52$	8
$52 \le c < 54$	12
$54 \le c < 56$	30
$56 \le c < 58$	44
$58 \le c < 60$	6

   a) Represent the data in a frequency polygon.

   b) If you are to give advice to the factory manager as to the size of hats to be produced what advice would you give? Explain.

2. a) Represent the following data in a frequency polygon. The data gives the mass of apples from one tree.

Mass of apple (g)	20 –30	30-40	40-50	50-60	60-70
Frequency	6	18	34	30	12

   b) Using the same scale and axes, represent the following data giving the mass of the same type apples from another tree in the orchard.

Mass of apple (g)	20 –30	30-40	40-50	50-60	60-70
Frequency	3	14	26	36	21

   c) Comparing the two polygons what conclusion(s) can you make? Explain.

Suggested answers are at the end of this unit.

Section D8: Stem-leaf diagrams

Most appropriate use: To represent ungrouped quantitative data. Also to compare two sets of ungrouped quantitative data (male and female data on the same variable for example). Stem-leaf diagrams are the only graphical representations that also display all the original data values.

How to draw: Part of the number, often the whole number part, is used as the stem (placed vertically under each other), the other part of the number forms the leaves. Place an explanatory legend beneath the diagram and a title above it.

Example:
Potato crisps come in packets marked 25 grams. The mass of 25 packets was found to the nearest 0.1 g.

26.4	25.2	26.3	26.0	24.1
25.3	25.6	26.2	27.8	24.5
25.0	27.5	25.8	26.0	25.7
25.5	26.4	25.5	24.7	26.9
27.3	25.3	25.1	27.7	26.8

Data can be organised in stem-leaf diagram. The whole number part of the mass can be used to form the stem shown at the left of the vertical line and the decimal part of the masses forms the leaves on the right. The leaves on each level or row in the diagram increase in value outwards from the stem.

Contents of packets of crisps (g)

24	157
25	0123355678
26	00234489
27	3578

$n = 25$ 24 | 1 represents 24.1 gram

The 'scale' is very important as 24 | 1 could mean 241, 24.1, 2.41, 0.241, etc., depending on the quantities displayed. The diagram has a title and the sample size $n = 25$ is noted.

The stem-leaf diagram can be stretched by choosing the stem to represent more levels.
For example
24.0 to 24.4
24.5 to 24.9
25.0 to 25.4
etc.

Contents of packets of crisps (g)

24	1
24	57
25	01233
25	55678
26	002344
26	89
27	3
27	578

$n = 25$ 24 | 1 represents 24.1 gram

The stem-leaf diagram can also be made 'double' allowing comparison.
For example, the following stem-leaf diagram gives the height of pupils in a class. The girls are on the left, the boys on the right.

Height of pupils in Form 2X

Girls		Boys
443310	15	2
9865	15	579
43220	16	12234
865	16	5566889
42	17	044
	17	58

$n = 41$ 16 | 8 represents 168 cm

The diagram makes visual that on the whole the girls are shorter than the boys. Or does it?

Self mark exercise 5

1. A sample of eggs from an one day's production has mass in grams:

40	50	72	51	60
55	67	46	57	53
55	42	51	59	49
52	46	64	43	66
54	64	48	58	52

   Draw a stem-leaf diagram using stems 4, 4, 5, 5, 6, 6, 7

2. Represent the following data in (a) a grouped frequency table (b) a histogram.

   Contents of packets of crisps (g)

24	14
24	57
25	0122233
25	55556778888
26	0000112233344
26	56789
27	001134
27	5678

   $n = 50$ 27 | 0 represents 27.0 gram

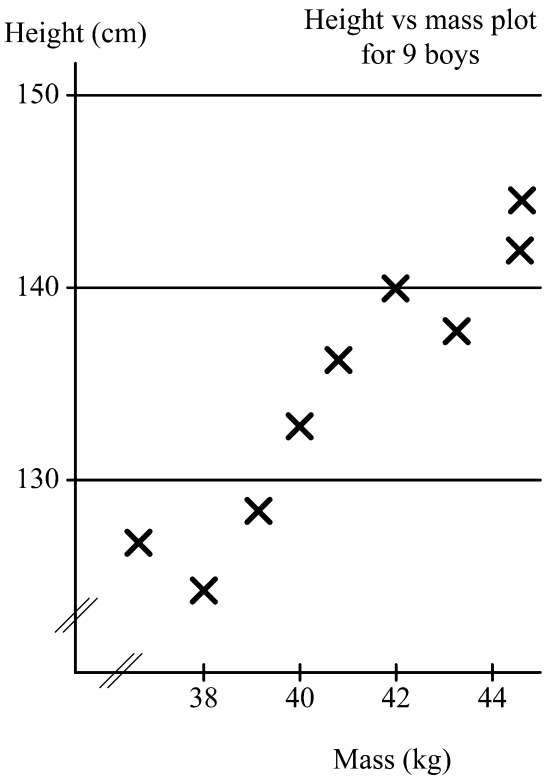
Suggested answers are at the end of this unit.

Section D9: Scatter diagrams

Most appropriate use: When looking at statistical data it is often observed that there are connections between sets of data. For example the mass and height of persons are related: the taller the person the greater his/her mass. To find out whether or not two sets of data are connected **scatter diagrams** can be used.

How to draw: In a scatter diagram each plotted point represent a pair, for example a (mass, height) pair of one person.

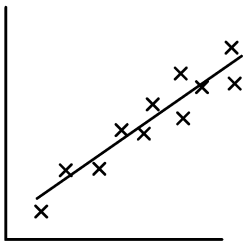
Example



The scatter graph illustrates that generally taller boys have greater mass.

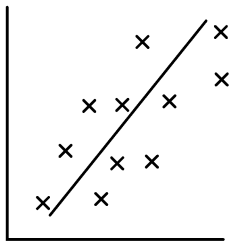
The relationship in a scatter diagram between the two sets of variables is described with the word **correlation**.

Strong positive correlation



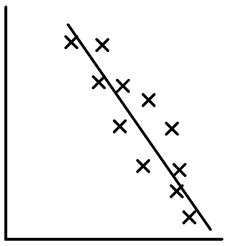
Points are clearly clustered around a line with positive gradient

Weak positive correlation



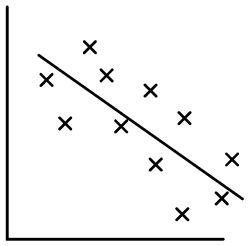
Points are roughly clustered around a line with positive gradient

Strong negative correlation



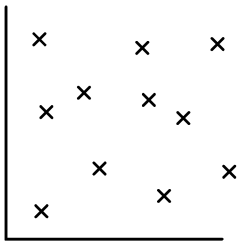
Points are clearly clustered around a line with negative gradient

Weak negative correlation



Points are roughly clustered around a line with negative gradient

No correlation



Points DO NOT cluster around a line

Positive correlation between two variables x and y can be described in words as: if x increases then y will also increase or if x decreases then y will decrease (x and y are directly proportional). Negative correlation between two variables x and y can be expressed in words as: if x increases y will decrease or if x decreases then y will increase (x and y are inversely proportional).

Strong positive or negative correlation between two sets of data *does not* prove that the two variables are **causal** related. For example, the length of a spring and the mass attached to it are likely strongly positively correlated and a greater mass attached causes the spring to extend more. If in a scatter diagram a positive correlation was found between the scores in mathematics of the pupils in a class and the distance they stay from the school (those staying close to school score low, those staying far from school score high) then it is very unlikely that there is any causal relationship (if a pupil moves to a place far from school his/her marks in mathematics are unlikely to increase!). This type of non causal correlation is called **spurious**

correlation, and is surprisingly common.

Estimating values:

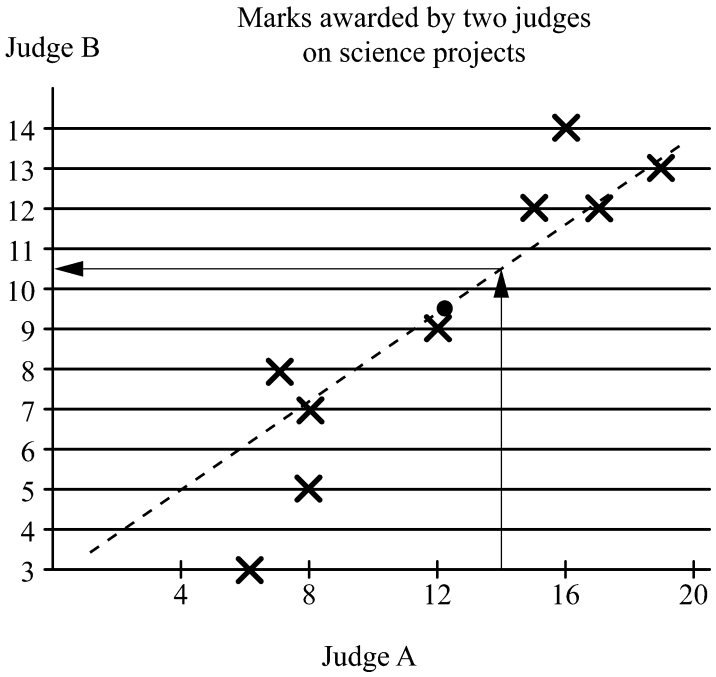
If two sets of data show correlation you can use your scatter graph to estimate missing values. You draw the 'best fitting' line through the point with co-ordinates (mean value of x, mean value of y).

Example:

Two judges awarded marks in a science fair for projects. Judge A scored out of 20 and judge B scored out of 15.

Judge A	15	12	8	19	7	6	17	8	15	16
Judge B	12	9	7	13	8	3	12	5	12	14

Plotting these data in a scatter graph gives the following graph.



The line of best fit is drawn through the point (12.3, 9.5) as the mean score of judge A is 12.3 and the mean score of judge B is 9.5 and such that about the same number of points is at each side of the line.

Judge A scored a project 14 marks but the project was not seen by judge B. You can now use the line of best fit to obtain an estimate for the score of judge B. You find 14 on the Judge A axis. Follow the arrows in the diagram to find the estimate for the mark judge B most likely should have given: 10.4 (or rounded to 10 as only whole marks were awarded).

Self mark exercise 6

1. Do you expect positive (strong or weak), negative (strong or weak) or no correlation between the following variables? Justify your answer.
   a) Number of days absent from school and the mark in the examination
   b) Number of boys and number of girls in a family
   c) Level of education and income
   d) Income and time spent in bars
   e) Age and number of pages read in one hour
   f) Number of rooms in a house and the number of doors
   g) Arm length and length of javelin throw
   h) Circumference of head and intelligence
   i) Length of spring and mass attached to its end
   j) Amount of pocket money and number of friends
   k) Age of a car and its second-hand selling price
   l) The shoe sizes of pupils and the distance they travel to school
   m) Depth of tread on a car tyre and the distance travelled
2. During the term 10 pupils obtained the following scores in mathematics and science (out of 50).

Maths	29	38	22	27	32	41	20	32	36	31
Science	27	34	26	34	27	42	21	28	34	29

   a) Draw a scatter graph to represent the data.
   b) Are pupils' score in mathematics and science correlated?
   c) Find the mean of the maths scores and the science scores to draw the line of best fit.
   d) A pupil scored 35 in mathematics. Work out an estimate for the pupil's score in science.
   e) Another pupil scored 40 in science. Work out an estimate for the pupil's score in mathematics.

3. A supermarket asked randomly some of their customers how many times they come to the supermarket in a three month period and how far away they live from the supermarket.
   The results are tabulated below:

Number of visits	Distance from supermarket (km)
9	10
7	8
12	6
11	7
14	4
12	5
6	12
8	11
15	3
13	4
5	13
10	8
4	15
9	11
2	16

   a) Plot a scatter graph to represent this data.
   b) Describe the correlation of the data.
   c) Find the mean number of visits and the mean distance. Use these to draw a line of best fit.
   d) Estimate the number of visits made by a customer living 10 km from the supermarket.
   e) How reliable do you think the estimate in (d) is?

Suggested answers are at the end of this unit.

Section E: Representing data for understanding

Representing data in diagrams is to enhance the understanding of the data. The question to be asked in each situation is: What kind of representation(s) would help you to make sense of the given data? Too frequently questions are set that prescribe to transform data (frequency table data for example) into a given format (bar chart for example). However it is important that pupils learn to decide what might be the most appropriate format to present their data. Different formats should be considered and within a given format (bar chart for example) the effect of re-scaling or changing class width. Transforming from one format to another is also a skill to be developed.

Practice task 1

1. Find some realistic raw data (3 different sets) and represent the data using several of the above-mentioned forms of representation. Discuss and justify which of the representations is the most appropriate to represent your raw data.
2. The table below gives different forms of representing data and the skill required to transform from one format to the other.

To / From	Verbal	Table	Graph/Chart/ Leaf-Stem/ Pictogram	Formula
Verbal	reformulating, expressing in own words	analysing, extracting data	modelling	analytical modelling
Table	reading	reorganising, regrouping	plotting	fitting
Graph/Chart/ Leaf-Stem/ Pictogram	interpretation	reading off	Re-scaling, using different class width	curve fitting
Formula	explaining	computing	sketching	changing subject

   Illustrate each transformation with an appropriate example.

3. Class based activity

   Split your class into two groups A and B. Present to each group the same data sheet.

   Group A is to use the data to make the country look as good as possible when compared with the other countries.

   Group B using the same data is to present the data such that it shows that the country has much to do to catch-up with the other countries.

   Groups are to be encouraged to use any diagram: bar charts, pie-charts, pictograms, histograms, frequency polygons, scatter graphs, stem-leaf plots.

   The data sheet is on the following page.

   Write an evaluative report on the activity.

Data sheet					
	Botswana	Namibia	Zimbabwe	Zambia	Mozambique
Population density (#/km^2)	2	2	29	52	20
Annual pop. growth	2.9%	3.1%	2.8%	2.7%	3.3%
Children per woman	5.2	6.0	5.5	6.5	6.5
Under 5 mortality	85/1000	120/1000	88/1000	200/1000	292/1000
Doctors	1/5000	1/4600	1/7000	1/10 000	1/38000
Safe water available to *x*% of population	54%	72%	66%	60%	24%
Access to health services	80%	52%	82%	50%	22%
Literacy, male	84%	72%	74%	81%	45%
Literacy, female	65%	48%	60%	65%	21%
Secondary enrolment	54%	41%	48%	20%	10%
University (students/inhabitants)	30/10 000	28/10 000	43/10 000	19/10 000	-
TV sets per 1000 inh.	16	21	26	26	3
Radios per 1000 inh.	122	127	84	81	47
Inflation rate	13%	12%	14%	48%	38%
Food import dependency	75%	31%	5%	7%	22%
Economic growth	6.1%	-1.0%	-0.9%	-2.9%	-3.6%

Pop - population
Inh - inhabitants

Representing data for understanding (continued)

An activity for use in the classroom

Use of visual representations of written text is frequently very helpful to understand the text. The 'standard' representations as used in statistics are not the only way data or text can be represented.

Practice task 2

1. Below are descriptions of five situations. Present these to pupils working in groups and ask them to come up with at least two diagrams to clarify the situation described. Each group could be given one or two situations to represent in diagrams, pictures, charts, etc.
2. After the groups have worked on the activity they are to present their work to the class for discussion. Some of the questions to be asked could be: What is the strength of the suggested representation? What is the weakness? How could it be improved? Are there other alternatives?

   The instruction given to pupils in each of the situations is:
   What kind of representation(s) would help you to make sense of each of the following passages (situation 1 – 5)?
3. a) Write an evaluative report on the activity. Questions to consider are: Did pupils meet difficulties? Were pupils well motivated to work on the activity? Were the objectives achieved? Did you meet some specific difficulties in preparing the lesson or during the lesson?

   b) Present the report to your supervisor.

Situation 1: Kidnapped

One of the most influential educationalists in Botswana, Cees, was kidnapped from outside his Gaborone home this morning by masked armed men. Although he was seized in broad daylight on one of the main streets leading to the station fly over, only two eye witnesses have been found by the police and they have been of little help.

Mr. Cees left his home in DO IT street just before 7 am. His driver saw him into the back of the UB INSET van and was driving down the road towards the station flyover when he was forced to pull out to overtake a champagne-coloured Toyota station car that seemed to be vary badly parked.

Immediately opposite was a minibus double-parked and therefore well out from the pavement. The car was forced into what seemed to be an innocent narrow passageway. As the driver was negotiating the gap, a motor cyclist pulled in front of him forcing him to stop. Two masked men jumped out of the back of the Toyota and the motor cyclist pulled a gun.

The men knocked out the driver who was thrown into the minibus that drove off towards Molepolole where he was found dumped along the road a few minutes later. One of the kidnappers jumped into the driver's seat of the UB INSET van and with another holding Mr. Cees at gun point in the back, drove off into the traffic heading for the station fly over. The car was found later at Tlokweng.

The kidnapping was all over in seconds and the witnesses have been able to give only vague descriptions of what they saw.

Situation 2: Ladybirds to the rescue

The Australian cottony cushion scale insect was accidentally introduced into America in 1888 and increased in number until it seemed about to destroy the Californian citrus orchards where it lived. Its natural predator, a ladybird, was artificially introduced in 1889 and this quickly reduced the scale insect population. Later, DDT was used to try to cut down the scale insect population still further. However the net result was to increase their number as, unfortunately, the ladybird was more susceptible to DDT than the scale insect! For the first time in fifty years the scale insect again became a problem.

Situation 3: What foods contain which vitamins?

Vitamin A is found mainly in fats and the fatty parts of some foods, so plenty of milk and butter will help to provide it. Other valuable sources are fish-liver oils and certain vegetables, especially carrots, tomatoes and dark green leafy vegetables. Vitamin D is also found in butter, cheese, milk and eggs but the richest source is fish-liver oils. Sunlight acting on the skin produces vitamin D in the tissue. Vitamin C is found in fresh fruit and vegetables, so plenty of these should be served. In addition orange juice should be given every day. The B vitamins are found in whole meal bread, oats, yeast, liver and dairy foods. One pint of milk a day will supply all the riboflavin (B2) that a child under five needs.

Situation 4: What are you doing in your holiday?

Well it depends. My mum and dad might buy a new car or a colour television or nothing. If we have a car, dad says we will just go for a few day-trips. If we buy nothing we might go to Harare or perhaps Cape Town to stay with Grandma and Grandpa. If we have a colour television, dad says he will only have enough money to travel to Durban and stay for free in the caravan of uncle Sam.

Situation 5: How time works

Emulo gave a description of how to work out what time it is in different parts of the world and completed a table of cities, their longitudes and hence their local times. She wrote:

How time works

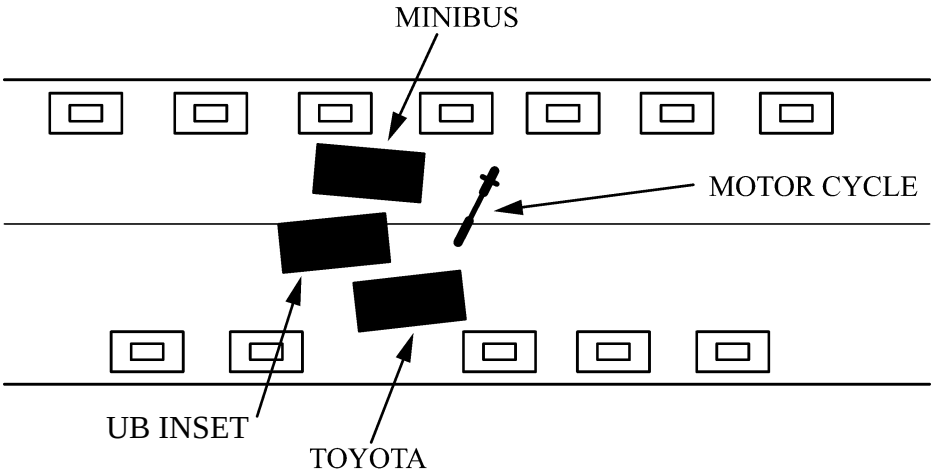
If you want to know what time it is in another place you see what the time is and then whether the country is east or west of Botswana. If it is east every 15° of longitude you pass over you add one hour and if it is west of Botswana you take an hour off Gaborone time as you pass over 15° of longitude. So if it is 10.00 am in Gabs and you want to know what time it is somewhere 60° west of Botswana, fifteen goes into 60 four times so this place is four hours behind Gabs time which it 6.00 am in somewhere 60° west of Botswana.

The verbal description process is rather cumbersome. Is not there an easier way to represent the situation?

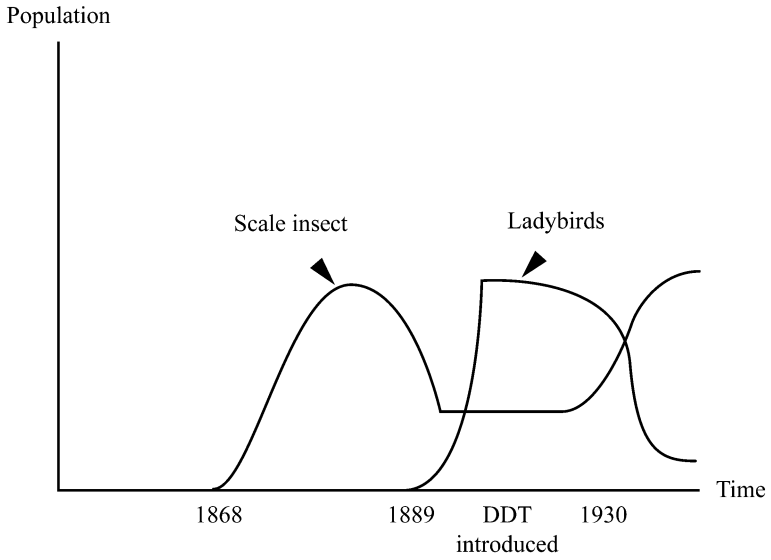
Information for the teacher on the above situations.

Here are a few possible suggestions, but pupils will come up with many others.

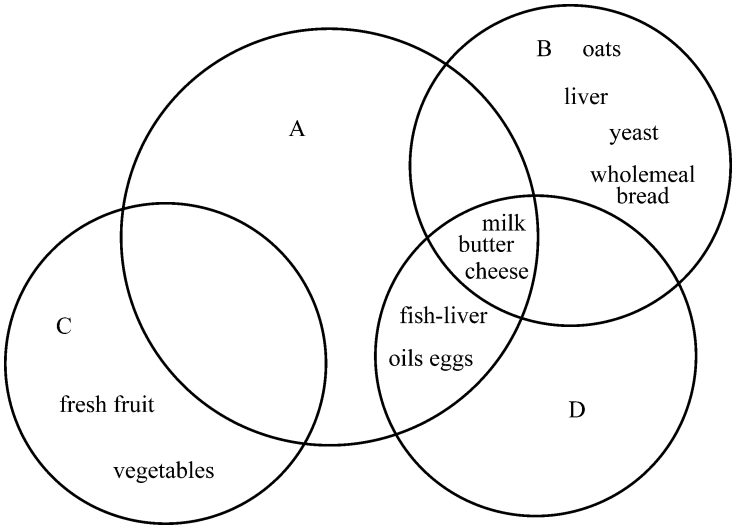
Cees kidnapped



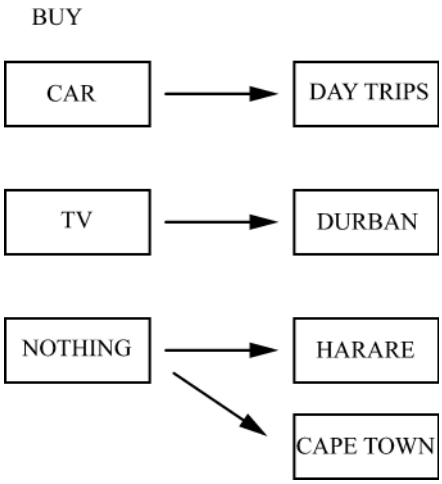
Ladybirds to the rescue



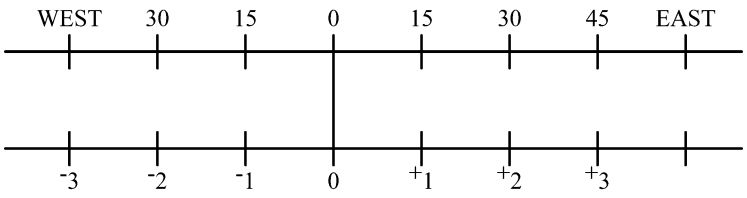
What foods contain which vitamins?



Where are you going for your holidays?



How time works



Section F: Misconceptions of pupils in descriptive statistics

Note: Points 1) through 5) in this list covered common misconceptions in inferential statistics. They were listed in Unit 2 Section C. Below are misconceptions in descriptive statistics.

6) Histograms and bar charts (categorical data: qualitative data) / line charts (discrete numerical data: the possible numerical values are separated from each other by impossible values)

   In a histogram the area is proportional to the frequency, while in a bar chart or line chart the height (or length) is proportional to frequency. The width of the blocks in a histogram need not be equal.

7) Confusing the observations with the frequencies (count) of the observation. Pupils often have a 'surface' concepts of mean, median and mode. Superficially the median is thought of as the 'middle value' and the mode as 'the most frequent value'. But the middle value of what? The most frequent value of what? In a frequency distribution pupils might erroneously take the 'middle' or the 'largest' *frequency* instead of the appropriate *observation*.

8) Difficulties in the interpretation of the meaning of something to *represent* another thing. There are two concepts of 'represent' to distinguish:

   a) the accurate representation. A histogram, for example, represents the sample accurately.

   b) the probabilistic representation. A sample represents a population probabilistically. The confidence in the representation will depend on randomness (being unbiased) of the sampling technique and the size of the sample.

   Pupils tend not to distinguish between these two types of representation. This results in the idea sample = population and if it differs they think the experimenter must have made a mistake.

Self mark exercise 7

In each of the following question 1 & 2:

a) Identify the error / misconception of the pupil

b) Develop activities using realistic data that will

- confront pupils with (common) misconceptions
- resolve the conflict
- consolidate the correct concept

1. An examination was taken by 1000 students and the overall average was 80%. A random sample of 10 examination scripts was taken from the 1000. The first script picked randomly had a score of 60%.

   What do you expect the average of the sample to be?

   Pupil answer: 80%

2. The following data are given:

   Height of 133 plants in cm

Height	160	161	162	163	164	165	166
Frequency	10	15	29	28	24	21	6

   Pupils are to find the mode and the median.

   Pupil answers: mode 29

   median 28

3. Which graphical representation(s) would you use in each of the following situations? Choose from: pie chart, bar chart, histogram, stem-leaf plot, scatter graph. Justify your answer.

   a. Testing the reaction speed of people after drinking a number of cans of beer.

   b. Representing the amount of money Government spends on health, education, armed forces, etc.

   c. Comparing the prices for the same brand of shoes in different shops.

   d. Comparing the salaries of workers in a factory.

   e. Comparing the height of boys and girls in your class.

Suggested answers are at the end of this unit.

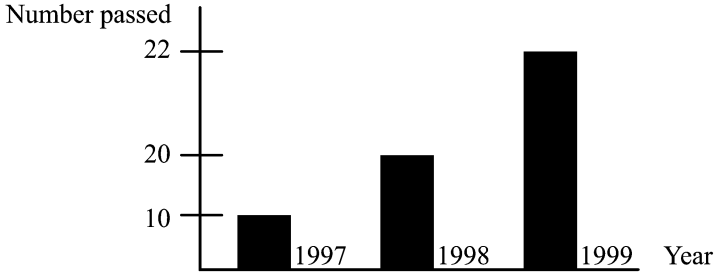
Section G: Making nonsense of statistics

Pupils are to learn to look critically at data presentations. Data is at times presented so as to carry misleading information to a careless observer. The irreverent name for this is "How to lie with statistics."

Common misleading techniques are:

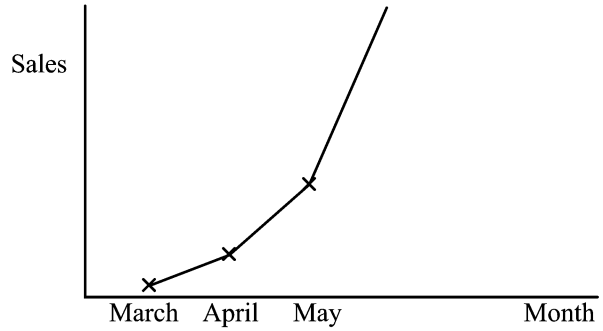
- vertical axis labelled unevenly

More learner drivers are passing in our driving school year after year.



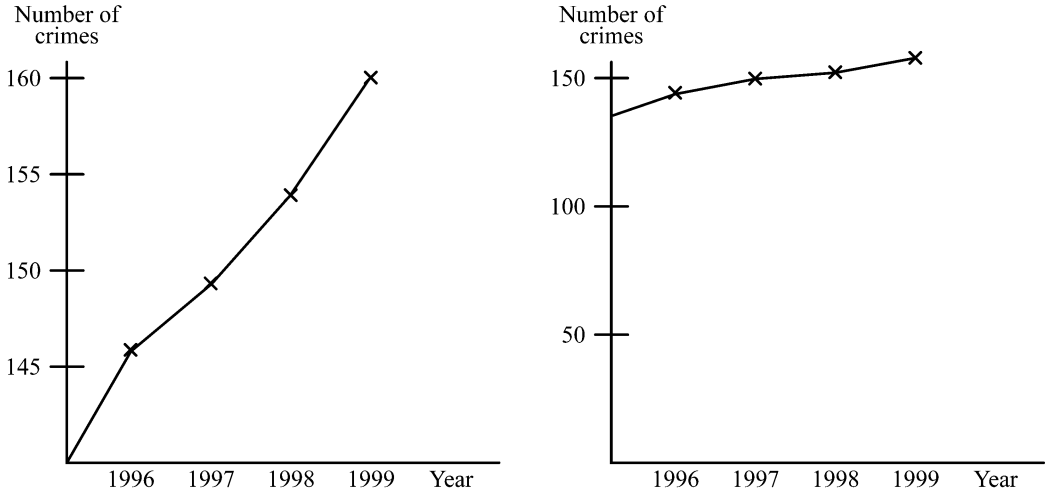
- no scale provided

The sales of our newspaper goes up. Join our readers!



- vertical scale not starting at 0, without clearly indicating this ("squeezed" line)

Crime rate in town has increased very rapidly over the past 5 years.

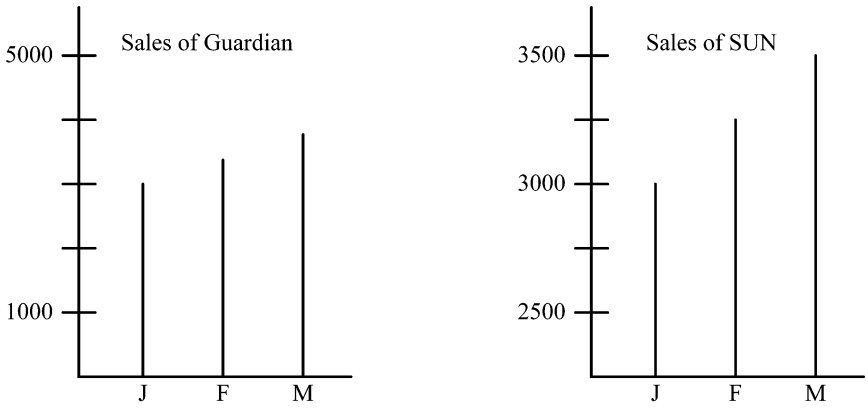


Because the vertical axis does not start at zero the number of crimes appear to be increasing quickly.

In the other graph the data has been redrawn using a different scale and starting from zero on the vertical axis. This illustrates that the number of crimes have very slightly increased.

- different scales used in two displays to 'show' that a certain company or product is doing better

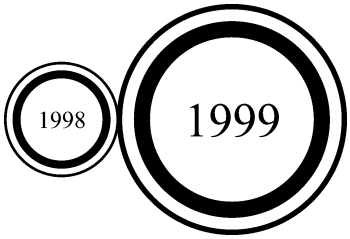
Sales of two newspapers "Guardian" and "SUN" over the first 3 months of 1996 are illustrated below. The SUN claims: Our readership is increasing faster than the readership of the Guardian.



As the scales in the two graphs to be compared differ, a false impression is created: both increased by the same amount from 3000 to 3500.

- using 2D- area or 3D- solid diagrams. To display 'doubling', for example, all dimensions are doubled and hence in the area case the area is 4x the original and the solid is 8x the original.

Our sales of tyres have doubled from 1998 to 1999.



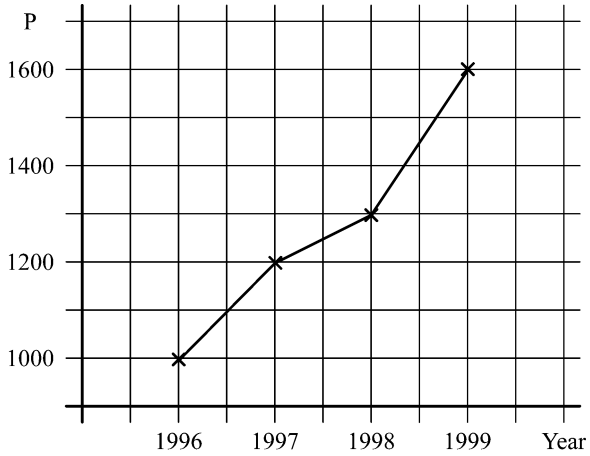
10 000 tyres 20 0000 tyres

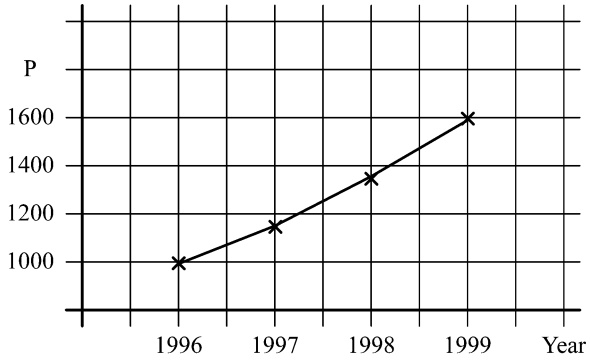
In 'picture graphs' it is the area (or volume that represents the quantities). The radius for the 1999 tyres sales is double that of the 1998 one. However that implies that the area of the circle representing the 1999 sales is 4-times the area of the circle representing the 1998 sales!

Self mark exercise 8

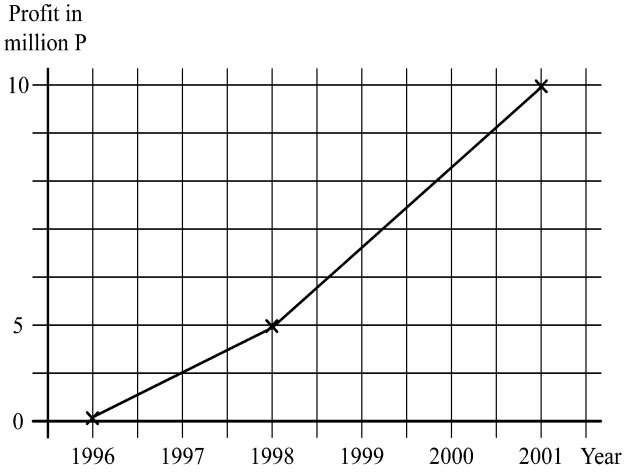
1. Look at the following graphs and explain why they are misleading. Give a better presentation of the data.

   a) "Invest with FAST GROWTH, your investment will grow faster than with any other company"



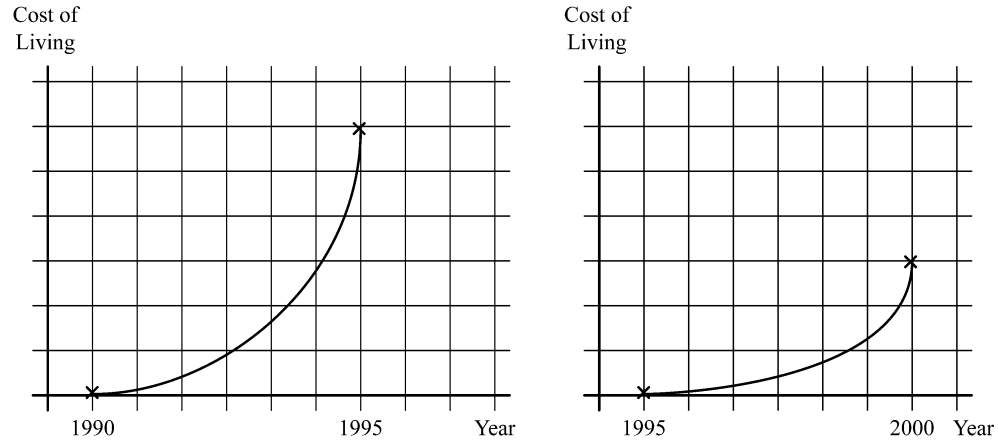


   b) "Our profits have increased faster over the past three years"



Continued on next page

c) "Cost of living increase has been reduced"



2. Chocco Sweets Company wants to impress their shareholders that their sales have increased from 20 000 kg in 1997 to 40 000 kg in 1998 to 60 000 kg in 1999.

   a) Represent the data in a bar chart.

   b) Design a picture graph to create an impression of much greater increase than actually is the case.

3. The following line graph was found in the newspaper with the heading "Unemployment figures have gone down rapidly over the past four years." Do you agree with the newspaper? Justify your answer.

4. The following table gives the number of cars sold in a garage during the first three months since a new sales manager took over.

Month	August	Sept	Oct
Number sold	80	74	64

The sales figures are clearly going down.

   a) You are the sales manager and are to present the figures to the board of directors in a bar chart. As much as possible you want to disguise the dropping sales figures. Draw the bar chart you would present to the board.

   b) You are the supervisor of the sales manager and want to impress on her that since she took over sales figures are dropping. Draw a bar chart you would use to get your message powerfully across.

Continued on next page

5. The following table displays the percent of households in some countries having television sets.

Country	Botswana	Namibia	Zimbabwe	Zambia
% of households with TV	20	26	32	18

a) You live in Zimbabwe and you want to make your country look wealthier than all the other countries. Draw a bar chart with a scale that will help you to make your point.

b) You are living in Zambia and want to make your country look as wealthy as possible as compared to the other countries. Draw a bar chart with a scale that will help you to make your point.

Suggested answers are at the end of this unit.

Section H: Interpreting data

In magazines and newspapers you frequently come across data representations in a variety of forms. Data representations need to be looked at critically. Reading and interpreting graphical representations of data is not a trivial task. Many diagrams, charts or graphs in newspapers and magazines have been designed to magnify differences or to emphasise minor points. You are to ask yourself questions such as: How was the data collected? Does the representation give a fair picture of the data? Are the data reliable? What purpose do the presenters of the data have?

Self mark exercise 9

For each of the following seven data representations answer the following questions:

1. What is the representation about?
2. What type of data is represented (qualitative, quantitative; discrete, continuous; grouped, ungrouped)?
3. What type of graph is it?
4. How do you think the data was collected?
5. What conclusions can you draw from the diagram?
6. What other representation would you consider appropriate for the data? Justify.

Suggested answers are at the end of this unit.

Fig 1:

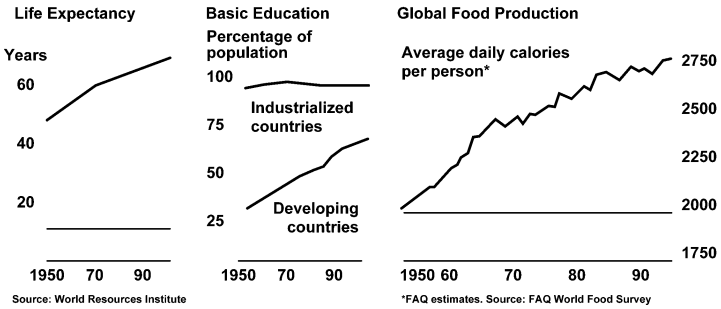


Fig 2:

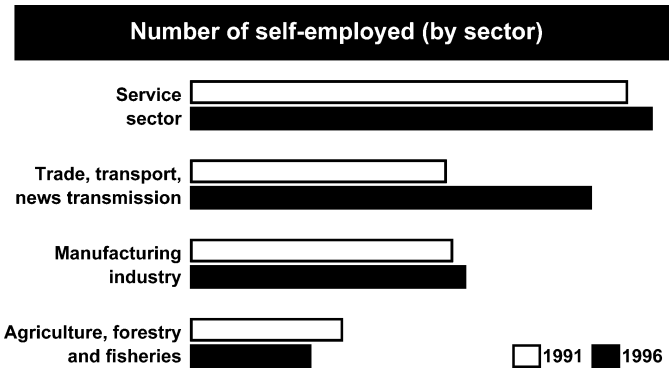


Fig 3:

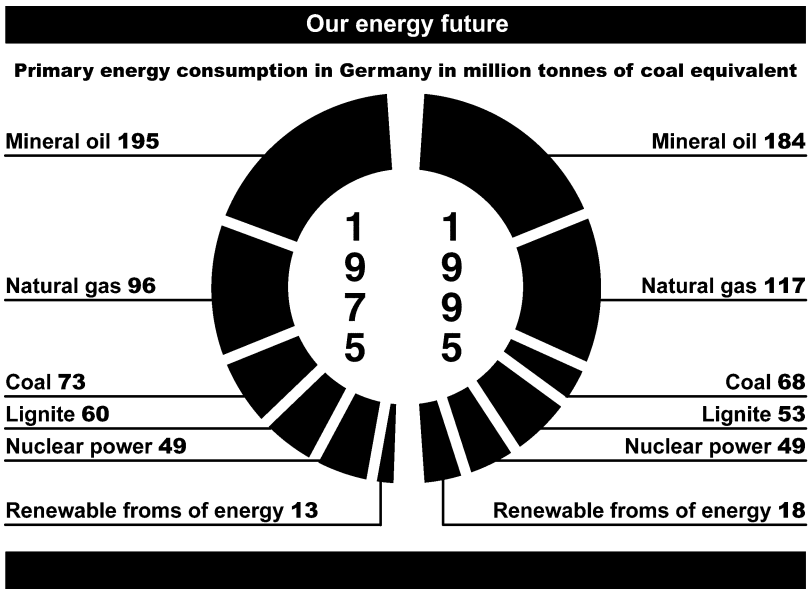
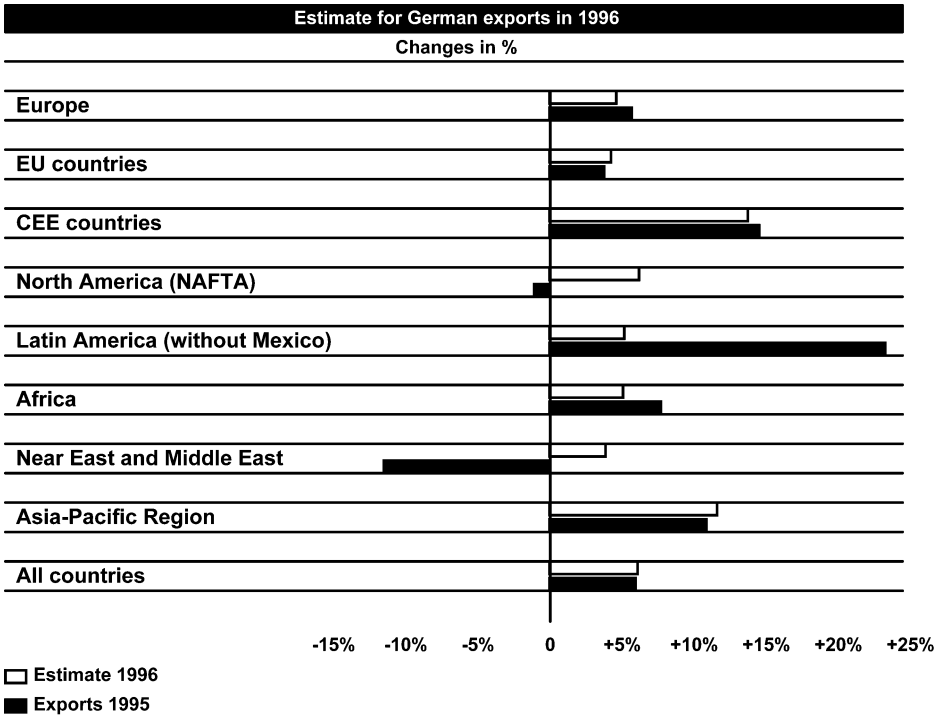


Fig 4:



Source: DIHT

Fig 5:

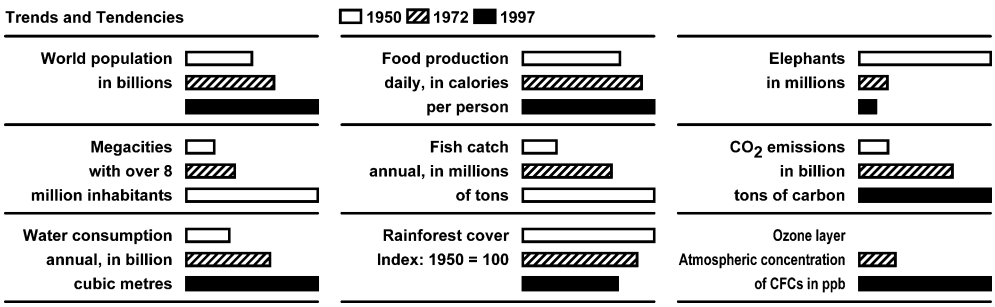


Fig 6:

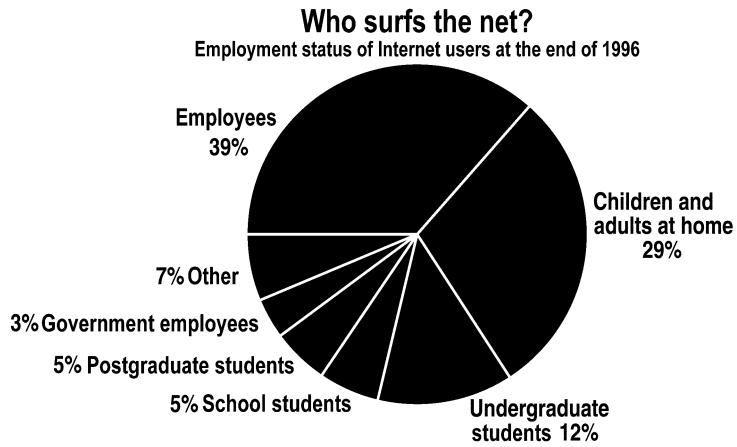
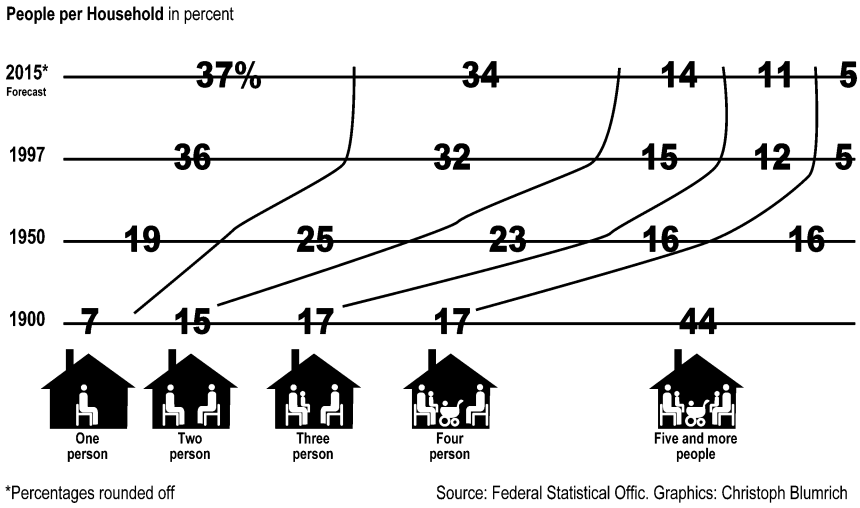


Fig 7:



Practice task 3

1. Choose one (or more) of the data representations in Section D, or an activity related to Section G or H.
2. Write a lesson plan with clearly stated objectives. Prepare worksheets for the pupils to work in groups.
3. a. Write an evaluative report on the lesson. Questions to consider are: Did pupils meet difficulties? Were pupils well motivated to work on the activity? Were the objectives achieved? Did you meet some specific difficulties in preparing the lesson or during the lesson?

   b. Present the lesson plan and report to your supervisor.

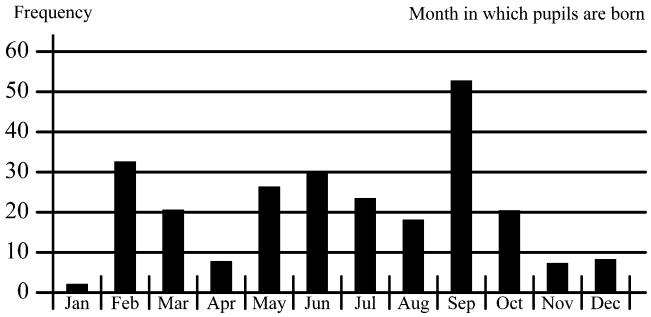
Summary

Data handling and interpretation is one of the few topics in Secondary Mathematics that impacts the daily lives of most thinking people. Graphs, like those in Section H and like the biased ones in Section G, abound in the media. Students should also produce representations of the data they gathered in their own projects—and, if possible, recommend a decision that could be taken on the basis of their work. Groups will learn a great deal from presenting their work to the class as a whole.

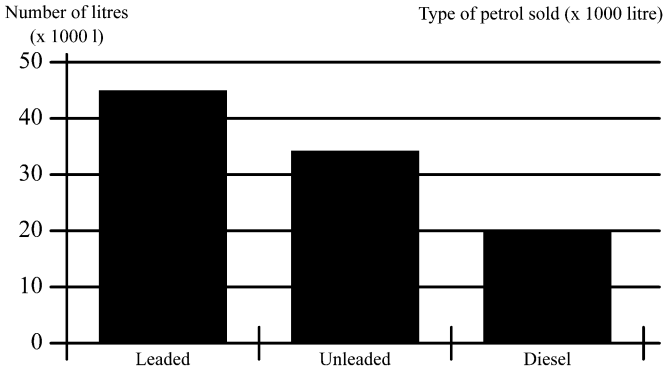
Unit 3: Answers to self mark exercises

Self mark exercise 1

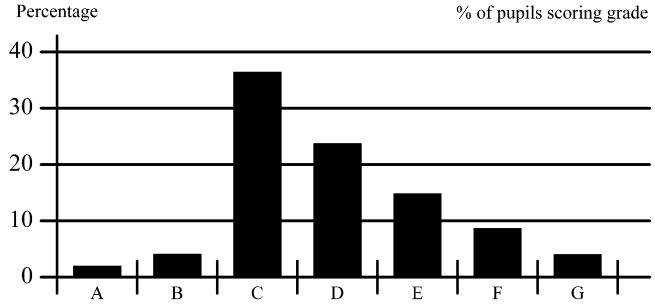
1.



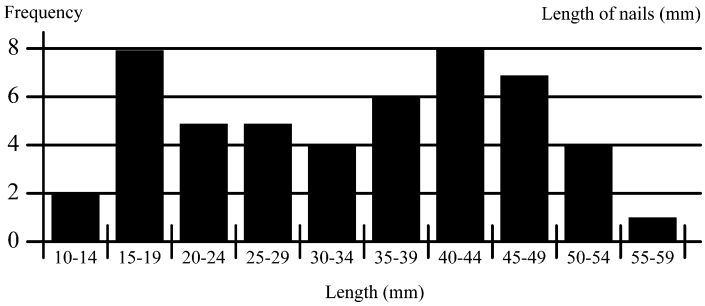
2.



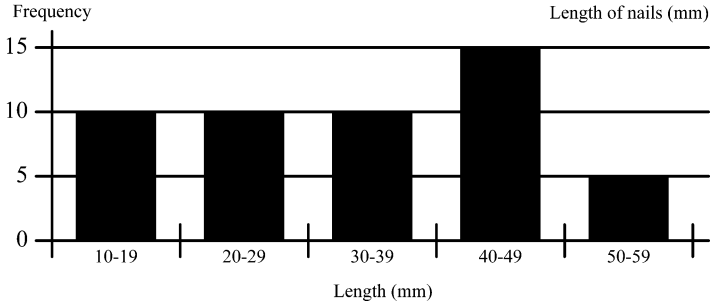
3.



4a Class interval	Frequency	4b Class interval	Frequency
$10 \le l \le 14$	2	$10 \le l \le 19$	10
$15 \le l \le 19$	8		
$20 \le l \le 24$	5	$20 \le l \le 29$	10
$25 \le l \le 29$	5		
$30 \le l \le 34$	4	$30 \le l \le 39$	10
$35 \le l \le 39$	6		
$40 \le l \le 44$	8	$40 \le l \le 49$	15
$45 \le l \le 49$	7		
$50 \le l \le 54$	4	$50 \le l \le 59$	5
$54 \le l \le 59$	1		



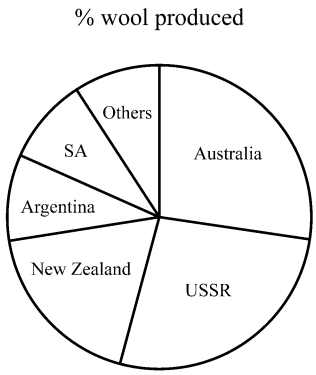
4b.

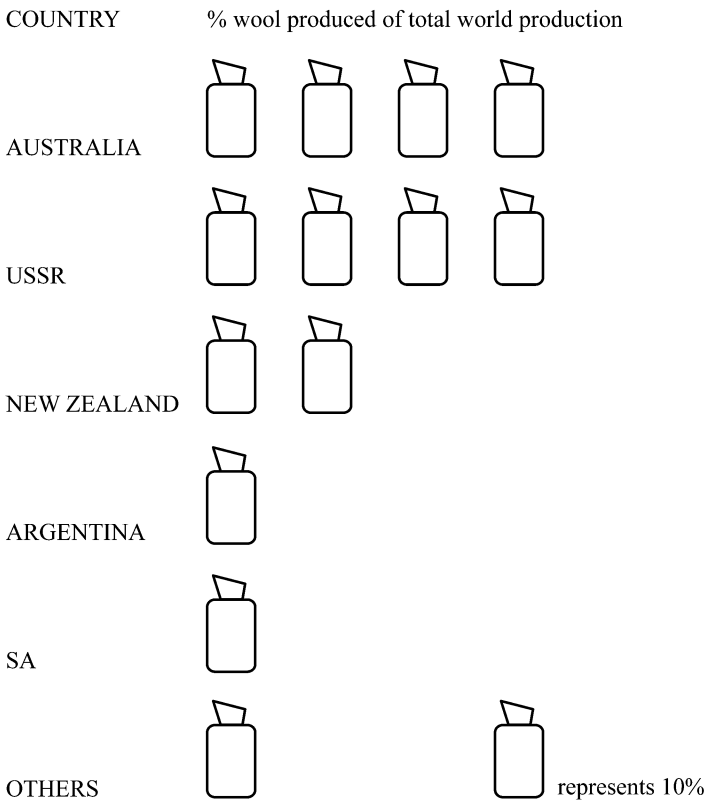


Changing the class width to 10 (as in 4b) gives a less informative representation of the data, as some information is lost. On the other hand, as the number of bars is reduced, the diagram becomes easier to read.

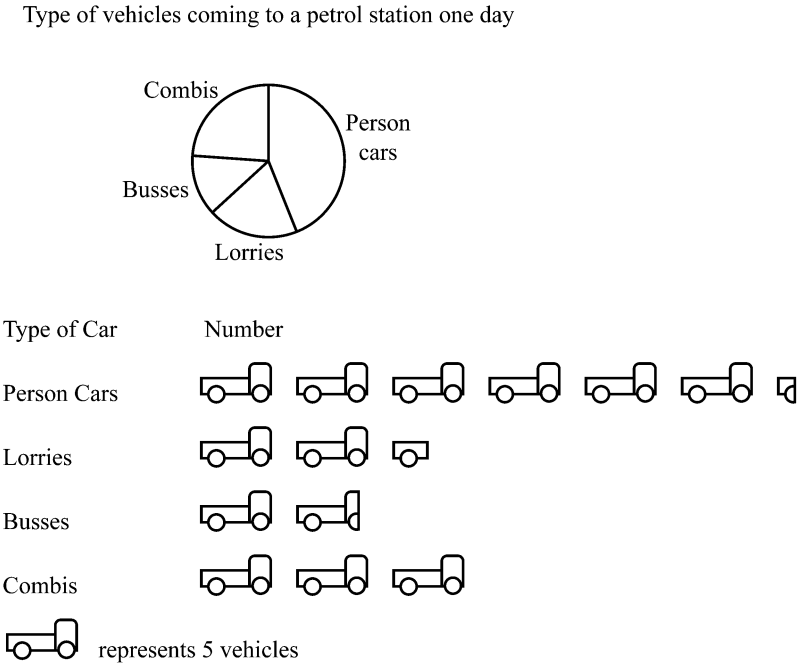
Self mark exercise 2

1.

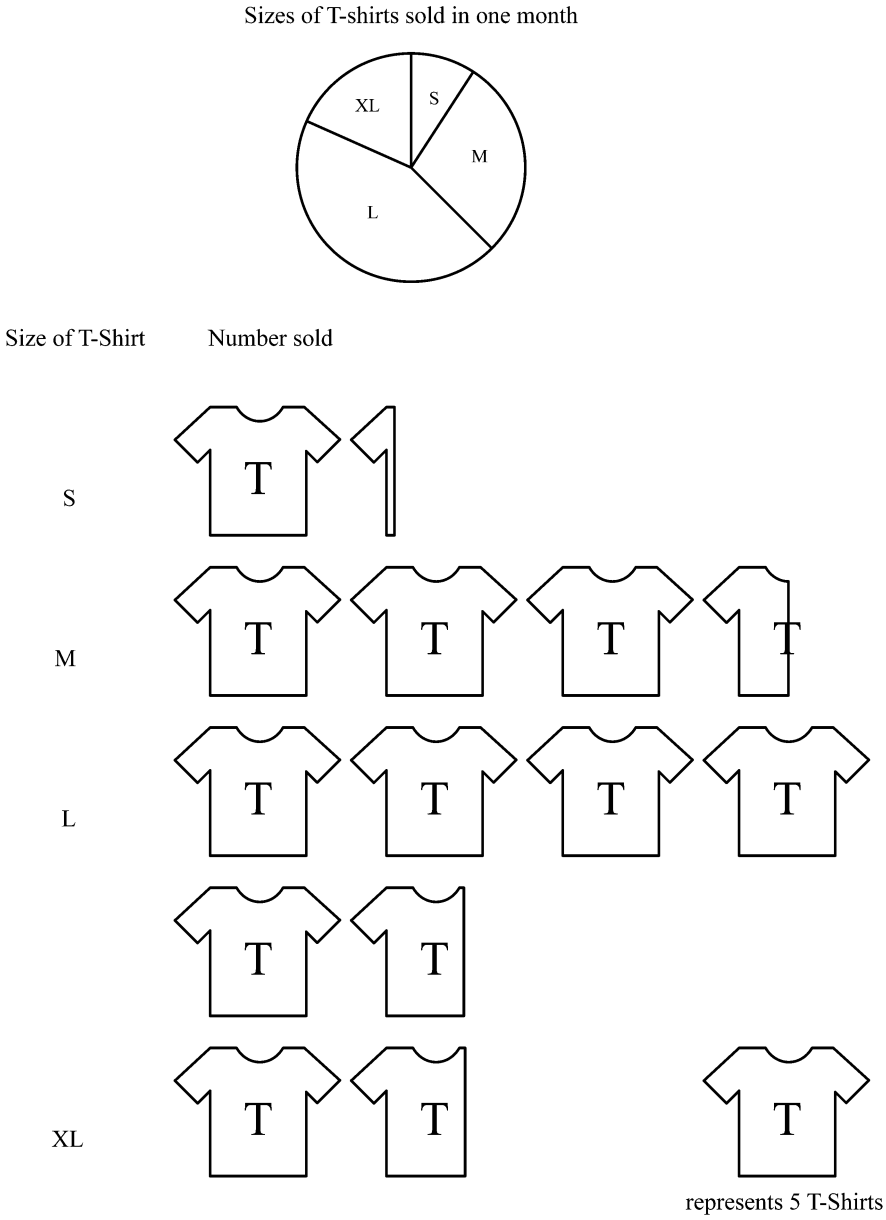



COUNTRY
% wool produced of total world production
AUSTRALIA
USSR
NEW ZEALAND
ARGENTINA
SA
OTHERS
represents 10%

2.


Type of vehicles coming to a petrol station one day
Combis
Person cars
Busses
Lorries
Type of Car
Number
Person Cars
Lorries
Busses
Combis
represents 5 vehicles

3.



4. **Pie Chart**

Advantages

- Allows easy comparison of parts with whole

Disadvantages

- At times tedious to calculate the sector angles
- The actual frequencies are not shown and need to be obtained by interpreting the chart

Pictogram

Advantages

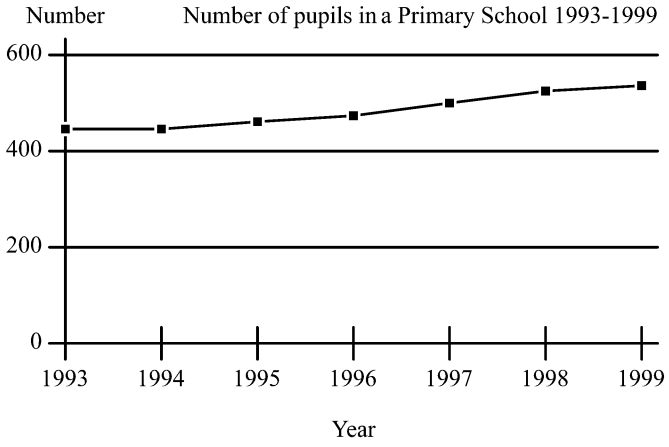
- Can be made visually attractive
- Pictures make 'topic' clear

Disadvantages

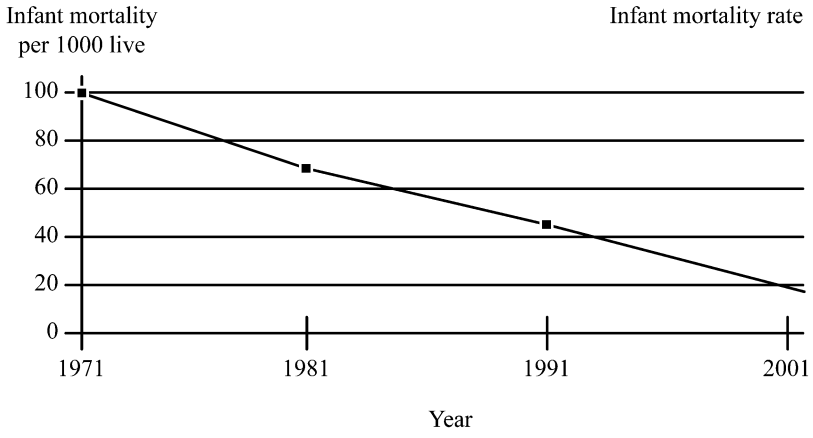
- Hard to draw
- 'Fractional' pictures difficult to interpret

Self mark exercise 3

1.

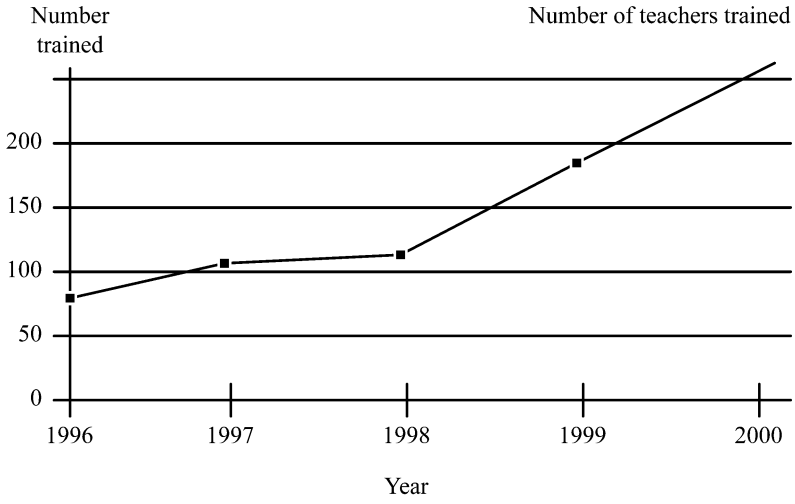


2a)



2b) 24

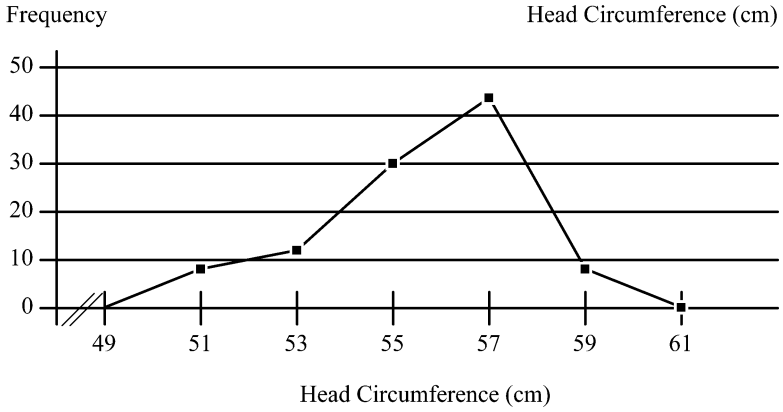
3a)



3b) 260

Self mark exercise 4

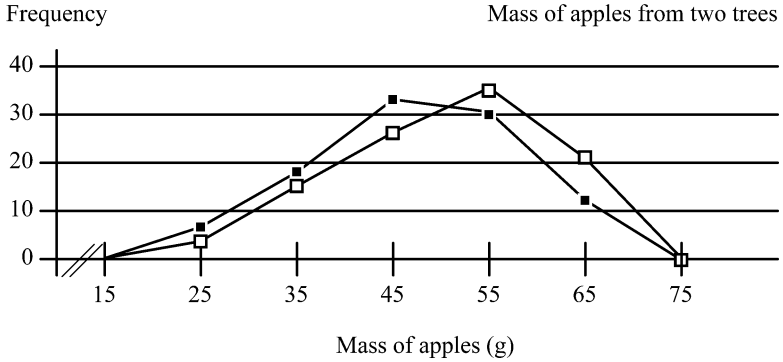
1. Plot the points ((49, 0), 51, 8), (53, 12), (55, 30), (57, 44), (59, 6), (61, 0)



1b) Produce 45% size 56 - 58, 30% of size 54 - 56, 15% of size 52 - 54, and 55 each of size 50 - 52 & 58 - 60.

2a) Plot (15, 0), (25, 6), (35, 18), (45, 34), (55, 30), (65, 12), (75, 0)

2b) Plot (15, 0), (25, 3), (35, 14), (45, 26), (55, 36), (65, 21), (75, 0)



2c) Both trees produced 100 apples. The apples from the second tree on average have a greater mass. The 'top' of the graph of the masses of the apples from the second tree is more to the right.

Self mark exercise 5

1.

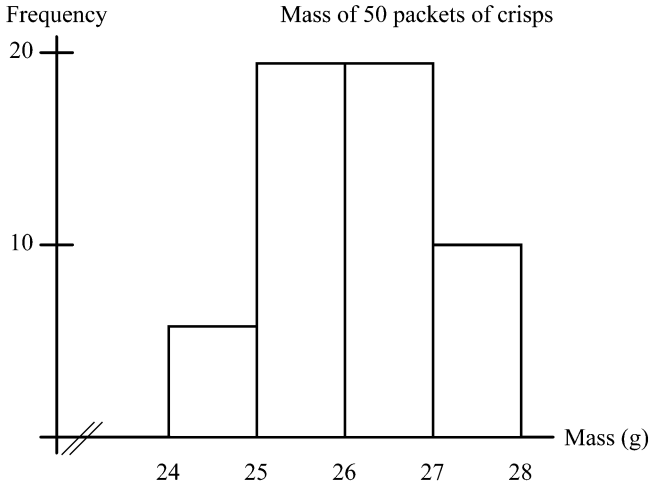
4	023
4	6689
5	0112234
5	55789
6	044
6	67
7	2

n=25 4 | 6 represents 46 g.

2a)

Mass	frequency
$24 \leq m < 25$	4
$25 \leq m < 26$	18
$26 \leq m < 27$	18
$27 \leq m < 28$	10

2b)



Self mark exercise 6

1a) Weak / strong positive correlation (as it will depend on the achievement level of the pupil. A high achiever will not suffer much from being absent, but an average or low achiever will be affected by missing school days).

b) No correlation.

c) Weak / strong correlation. People with higher levels of education generally will get better paid positions. However in business, people without much formal education can become business tycoons.

d) Difficult to predict. One might argue for different types of correlation. E.g., positive correlation - people with higher incomes have more money and can spend more time in bars.

Negative correlation - people with low incomes spend the little they have on drinks in bars, while the people with higher incomes prefer to drink in their houses and not in public places.

No correlation - a group of people with low incomes as well as a group of people with high incomes spend time in bars.

e) No linear correlation. Children take a long time to read a page and with an increase in age they might start to read faster, but at an older age the speed might go down again.

f) Strong positive correlation. Each room will need (at least one) doors to enter. More rooms leads to more doors.

g) No correlation expected. Technique and strength of muscles might be more relevant than arm length.

h) No correlation.

i) Strong positive correlation. The more mass attached the more the spring will extend.

j) No correlation.

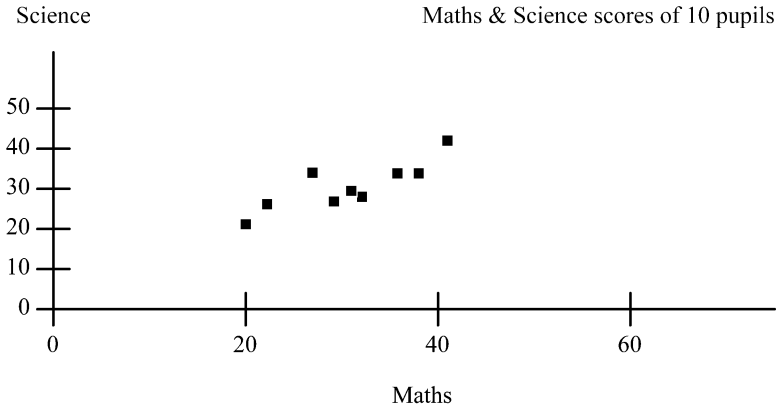
k) Negative correlation. The older the car the less its value will be.

l) No correlation.

m) In the form the question is set: no correlation as a car might have covered many thousands of kilometres but just have new tyres.

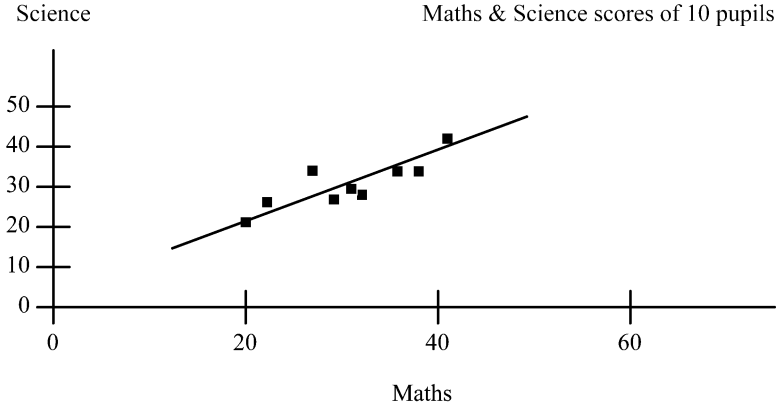
If the question is intending to say: distance travelled with those tyres, then there is a strong negative correlation. Depth of tread reduces if number of kilometres travelled with those tyres increases.

2a)



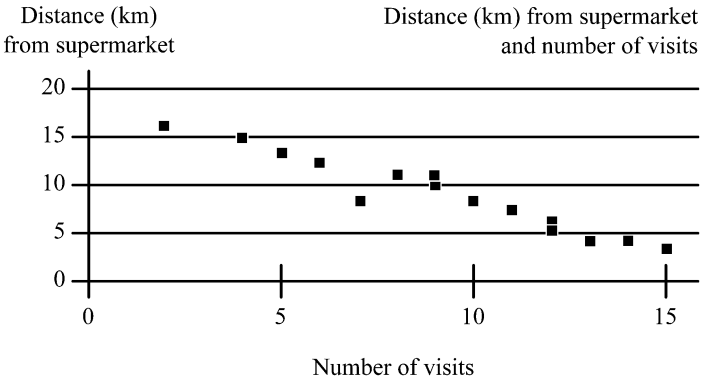
b) A positive correlation exists between the scores.

c) Mean Maths 30. 8, mean science 30.2.



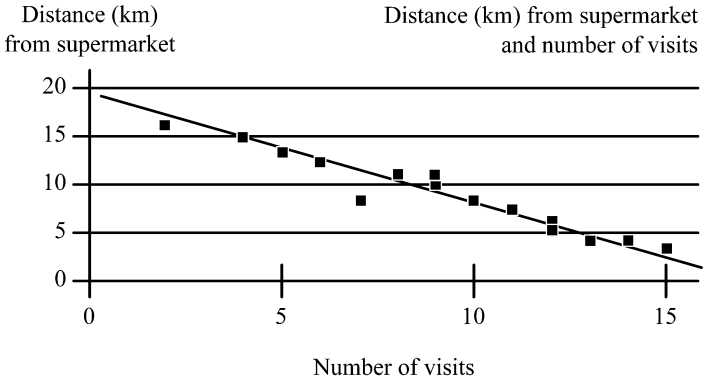
d) 35 e. 40

3a)



b) The data are negatively correlated, i.e., the number of visits to the super market decreases when the distance the person is away from the supermarket increases. Or the number of visits increases if the distance the person is away from the supermarket decreases.

c) Mean number of visits 9.13, mean distance 8.87



d) 8

e) As the correlation is rather strong the estimate is reasonable reliable.

Self mark exercise 7

1a) The pupil thinks that the sample must have the same mean as the population.

b) (i) Ask pupil to explain how (s)he worked the question.

(ii) Create mental conflict. Suppose only 2 scripts were taken at random. The first has a score of 55%. What is the expected mean of the sample of two?
Here the pupil will discover it cannot be 80% because then the second script would need a mark of 105%, more than is possible.

(iii) Guided questioning should bring out:
One script has score 60%, the other nine scripts, we don't know, but looking at the mean we have as our only option to assume that those nine will have an mean of 80%. Hence the expected mean of the sample will be $(60\% + 9 \times 80\%) \div 10 = 78\%$.

(iv) Consolidate the 'new' knowledge of the pupil by setting similar questions.

2a) Pupil confuses observations with frequencies. Mode and median are observation related, not frequency.

b) (i) Ask pupil to explain how (s)he worked the question.

(ii) Create mental conflict by looking at data set
160, 160, 160, 161, 161, 162, 163

Ask pupil for mode and median.

Next ask pupil to place the data in a frequency table thus:

Height	160	161	162	163
Frequency	3	2	1	1

Ask again for mode and median.

Ask pupil what is to be looked at for mode and median, the height row or the frequency row.

(iii) Guide the pupil through the set problem, which will be not so hard after step (ii). The pupil is now aware of the error and will work with the correct row.

(iv) Set similar questions—given a distribution table of discrete data, to find median and mode—for consolidation purposes.

3a) Line graph to see whether or not there is a trend.

b) Bar graph if the number of categories turn out to be large (8 or more). Otherwise use pie chart to compare the amounts spend on each category.

c) Bar chart or bar line chart, the height of the bars giving the prices.

d) Pictogram / bar chart if comparing discrete salaries (or range of salaries) of different groups of workers. If the data is to display how many workers earn a salary in a particular range a histogram might be appropriate (unequal classes).

e) Frequency polygons on the same axes will allow easy comparison.

Self mark exercise 8

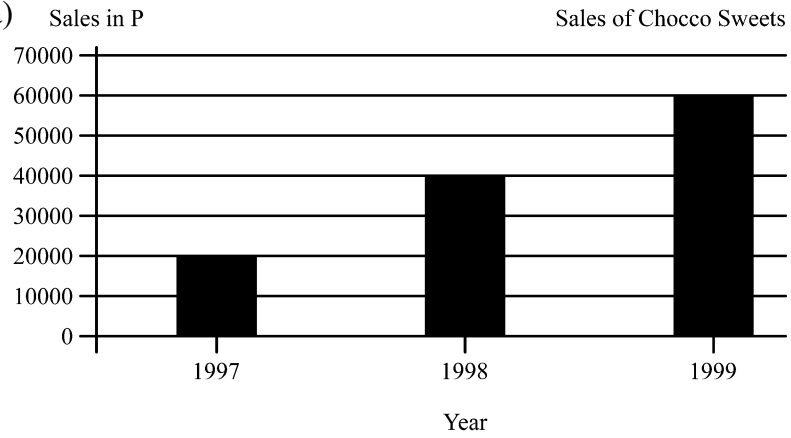
1a) The two line graphs use different scales on the vertical axis (that they do not start at zero is less serious although the 'squeeze' should have been indicated).

Draw the two line graphs on the same axes system (or if on different then use the same scale on the vertical axis). The FAST GROWTH company has the same trend as the others.

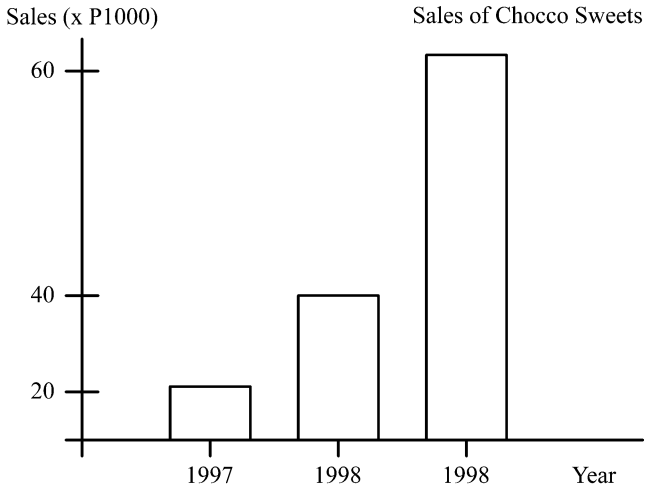
b) The scale along the vertical axis is uneven. First two squares represent P5 million, next 5 squares are used to represent the same amount. Draw the graph using consistent scale along vertical axis and note that the increase is the same for both 3 year periods.

c) No scale is available, so no conclusion can be drawn from the graphs. They should be drawn on the same axis—but due to missing scale you cannot do it.

2a)

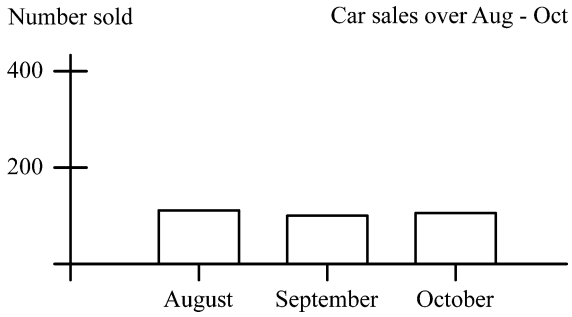


2b)

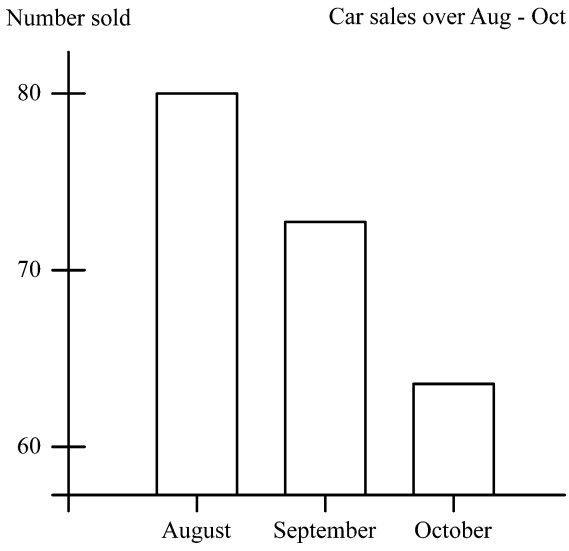


3) The vertical axis is not labelled. Hence no valid conclusions can be made.

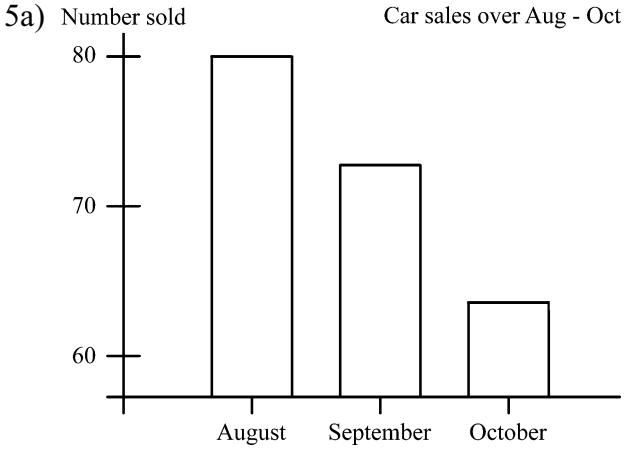
4a)



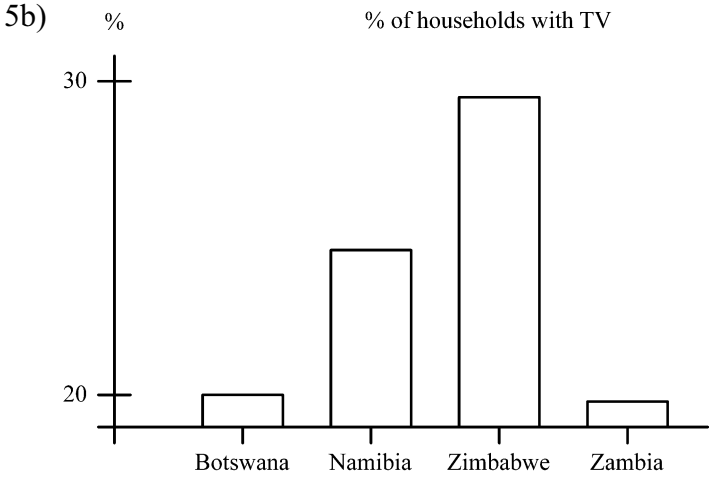
4b)



5a)



5b)



Self mark exercise 9

	Fig. 1	Fig. 2	Fig. 3	Fig. 4	Fig. 5	Fig. 6	Fig. 7
2. What type of data? (Answers may vary)	Quant, contin., ungrouped	Qual, discrete, grouped	Quant, contin., grouped	Quant, contin., grouped	Quant, mixed, grouped	Qual, contin., grouped	Qual?, contin., grouped
3. What type of graph?	Line	Bar	Bar*	Bar	Bar	Pie	Pie/Bar**

* Not Pie, because (a) two sets of data are displayed and a pie chart can show only one, and (b) each bar is separate from its neighbour, as in a bar chart but not in a pie chart. Another description of this is a "stacked bar" chart consisting of two bars, except that the bars are not vertical. Incidentally, the graph shows bias! Compare the two bars for "renewable forms" and for "lignite": they are clearly not sized correctly for the numbers they represent.

** This clever graph could be considered a line graph turned on its side, or as four pie charts which each add up to 100%. Although not mentioned in this course, this graph is also a "stacked bar" graph turned on its side, since stacked bar graphs often have tie-lines that connect a level in one bar to the same level in the next bar. It is *not* a pictogram, since the pictures of houses do not represent counts or frequencies—only categories.

Unit 4: Measures of central tendency

Introduction to Unit 4

The use of the mind should precede the use of the mean. This is to say that one has to look further than the mere calculation of averages.

There are four aspects to consider when it comes to giving meaning to the represented data:

1. Look behind the data
   Data sets are related to a context and gathered and presented by someone who might have a particular agenda. It is therefore important to look behind the data. Questions that are to be considered are: is there bias, attempts to disguise some data, attempts to mislead with data, attempts to present the data from only one point of view? Misuse and abuse of statistics are to be an important aspect of pupils' data handling experiences.
2. Look at the data
   This covers computational and representation aspects: which statistics are meaningful to compute and what is the best way to represent the data in chart or diagram.
3. Look between the data
   This is the comparison aspect of the analysis: looking for differences and similarities.
4. Look beyond the data
   This is to cover the inference part of the analysis: what conclusion can be (safely) drawn from the results.

Purpose of Unit 4

In this unit you are going to look at different averages: mean, mode and median and how to calculate them when data is grouped or ungrouped, as well as which average is most appropriate to use in a given situation. The unit begins with reasonably standard material on central tendency. However, it includes classroom assignments with a twist: students determine the central tendency of their own understanding of central tendency! Thus, an underlying purpose of this unit is to help you teach statistics by means of statistics.

Objectives

After completing this module, you should be able to:

- find the measures of central tendency (mean, median and mode) of ungrouped data (frequency tables)
- find an estimate of mean, median and mode of grouped data given in a frequency table and /or histogram or cumulative frequency curve
- justify which measure of central tendency is most appropriate to use in a given context

- investigate the effect of increasing/decreasing all data by a constant, or increasing/decreasing all data by a given factor on the mean, median and mode
- investigate the effect of data value of 0 on mean, median and mode
- set activities for pupils to enhance their understanding of measures of central tendency
- state the common misconceptions of pupils related to measures of central tendency

Time

To study this unit will take about 10 hours.

Unit 4: Measures of central tendency

You must be familiar with the three averages: mean, median and mode. Write down how you have been teaching these concepts to your pupils. Illustrate with the examples you generally use.

When reading through the next section refer to what you wrote down.

Section A: Averages: mean, mode, median

Many questions related to mean, mode and median merely test a pupil's ability to recall a formula, to substitute the values into the formula and to compute an arithmetically correct answer (operational or instrumental understanding) while pupils lack a relational or functional understanding of these measures of central tendency. Many questions in data handling deal only with **arithmetical aspects** and not real statistical questions. How to compute measures for central tendency is of limited value when the pupil does not know how to interpret the values. Interpreting the meaning of a stated measure of central tendency should be part of the activities presented to pupils.

A question such as: Find the mean of 6.3, 5.4, 4.9, 4.3 and 0.8 does nothing to test a pupil's understanding of the functional characteristics of the mean as a representative statistic, or model of the given data, and so is a bad question. It can also be criticised because it is based on a small sample of meaningless figures. Data is to be interpreted in a context.

For example:
A train is to leave the station at 8.30 each morning. The departing time on Monday was 35 minutes late due to a fire in the restaurant car. On Tuesday the train was delayed 5 minutes, on Wednesday 3 minutes, Thursday 3 minutes and Friday 4 minutes.

What was the average number of minutes that the train was late leaving? (10 minutes)

Is this 'average' a good figure to use to represent the week's data? (No, because the delay on Monday is an exceptional instance, and hence it should NOT be included in the average; the 10 minutes are not a reflection of what passengers might expect.)

The problem of when to consider an outlier as an outlier and when as an extreme (but possible) value is seldom considered by teachers. In the first example based on meaningless data, pupils with real understanding of the statistical concept 'average' face serious problems: is 0.8 an outlier or not? There is no context to provide clues.

Data handling has to relate to a context to make it meaningful. Providing pupils with meaningful and realistic problems is essential for developing understanding of measures of central tendency or more generally for the understanding of data handling.

Section B: The concept of the mean

The following aspects need attention.

1) The mean is located between the extreme values.

   For example: The number of pupils present in class during a week are as follows:

Monday	26	Tuesday	18	Wednesday	24
Thursday	29	Friday	28		

   The mean is $(26 + 18 + 24 + 29 + 28) \div 5 = 25$

   This is between the extreme values of 18 and 29.

2) The sum of the deviations from the mean is zero.

   Using the same data as in the example above, the deviations from the mean 25 are

   $+1, {}^{-}7, {}^{-}1, +4, +3$. The sum being zero.

   This property is important for estimating the mean of a set of numerical data. It can be taught effectively by reversing the traditional order: instead of giving data and asking to compute the mean, the teacher can give the mean (say 25) and ask to find a data set (of say 10 data) with that mean. Pupils quickly discover the "balance strategy", i.e., if I include 23 (two below the mean of 25), it is to be balanced by say 27 (two above the mean).

3) The average is influenced by values that deviate from the average.

   Using the above data and assuming that on Saturday 28 pupils attended, the mean attendance over the six days will be different from 25. However if on Saturday 25 pupils attended (the mean over the first five days) the mean over the six days will remain 25.

4) The average does not necessarily equal one of the values that was summed.

   The mean over the six days is $(26 + 18 + 24 + 29 + 28 + 28) \div 6 = 25.5$. However 25.5 pupils can never attend. The mean is not a value equal to any of the values averaged.

5) The average can be a fractional value with no counterpart in reality.

   The example in 4 shows that the mean (25.5) can be a decimal with no real object it can refer to in reality: 25.5 pupils do not exist.

6) The average value is representative of the values that were averaged.

   This is an important property used when interpreting data. The mean represents all the data in the data set.

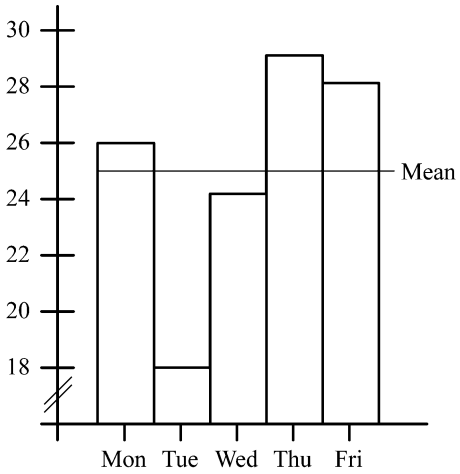
7) In computing an average, a value of zero, if it appears, is to be taken into account.

   Some pupils have the misunderstanding that 0 is 'nothing' and hence need not to be included in calculation of the mean. However 0 is a legitimate numerical value. For example somebody might have 0 brothers and sisters, the temperature might be 0^{o} C.

8) Relating the mean to the representations of data.

In a bar chart the mean frequency will be the height of a rectangle enclosing an area equal to the area of the bars.

For example: The bar chart represents the number of pupils present in class from Monday to Friday.



The mean number of pupils in class during the five days was 25. The area from the "mean" line down to the lower axis equals the area of the five bars.

Section B1: The average or mean has three meanings

Take for example the statement that "On average a box of matches contains 35 matches."

1) measure of location

   Words such as 'around', 'near', 'about', 'close to' were used by pupils to explain the above statement. These responses (the great majority) indicate that pupils understand mean as a measure of location.

2) representative number

   This notion is usually lacking among 13 -14 year olds. They do not understand the mean as a measure resulting from a stochastic process, i.e., a random process based on chance.

3) expected value

   Words such as 'normally' 'the usual amount' express the idea of expectation. That one might expect 35 matches in the box.

Section B2: Misconceptions and pupils' errors

The following are some misconceptions that can be expected.

1) any difference in the means between two groups is significant

   If the average height of pupils (12 - 13 years old) in Form 1A is 158 cm and in Form 1B the average is 160 cm, this difference need not be significant. It would be incorrect to conclude that the pupils in Form 1B are taller than the pupils in Form1A (additional information would be needed to make such a statement). It could be that there is one very tall girl in Form 1B causing the mean to go up, while all other pupils might in fact be below the height of 158 cm.

2) a lack of awareness of regression to the mean in everyday life

   If a process is repeated many times (throwing a dice) the number of times a six is thrown will come closer to the theoretical mean.

3) errors due to inconsistent notation used in different textbooks

   The conventions used in different textbooks are not the same and are problematic to pupils that are not fluent in the use of mathematical symbols. As project work implies consulting different textbooks a teacher should be aware of this possible source of errors. To avoid errors caused by differences in conventions a teacher should make it a rule NEVER to present concepts exclusively in symbolic algebraic format. A word format should be presented as well.

Section C: Mean, median and mode for ungrouped data

I(a) The mean of n listed values:

The mean of n numbers $a_1, a_2, a_3, \ldots a_{n-1}, a_n$ is (sum of all the data values) ÷ (number of data values) . In formula form

$$\frac{a_1 + a_2 + a_3 + \ \ldots \ + a_{n-1} + a_n}{n} = \frac{\sum_{i=1}^{n} a_i}{n}$$

Σ(pronounced: sigma) is the Greek letter for S. It is used to mean "the sum of ."

$\sum_{i=1}^{n} a_i$ means the sum of all the a_i where I take values from 1 to n.

It is short for $a_1 + a_2 + a_3 + \ldots + a_{n-1} + a_n$

I(b) The mean of numbers given in a frequency distribution

Number	a_1	a_2	a_3	.....	.....	a_{n-1}	a_n
Frequency	f_1	f_2	f_3	....	.....	f_{n-1}	f_n

The mean of $a_1, a_2, a_3, \ldots a_{n-1}, a_n$ if the numbers have frequencies of respectively $f_1, f_2, f_3, \ldots f_{n-1}, f_n$ is $\frac{\sum_{i=1}^{n} f_i a_i}{\sum_{i=1}^{n} f_i}$

The mean can be obtained from discrete or continuous quantitative data. Qualitative data *cannot* be summarised by a mean. If you have data on the favourite type of music of the pupils in your class it does not make sense to average 'gospel music' with 'reggae'.

II The **mode** is the observation with the highest frequency. The mode uses the frequencies and hence a mode can be obtained for both quantitative and qualitative data.

II(a) The mode of n listed observations.

For example:

(i) 10 people were asked for their favourite drink and the responses were tea, fruit juice, milk, fruit juice, fizzy drink, fruit juice, fizzy drink, milk, fruit juice, tea.

The mode being: fruit juice.

(ii) The height of ten plants was 1.64 m, 1.58 m, 1.67 m, 1.71 m, 1.65 m., 1.68 m, 1.65 m, 1.60 m, 1.65 m, 1.71 m.

The most frequent height is 1.65 m. The mode is 1.65 m. Both qualitative and quantitative data can have a mode.

II(b) The mode read from a frequency distribution.

Number of words in a sentence in a nursery rhyme.

No. of words	5	6	7	8	9	10
Frequency	3	3	6	5	4	4

The highest frequency is 6. The mode is 7. Seven words per sentence occurred most frequently.

III The **median** is the middle observation if the number of observations is odd or the mean of the two middle observations if the number of observations is even provided the data is written in increasing (or decreasing) order. Median can be obtained for quantitative data; qualitative data has no median.

III(a) The median of *n* listed values:

The scores of the school's netball teams during a tournament were
Team A: 2, 4, 5, 3, 4, 1, 7 Team B: 3, 6, 4, 2, 4, 5, 8, 6.

The median number of goals scored by the school's netball teams requires ordering the given data first:

Team A: 1, 2, 3, 4, 4, 5, 7

Team B: 2, 3, 4, 4, 5, 6, 6, 8.

The first team A had a median score 4, the middle of the seven ordered scores.

The middle value of *n* observation, when *n* is odd, is the $\frac{1}{2}(n+1)$th observation.

The median of team B is the mean of 4th and 5th observation (as the number of observation is even there are two numbers in the middle).

The median is therefore $\frac{4+5}{2}=4.5$.

The middle value of *n* observation, when even, is the mean of the $\frac{1}{2}n$th and the $\frac{1}{2}(n+1)$th observation.

III(b) The median of quantitative data given in a frequency distribution.

Using the same data as above

No of words	5	6	7	8	9	10
Frequency	3	3	6	5	4	4

The median is found by first finding the total number of observations $(3 + 3 + 6 + 5 + 4 + 4) = 27$.

The median is therefore the 13^{th} observation: the 3^{rd} observation is 5, the sixth $(3 + 3)$ is 6, the 12^{th} $(3 + 3 + 6)$ is 7, the 13^{th} is the first of the five, 8. The median is 8.

Self mark exercise 1

1. The heights of three friends are 1.56 m, 1.67 m and 1.61 m.
   Find the mean height.

2. The frequency distribution table shows the number of pips in 50 oranges.

number of pips	8	9	10	11	12	13	14	15
frequency	2	5	12	10	6	7	5	3

   Find the mean number of pips in the 50 oranges.

3. A survey in the class on the number of hours spent by pupils to study for the end of year Science examination gave the following data (to the nearest hour).

1	3	1	2	5
3	2	3	3	1
4	2	2	1	6
3	2	2	4	5

   Find the mean, mode and median of this data.

   Which of the three averages best represents the data? Justify your answer.

4. Number of days pupils were absent during one week in form 2A4.

Days absent	1	2	3	4	5	6
Frequency	6	0	2	2	1	1

   Find the mean, mode and median of this data.

   Which of the three averages best represents the data? Justify your answer.

5. The stem-leaf diagram represents the height of boys and girls in a class.

   Height of pupils in form 2X

Girls		Boys
443310	15	2
9865	15	579
43220	16	12234
865	16	5566889
42	17	044
	17	58

   $n = 41$ 16 | 8 represent 168 cm

   Find the mean, mode and median of (i) girls (ii) boys.

Continued on next page

Compare the averages of boys and girls and make some statements based on the averages calculated.

Represent the data in one bar chart (150-154, 155 - 159, 160-164, etc.) placing the bars for boys and girls next to each other (double bar chart).

6. Tashata collected data on the number of people in a car, parking in front of a shopping centre between 09 00 h and 10 00 h.

Number of people	1	2	3	4	5	6
Number of cars	42	20	14	12	10	2

a) Find the mean, mode and median.

b) Represent the data in a graph/chart. Justify your choice.

7. Packets of crisps are marked 30 g. A sample of packets was taken and the mass of crisps determined to nearest gram. The results are as listed:

Mass (g)	29	30	31	32	33
Frequency	20	50	45	10	25

a) Find the mean, mode and median.

b) Represent the data in a histogram.

8. Two running teams A and B took part in a competition. The times, to nearest 0.1 of a second, of the members of team A and team B in the 100 m are listed below.

Team A	Team B
11.9	12.0
12.4	12.5
13.1	13.2
13.3	13.8
13.6	14.1
12.9	12.6
13.8	13.5
13.3	12.7
12.8	12.1
12.4	13.7
12.0	11.8

a) Find the mean, mode and median.

b) Which average best represents the data? Explain.

c) Which is the ‘better’ team? Justify your answer.

Suggested answers are at the end of this unit.

Section D: Which is the best average to use?

Which is the best average to use depends on the situation and what you want to use the average for. The mean is the most commonly used measure of central tendency as it is the only one of the three averages using all the data. It takes all the data in the distribution into account. The mean—being arithmetically based—can be combined with the means of other groups on the same variable. For example: If you found that the average score on a mathematics test in class 1A was 68% and in 1B the average score was 71%, the average score of the combined class 1AB can be computed. The median and the mode, not being arithmetically based, do not have such a property. If the modal height in form 1A is 165 cm and in 1B is 167 cm, you cannot draw any conclusion with regards to the mode of the combined class 1AB.

However using the mean can give a rather distorted picture of the data if there are outliers, or if the mean is not meaningful in the given context.

Examples:

1. The leader of a youth club can get discounts on cans of drinks if she buys all one size. She took a vote on which size the members of the club wanted.

Size of can (ml)	100	200	330	500
Number of votes	9	12	19	1

   Mode = 330 ml, median = 200 ml and mean = 245.6 ml (1 decimal point)

   Which size should she buy?

   The mean is clearly of no use—cans of size 245.6 ml do not exist. The median would be possible as 200 ml cans are for sale. However only 12 out of the 41 club members want this size. In this case the mode is the best average to use as it is the most popular one among the club members.

2. A teacher wants 50% of his pupils to get a credit in a test and wants to set the minimum mark for the credit. Which average should the teacher use? In this case the median is the only appropriate average as this is the middle score—50% will be above and 50% below this score. (N.B. the median mark can only be set after the test is written).

3. In a small butchery the four labourers earn each R 400 per month, the supervisor earns R 1200 and the manager R 2600. Which average best represents the monthly wages earned?

   The mean (R 900) is misleading as it is more than twice the salary earned by most workers. The median (R 400) is representative. The other appropriate average to use is the mode: it gives the wages of most of the workers. When datasets are skewed to one side, like wages or house prices, the median and mode are more realistic than the mean.

4. The time taken (in hours) by 6 pupils to complete their project was 20, 25, 31, 35, 87, 87.

   The mean is 47.5 but most pupils worked less than that on their project. The mode is 87 but is also misleading. The median is

$33\left[\frac{1}{2}(31+35)=33\right]$ which is the best to use as it tells us that half of the number of pupils needed less than 33 h and half needed more.

5. The number of border crossings at 5 border posts between Botswana and Zimbabwe on a certain day were 40, 60, 60, 80 and 810. The median is 60, the mode is 60 and the mean is 210. The outlier 810 makes the mean move to 210, a value atypical for the data. The median would be a better value to represent the data.

6. Occasionally, distributions arise for which none of the averages is particularly informative. For example:

   The table shows the number of cigarettes smoked per day by 50 persons.

Number of cigarettes	Number of people
0	30
1	10
2	5
3	3
4	2

   The mean is 7.4, the mode is 0 and the median is 0. None of these averages represent the data well. In this case it would be better to state that 60% are non-smokers and that the smokers smoke on average (mean) 18.5 cigarettes a day.

The *mean* is generally used if the data is more or less symmetrically grouped about a central point, i.e., the data do not contain outliers. If further calculation is required (e.g., measures of dispersion) or comparison with a similar measure on another group is intended, or the (sample) mean is to be used in estimating parameters of the population then the mean is to be used as mode and median cannot be used in 'further' calculations.

A distribution with outliers is frequently best described by using the *median*.

The *mode* is used when the context suggests 'most usual' or 'typical' value.

Self mark exercise 2

1. Decide which of the following averages—mean, mode or median—is the most appropriate to use to summarise the following data. Justify your answer.

   a) number of children in the family
   b) number of letters in pupils' surnames
   c) number of pupils born in a certain month
   d) shoe size of the boys in the class
   e) favourite subject in school
   f) number of days each pupil in a class was absent during the term
   g) method of payment for goods bought in a furniture shop
   h) most popular activity of pupils during a long weekend

Continued on next page

2a. Calculate mean, median and mode for each of the following sets of data.

b. Decide and justify which of the three best represents the data.

(i) Modise scored the following number of goals during the eight matches of the school's football team:

1, 1, 0, 6, 2, 1, 3, 0

(ii) A pupil scored the following percent grades in geography:

75%, 72%, 68%, 57%, 62%, 10%, 75%.

3. There are six possible ways of listing the three averages in ascending order of size. For example: mean, mode, median; mean, median, mode, etc. Write down all six and then try to find sets of numbers which will fit each arrangement.

For example, to make the order mode < median < mean you could choose the four numbers 1, 1, 3, 11 with mode 1, median 2 and mean 4.

4. Averages are meant to represent the data. List advantages and disadvantages of the mode, median and mean and give examples when you would use one instead of the others.

5a. (Challenge question for the strong mathematicians.) In mathematics you might also meet the **geometric mean** of numbers. The geometric mean of p and q is $\sqrt{(pq)}$.

Show that always for two numbers (arithmetic mean) ≥ (geometric mean).

b. The harmonic mean H of two numbers p and q is defined as $\frac{1}{H} = \frac{1}{p} + \frac{1}{q}$.

Express H in terms of the geometric and arithmetic mean.

Suggested answers are at the end of this unit.

Section E: Mean, mode and median: classroom lessons

The following pages give suggestions for an investigative approach to mean, mode and median. It is assumed that pupils have basic knowledge of what each of these measures stands for. The sequence of activities starts with a diagnostic test. This will give feedback to the teacher as to the ideas of the pupils and help to plan the lessons: which points need to be emphasised. At the end of the lesson sequence the teacher might decide to discuss some of the questions in the diagnostic test with the pupils.

The main objective is to make pupils aware of the effect on the three measures (mean, mode and median) when values are added to the group. For example, a new pupil, age 12 years 3 months, height 1.58 m, mass 58 kg joins the class. How will the mode, median and mean age / length / mass of the class change? The lessons also aim at developing in pupils an awareness as to which measure is most appropriate to use in a given context. The lessons use an investigative method, with pupils working in groups, to discuss points with

each other before (if needed) whole class discussion is used to round up the activities.

Practice task 1

1. Work through the diagnostic test and lesson outlines by yourself. Write down any problems you encounter.
2. Administer the diagnostic test to your class and analyse the results. Write a report on your findings.
3. Based on the outcomes of the diagnostic test make, if necessary, changes in the lesson outline and work out the lessons in detail. Prepare detailed lesson plans and notes. Prepare the worksheets for the pupils. You might need different versions for different levels of achievement of your pupils (differentiated worksheets to meet 'mixed ability' of your class).
4. Try out the planned activities / lessons and write an evaluative report. Cover questions such as:

   Were the objectives attained? How do you know?

   Did pupils enjoy the activities?

   What needs changing in the material?

   Was the method different from what you used to use?

Present the assignment to your supervisor.

A diagnostic instrument

Answer the following questions. Give reasons for your answers.

1. Suppose you have calculated the average age of a family. Afterwards a baby is born in that family. If you were asked to re-calculate the average age of that family, your task is very simple. Since the baby is 0 years old the average age will be exactly the same.

   TRUE / FALSE (circle the correct answer)

   Reason:

2. Somebody has calculated the average of a set of numbers. She tells her friend that if you take the total of all differences between the average and the numbers in the set you will always find zero.

   TRUE / FALSE (circle the correct answer)

   Reason:

3. A mode of a set of scores is the score with the highest frequency. What is the mode of the following set of scores: 60, 58, 33, 98, 58, 60, 42, 58

   A) 60 B) 58 C) 33 D) 98 E) 42 (circle correct answer)

   Reason: