basic education

Department:
Basic Education
REPUBLIC OF SOUTH AFRICA

Mathematics Teachers' Self Study Guide on the national Curriculum Statement

Book 2 of 2

WORKING WITH GROUPED DATA

Material written by
Meg Dickson and Jackie Scheiber
RADMASTE Centre, University of the Witwatersrand

The National Curriculum Statement for Grade 10, 11 and 12 (NCS) mentions grouped data in the following Assessment Standard in Grades 10:

10.4.1

a) Collect, organise and interpret univariate numerical data in order to determine
   - Measures of central tendency (mean, median, mode) of **grouped and ungrouped data** and knows which is the most appropriate under given conditions
   - Measures of dispersion: range, percentiles, quartiles, interquartile and semi-interquartile range

b) Represent data effectively, choosing appropriately from:
   - Bar and compound bar graphs
   - **Histograms (grouped data)**
   - **Frequency polygons**
   - Pie charts
   - Line and broken line graphs

UNIVARIATE DATA

Univariate data is data concerned with a single attribute or variable. When we graph univariate data, we do so on a pictogram, bar graph, pie chart, histogram, frequency polygon, line or broken line graph. Univariate data looks at the range of values, as well as the central tendency of the values.

Examples of univariate data are:
- Height of learners in Grade 11
- Length of earthworms in a soil sample
- Number of cars manufactured in a particular year
- Number of people born in a particular year

There are 2 forms of numerical data:

a) Information that is collected by counting is called **discrete data**. The data is collected by counting exact amounts and you list the information or values.

   e.g. the number of children in a family; the number of children with birthdays in January; the number of goals scored at a soccer match.

b) **Continuous data** values form part of a continuous scale and the values can not all be listed,

   e.g. the height of learners in a Grade 8 measured in centimetres and fractions of a centimetre; temperature measured in degrees and fractions of a degree.

   The mass of a baby at birth is **continuous** data, as there is no reason why a baby should not have a mass of 3,25167312 kg – even if there is no scale that could measure so many decimal places. However, the number of children born to a mother is **discrete** data, as decimals make no sense when counting babies!

TABLES, LISTS AND TALLIES

When you first look at data, all you may see is a jumble of information. You need to sort the data and record it in a way that makes more sense. Some data is easy to sort into lists that are either numerical or alphabetical. Other data can be sorted into **tables**. Some tables can be used to keep **count** of the number of times a particular piece of data occurs; such a table is called a **frequency table**. In a frequency table you can also find a 'running total' of frequencies. This is called the **cumulative frequency.** It is sometimes useful to know the running total of the frequencies as this tells you the total number of data items at different stages in the data set.

1) STEM AND LEAF DISPLAY

Example:
Suppose the members of your Grade 11 maths class scored the following percentages in a maths test:

32 ; 56 ; 45 ; 78 ; 77 ; 59 ; 65 ; 54 ; 54 ; 39 ;
45 ; 44 ; 52 ; 47 ; 50 ; 52 ; 51 ; 40 ; 69 ; 72 ;

36 ; 57 ; 55 ; 47 ; 33 ; 39 ; 66 ; 61 ; 48 ; 45 ;
53 ; 57 ; 56 ; 55 ; 71 ; 63 ; 62 ; 65 ; 58 ; 55 ;

This data is **discrete data.** The percentages are numbers representing the count of marks on the test scripts.

This list of numbers has little meaning as it is. However, by organising the data into tables we can begin to make some sense out of the numbers. One way of organising them would be in a stem & leaf plot.

3	2	3	6	9	9											
4	0	4	5	5	5	7	7	8								
5	0	1	2	2	3	4	4	5	5	5	6	6	7	7	8	9
6	1	2	3	5	5	6	9									
7	1	2	7	8												

Key: 6/2 = 62

Notice the stem and leaf display is visual representation of the data. It is easy to see that there are more marks in the fifties than in the seventies.

2) GROUPED FREQUENCY TABLE

Another way of organising the list of marks would be to write them in a **grouped frequency table**. In this sort of table the numbers are arranged in **groups** or **class intervals**.

Maths marks:
32 ; 56 ; 45 ; 78 ; 77 ; 59 ; 65 ; 54 ; 54 ; 39 ; 45 ; 44 ; 52 ; 47 ; 50 ; 52 ; 51 ; 40 ; 69 ; 72 ; 36 ; 57 ; 55 ; 47 ; 33 ; 39 ; 66 ; 61 ; 48 ; 45 ; 53 ; 57 ; 56 ; 55 ; 71 ; 63 ; 62 ; 65

Rewrite the list into groups of multiples of ten like this:

marks	**tally**	**frequency**
30 - 39	//// /	5
40 - 49		
50 - 59		
60 – 69		
70 - 79		

There are 5 marks in the class interval 30 -39

- *Now complete the grouped frequency table*

Note:

- You list the number of times each number in the group occurs i.e. the frequency of each of the number.

- The groups do not overlap at all. Notice how the groups are written **30 – 39** and then **40 – 49** . You cannot write the groups as **30 – 40** and then **40 – 50** etc, as you wouldn't know where to record a mark of 40.

- This data is discrete data. You can also group **continuous data**. For continuous data you would write the groups with inequality signs like this:
  20 ≤ m < 30 to show that in this group the marks can equal 20 but not equal 30. When grouping measurements the class intervals might include decimal values.

Activity 1:

1) 30 learners were asked in a survey to say how many hours they spent watching TV in a week. Their answers (correct to the nearest hour) are given below:

   12 ; 20 ; 13 ; 15 ; 22 ; 3 ; 6 ; 24 ; 20 ; 15 ; 9 ; 12 ; 5 ; 6 ; 8
   30 ; 7 ; 12 ; 14 ; 25 ; 2 ; 6 ; 12 ; 20 ; 20 ; 18 ; 3 ; 18 ; 8 ; 9

   a) Complete the grouped frequency table for the data shown below:

group	tally	frequency
0 – 10 hrs		
11 – 20 hrs		
21 – 30 hrs		
	TOTAL	30

   b) Write down two questions that you could ask about this data.

Activity 1 (continued):

2) Sipho wanted to find out about the heights of learners in his Grade 12 class. He collected the data below: (the measurements are all in metres)

1,82	1,64	1,71	1,77	1,64	1,67	1,73
1,80	1,76	1,78	1,52	1,63	1,65	1,67
1,86	1,90	1,71	1,64	1,58	1,81	1,87
1,67	1,74	1,69	1,75	1,68	1,74	1,61
1,79	1,83	1,69	1,58	1,57	1,73	1,54

a) What sort of data is this?

b) Use groups intervals of $1{,}50 < x \leq 1{,}55$; $1{,}55 < x \leq 1{,}60$; etc to draw up a grouped frequency table for Sipho's data.

c) How many learners are taller than 1,75m?

MEASURES OF CENTRAL TENDENCY FROM A FREQUENCY TABLE

You can easily find the mean, median and mode of a data set when it is shown in a frequency table.

1) The mean

The **mean** is the sum of the data items divided by the number of items.

- We use the following formula to find the mean of **ungrouped data**:
  Mean = $\bar{x} = \frac{\sum x}{n}$, where $\sum x$ = sum of the data items, and n = the number of items

- We use a modified formula when finding the mean of **grouped data**:
  Mean = $\bar{x} = \frac{\sum f.x}{n}$, where f = the frequency, x is the value of the item, and
  n = the number of items

2) The median

The **median** is the middle data item when the data is listed in order. We sometimes use the formula $\frac{n+1}{2}$ to find out which item is the middle item, and can also find the **median** from the frequency table.

3) The mode

The **mode** is the data item with the highest frequency.

Example:
You collect the marks of the learners in Grade 9 in a general knowledge quiz. You record the marks in a frequency table where x is the value of the mark and f is the frequency.

Mark obtained x	**Frequency** f	$f.x$
0	10	0
1	20	
2	40	
3	50	
4	30	
5	30	150
6	20	
7	20	
8	10	
9	10	
10	10	
	$n = 250$	$\Sigma f.x =$

1) Complete the last column by multiplying together the mark obtained and the frequency.
2) Calculate $\sum f.x$
3) Calculate the mean using the formula:

$$\bar{x} = \frac{\sum f.x}{n} =$$

4) There are 250 items in this data set. Substituting, we get:
$\frac{n+1}{2} = \frac{250+1}{2} = \frac{251}{2}$ = 125 ½
So the median is the 125 ½ th item.

Add the frequencies together to find out which data item is the median

The median = ...

5) The mode = ...

Activity 2:

The numbers of children per family in a township are recorded in the table below.

No of children	0	1	2	3	4
No of families	12	15	5	2	1

1) Redraw the frequency table so you can find the mean number of children per family.

2) Determine the median of the data from the frequency table

3) Determine the mode of the data from the frequency table.

MEASURES OF CENTRAL TENDENCY FROM A GROUPED FREQUENCY TABLE

You can also find averages when the data is grouped.

Example:
The height of 40 trees is measured. The measurements, given in metres, are grouped as shown in the table below:

Height (h) In metres	frequency
$2 \le h < 3{,}99$	2
$4 \le h < 5{,}99$	6
$6 \le h < 7{,}99$	11
$8 \le h < 9{,}99$	12
$10 \le h < 11{,}99$	8
$12 \le h < 13{,}99$	1
	n = 40

The heights of the trees are measurements so this is **continuous**

This means 12 trees were between 8 and 9,9 ...

The group with the highest frequency in the table is the group $8 \le h < 9{,}99$.

∴ Modal class is $8 \le h <$

Total frequency = 40
Median is 20 ½ item
(2+6+11=19)
∴ Median is in interval $8 \le h < 9{,}99$.

1) The modal class:
The mode is the item of data that occurs most often. The modal class is therefore the group (or class interval) that occurs most often. For this data the **modal class** (the height that occurs most often) is **$8 \le h < 9{,}99$ m**

2) The median from a grouped frequency table.
It is also relatively easy to find the median from a frequency table. When the data is grouped you use the same method – counting up the items. It does not matter whether the data is discrete or continuous, the same method applies.
In this case the median will lie half-way between the 20th and the 21st item.
So the **median will lie in the class $8 \le x < 9{,}99$ m**

3) The mean from a grouped frequency table
As you saw previously, it is relatively easy to find the mean from a frequency table. When the data is grouped you use a similar method.

However, it is first necessary to find a single value to represent each class. This single value is the **midpoint** of the interval.

The next activity illustrates how you can find these averages from grouped data in a frequency table.

Activity 3

1) Suppose you asked a group of men to count the number of items in their pockets.

Number of items	0-4	5-9	10-14	15-19	20-24
Frequency	6	11	6	4	3

- This frequency table has been redrawn below as a vertical table. Notice there are 2 extra columns this time.
- When you find the mean from a frequency table you need to multiply the frequency ***f*** by the data item ***x.*** But the data items in a grouped table are groups, so first you need to find the **midpoints of the groups**. Notice that the numbers 0, 1, 2, 3 and 4 are included in the group 0 – 4. The middle score is thus 2.
- Notice that we use x to represent the actual value and X to represent the midpoint. We therefore use $\bar{x}$ to represent the mean when using actual values, and $\bar{X}$ to represent the approximate mean which is found using the midpoint of the interval.

No of items	**frequency** ***f***	**midpoint of groups** ***X***	***f.X***
0 – 4	6	2	
5 – 9	11	7	
10 – 14	6		
15 – 19	4		
20 – 24	3		
	***n* = 30**		**Σ *f.X* =**

a) Is this discrete or continuous data?

b) Calculate the midpoint ***X*** of each of the groups, and fill it in on the table.

c) Complete the last column of the table.

d) Calculate the mean using the formula below:
Mean = $\frac{\sum fX}{n} = \frac{\qquad}{30}$ =

e) Determine the median.
The median is $\frac{n+1}{2} = \frac{\qquad}{2}$ = item
∴Median is in the interval .....

f) Determine the modal class:

Modal class is: .....

Activity 3 (continued)

2) A survey was done to find out the number of trees of various heights

Notice:
- The group 2 ≤ h < 3,99 contains all the measurements from 2 m to 3,99 m
- The midpoint of the group = $\frac{2+3,99}{2} = \frac{5,99}{2}$ = 2,995 ≈ 3 (correct to the nearest cm)

Height (in metres)	**midpoint *X***	**Frequency *f***	***f.X***
2 ≤ h< 3,99	3	2	6
4 ≤ h< 5,99	5	6	
6 ≤ h< 7,99		11	
8 ≤ h< 9,99		12	
10 ≤ h< 11,99		8	
12 ≤ h< 13,99		1	
		***n* = 40**	**Σ *f.X* =**

a) Is this discrete or continuous data?

b) Complete the frequency table by:
i) Finding the midpoints of each interval, ***X***
ii) Multiplying the midpoints ***X,*** by the frequency ***f***, to obtain the value ***f.X***.

c) Calculate the mean using the formula below:

Mean = $\frac{\text{total of 40 scores}}{\text{total frequency}} = \frac{\sum fX}{n} = \frac{.....}{40}$ = ……

d) Determine the median from the frequency table.

Median is $\frac{n+1}{2} = \frac{\quad}{2} =$ item

∴Median is in the interval ……

e) Determine the modal class.

Modal class = ……

THINGS TO NOTICE WHEN FINDING THE MEAN FROM A FREQUENCY TABLE:

1) **For ungrouped data:**
   - Multiply the data item or score (***x***) by the frequency (***f***) and record this in an extra column in the frequency table (***fx***)
   - Calculate the arithmetic mean of *fx* i.e. calculate
     $\frac{\text{total of scores}}{\text{total frequency}} = \frac{\sum fx}{n}$
2) **For grouped data**
   - First find the **midpoint** (***X***) of each group or class interval
   - Multiply each midpoint (***X***) by the frequency of that group (***f***) and record this in an extra column in the frequency table (***fX***)
   - Calculate the arithmetic mean of *fX* i.e. calculate
     $\frac{\text{total of scores}}{\text{total frequency}} = \frac{\sum fX}{n}$

Activity 4:

1) The local bus company went through a period when its buses always left the city-centre late. The data is shown in the table below:

Minutes late	0-10	11-20	21-30	31-40	41-50
Frequency (no of buses)	5	8	21	14	5

a) What sort of data is this?
b) Redraw the frequency table and then find:
   i) the mean number of minutes late
   ii) the modal class
   iii) the median of the data

Activity 4 (continued):

2) The table below shows the distribution of the heights of 50 female teachers in your district.

Height (cm)	$149{,}5 \leq h < 154{,}5$	$154{,}5 \leq h < 159{,}5$	$159{,}5 \leq h < 164{,}5$	$164{,}5 \leq h < 169{,}5$
Frequency	4	21	18	7

a) What sort of data is this?
b) Redraw the frequency table and then find:
   i) the mean height of the teachers
   ii) the modal class
   iii) the median of the data

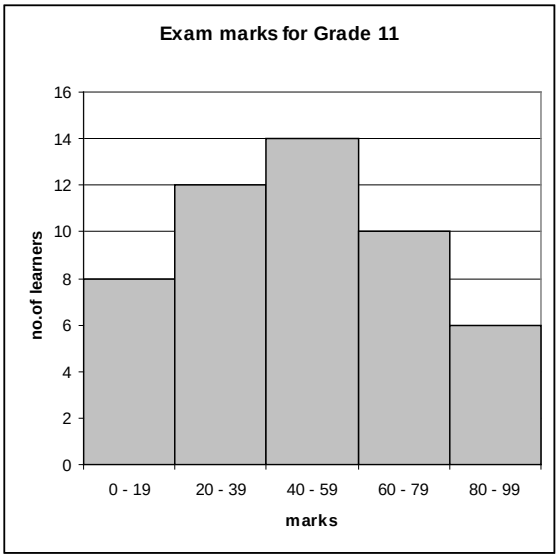
DRAWING HISTOGRAMS

Graphs of **grouped data** can be drawn for discrete data (**bar graphs**) and continuous data (**histograms**). If the class intervals of the grouped data are equal the bars of the histogram will be of equal width.

Histograms are similar to bar graphs but, because they represent continuous data, the 'bars' (or columns) are joined together. The horizontal (x) axis will always be a number line.

This histogram represents the exam marks of Grade 11 learners at the end of the year.
Notice:

- the 'bars' are joined
- the 'bars' are the same width
- the scale on the horizontal axis is continuous.



Activity 5:
Look at the data that Sophia collected about the height of learners in her Grade 12 maths class. The heights are given in metres.

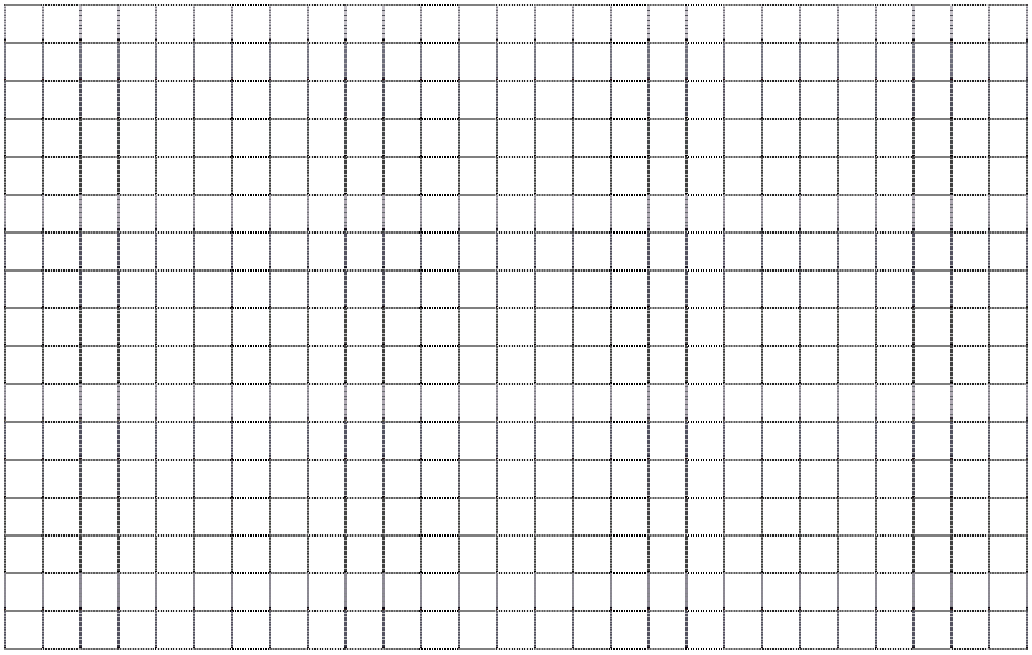
1,82	1,64	1,71	1,77	1,64	1,67	1,73
1,80	1,76	1,78	1,52	1,63	1,65	1,67
1,86	1,90	1,71	1,64	1,58	1,81	1,87
1,67	1,74	1,69	1,75	1,68	1,74	1,61
1,79	1,83	1,69	1,58	1,57	1,73	1,54

1) Complete the grouped frequency table for Sophia's data.

groups	**Midpoint** ***X***	**frequency** ***f***	***f.X***
$1{,}50 \leq x < 1{,}55$	$1{,}525 \approx 1{,}53$		
$1{,}55 \leq x < 1{,}60$			
		***n* =**	$\sum f.X$ =

Activity 5 (continued):

2) Draw a histogram illustrating Sophia's data on the grid below.



3) a) Use the frequency table to find the mean of the data.

b) Mark this on the histogram

4) Find the median and the modal class of the data and mark them on the histogram

FREQUENCY POLYGONS

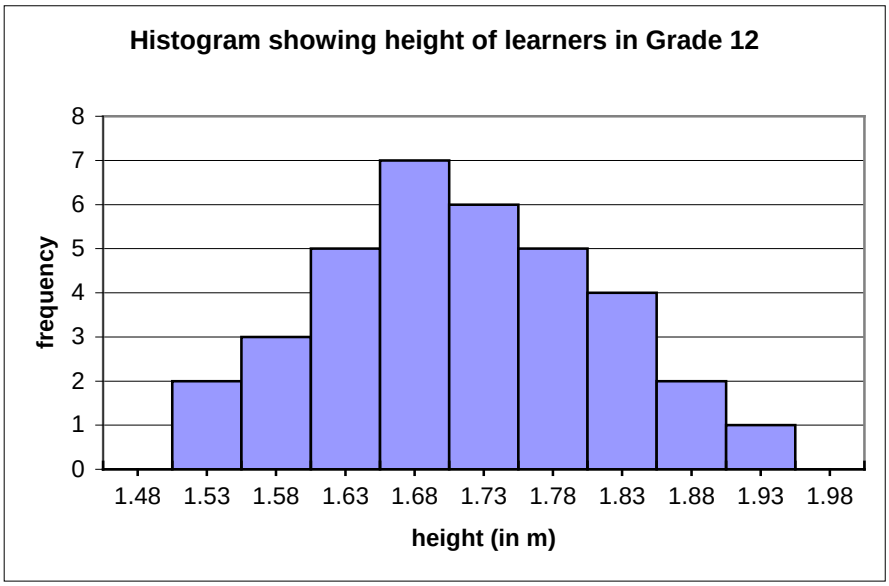
A **frequency polygon** is another representation of data that is closely linked to a histogram.

A frequency polygon is constructed by plotting the **middle point** of each class interval (i.e. each bar) of the histogram. The midpoints are then joined by straight lines to form a polygon. In order to create a polygon (i.e. a closed 2-D shape made up of straight lines), it is important to include an extra interval to the left and to the right of the required intervals.

Example:
Look again at the data Sophia collected about the height of learners in her Grade 12 maths class. The heights are shown in the table below and are given in metres.

1,82	1,64	1,71	1,77	1,63	1,64	1,58
1,80	1,76	1,78	1,52	1,65	1,57	1,67
1,86	1,90	1,71	1,64	1,54	1,73	1,73
1,67	1,74	1,69	1,75	1,67	1,68	1,81
1,79	1,83	1,69	1,58	1,61	1,74	1,87

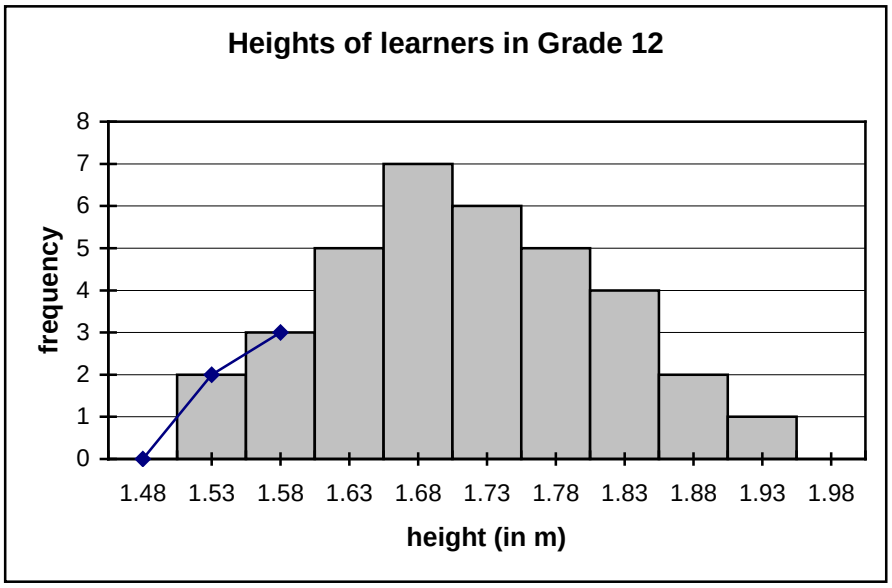
You drew a grouped frequency table and a histogram to represent the data. Did your histogram look like this?



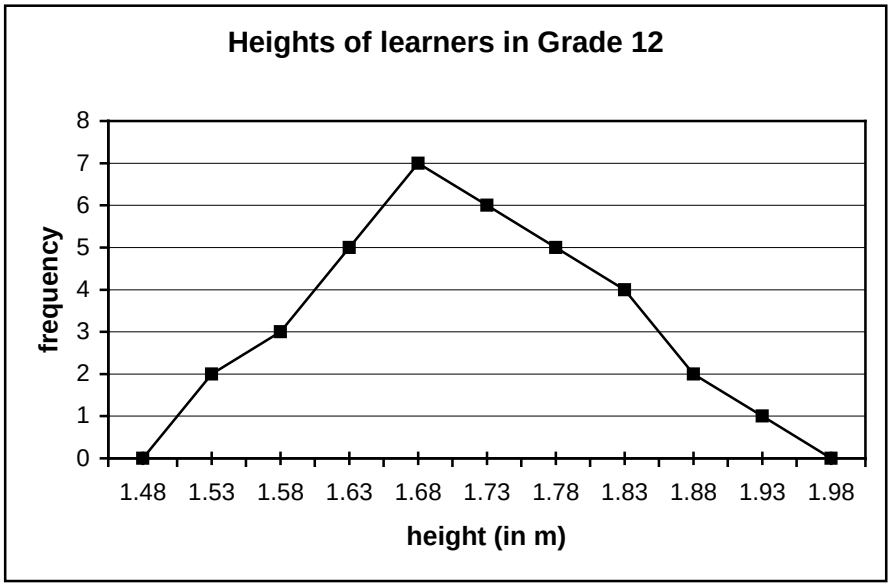
Notice the following:

- The horizontal axis is actually a number line.
- You can mark in values on the horizontal axis at the beginning of an interval, or in the middle of an interval. If you use the computer for drawing the graph, the package you use may determine where the values are placed on the horizontal axis. With MSWord, the values are placed in the middle of the interval.

On the following histogram **join up the midpoints** of the bars and construct a frequency polygon. (The first two have been done for you.)



Here is the completed frequency polygon without the histogram.



Note:

- **The frequency polygon starts and ends on the horizontal axis.** The beginning point of the polygon is the midpoint of the class interval below the first class interval of data. The end point is the midpoint of the class interval after the last group of the data.

- It is not necessary to first draw a histogram before drawing a frequency polygon.
  - Insert a class interval at the beginning and the end of the frequency table with a frequency of zero
  - Find the mid points of the class intervals
  - Plot the frequencies for each midpoint
  - Join the points with straight lines to form the polygon.

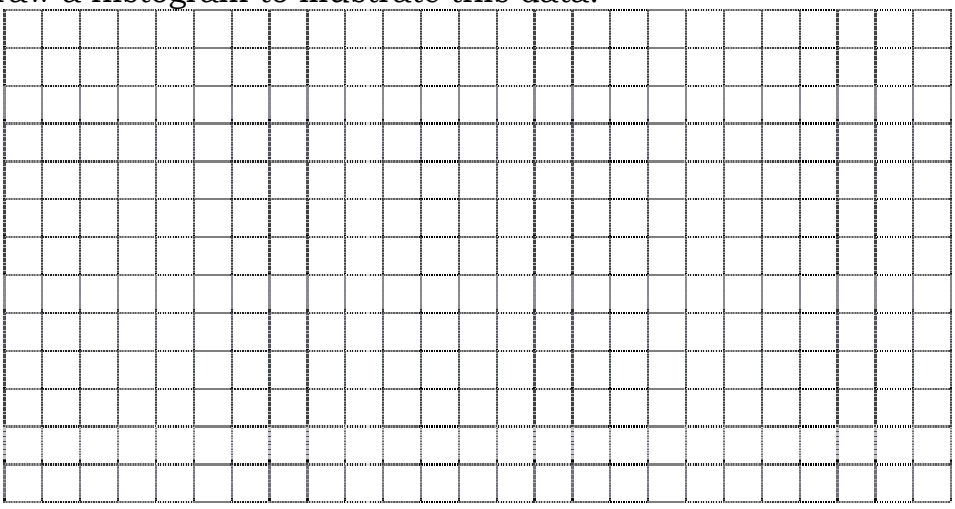
Activity 6

1) The frequency table below shows the number of people (by age) attending a film premier:

   Notice: two extra columns have been included in the table where the frequency is zero – at the beginning and the end of the data

Age, in years	$10 < n \leq 20$	$20 < n \leq 30$	$30 < n \leq 40$	$40 < n \leq 50$	$50 < n \leq 60$	$60 < n \leq 70$	$70 < n \leq 80$
frequency	0	42	57	63	26	22	0

   a) Draw a histogram to illustrate this data.



   b) Draw a frequency polygon on the same set of axes.

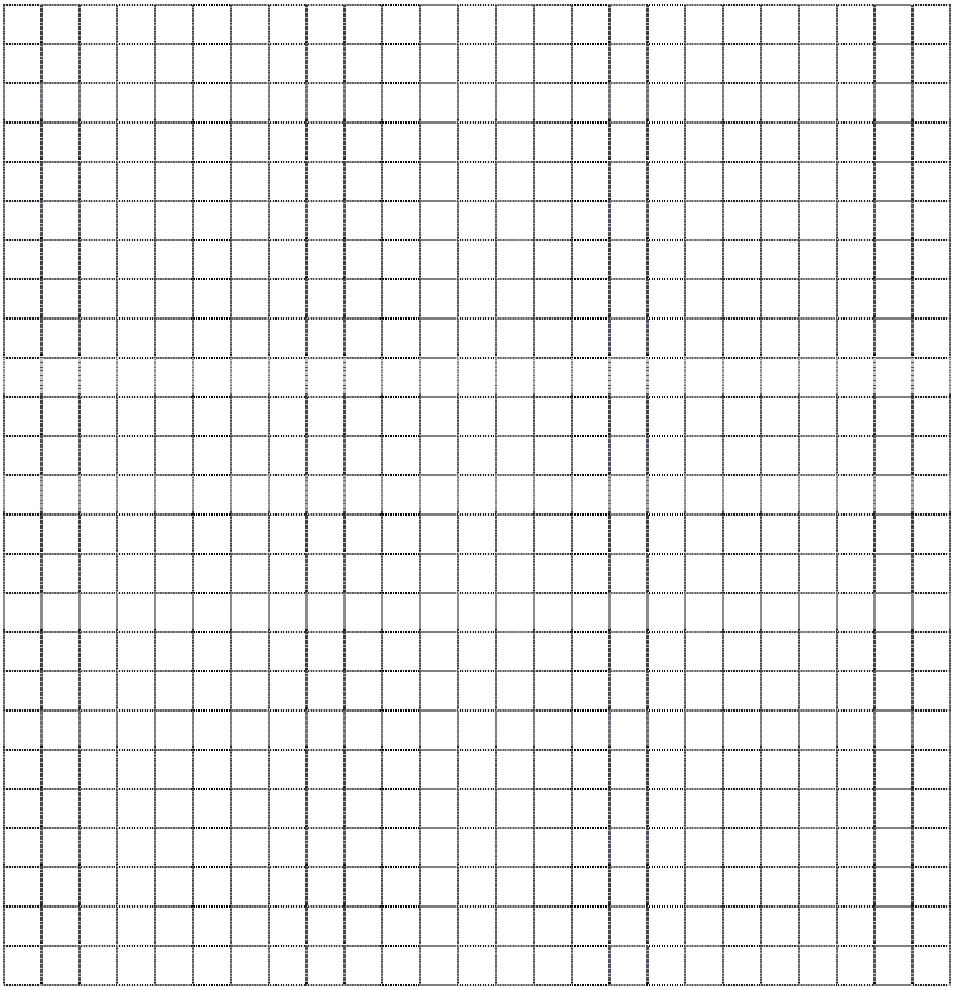
Activity 6 (continued)

2) The heights of 80 learners were recorded. The data is shown in the table below:

Height (in cm)		150 < x ≤ 160	160 < x ≤ 170	170 < x ≤ 180	180 < x ≤ 190	190 < x ≤ 200	200 < x ≤ 210	
No of learners	0	4	7	15	47	6	1	0

a) Illustrate this data by means of a histogram

b) On the same set of axes draw a frequency polygon to illustrate the data.



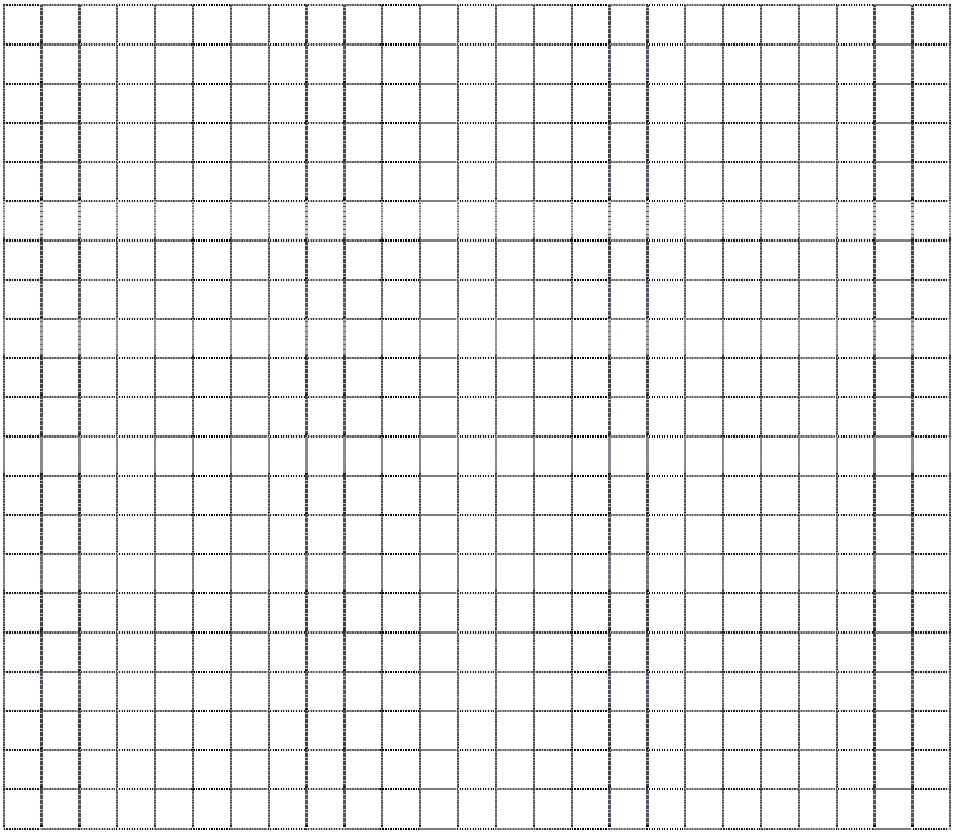
Activity 6 (continued)

3) The maths exam results for two successive Grade 12 years are recorded in the table:

Marks as %		$1 < x \leq 20$	$21 < x \leq 40$	$41 < x \leq 60$	$61 < x \leq 80$	$81 < x \leq 100$	
Group A frequency	0	5	12	35	28	20	0
Group B frequency	0	7	26	48	9	10	0
Mid point of interval							

a) Complete the last row of the table

b) Draw a frequency polygon for each set of data on the same set of axes. If possible, use two different colours for the two frequency polygons.

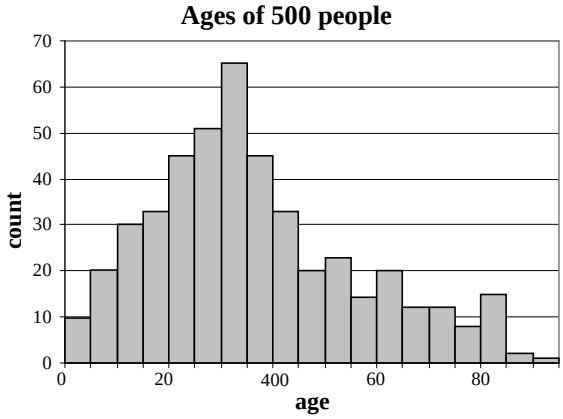


c) Assuming that the ability of the learners was the same in each year, what can you say about the exam papers?

Activity 6 (continued)

4) The histogram below shows the ages in 1990 of 500 people chosen at random in South Africa. Study the histogram then answer the questions below

a) Describe the main features that you see



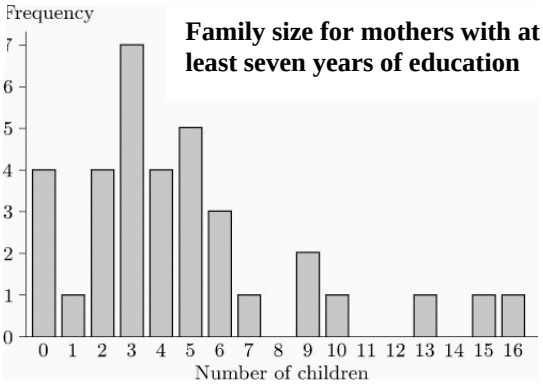
b) The tallest column shows people between the ages of 30 and 35. When were these people born?

c) Sketch how you think this distribution might look in 2010.

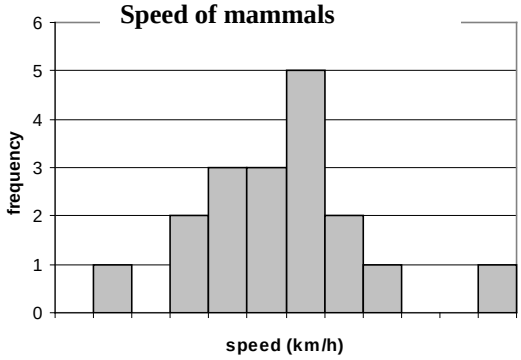
DIFFERENT SHAPES OF HISTOGRAMS

The 'shape' of a histogram can vary considerably depending on the data it represents. The four histograms below illustrate this.

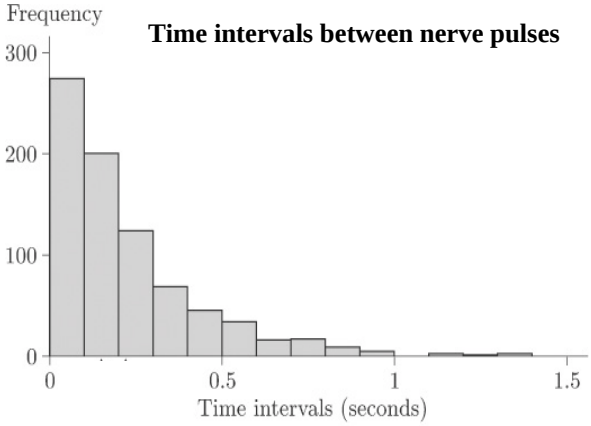
A



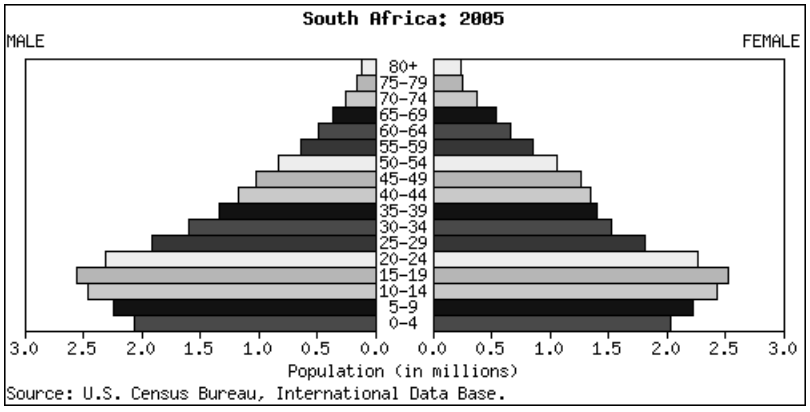
B



C



D



- Diagrams **A** and **C** show histograms in which the distributions that are 'skewed' to one side
- Diagram **B** shows a histogram representing an almost symmetrical distribution.
- Diagram **D** shows a double histogram of the female and male population of South Africa

NORMAL AND SKEWED DATA

Material written by
Meg Dickson and Jackie Scheiber
RADMASTE Centre, University of the Witwatersrand

The National Curriculum Statement for Grade 10, 11 and 12 (NCS) mentions normally distributed data in the following Assessment Standards in Grades 11 & 12

11.4.1 (a)
Calculate and represent measures of central tendency and dispersion in univariate numerical data by calculating the variance and standard deviation of sets of data manually (for small sets of data) and using available technology (for larger sets of data), and representing results graphically using histograms and frequency polygons.

11.4.4
Differentiate between symmetric and skewed data and make relevant deductions

12.4.4
Identify data which is normally distributed about a mean by investigating appropriate histograms and frequency polygons

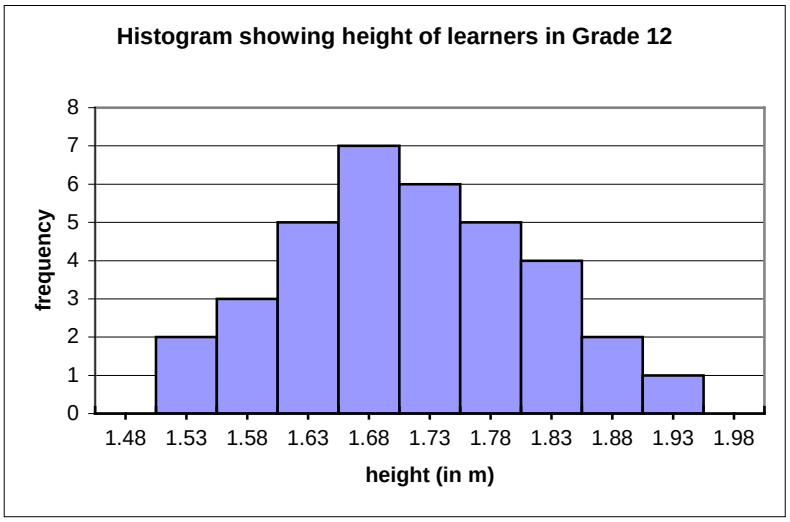
HISTOGRAMS and FREQUENCY POLYGONS

You already know how to represent grouped data as a **histogram** and/or a **frequency polygon**. Remember the class intervals are equal so the histogram is similar to a bar graph, but the columns 'touch' one another. A histogram is drawn from a frequency table. The following example is given to reminder you about histograms and frequency polygons.

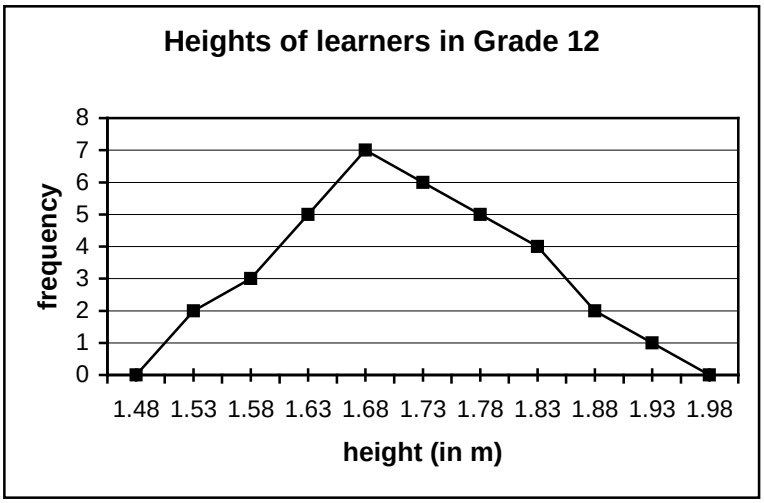
Example:
Look again at the data Sophia collected about the height of learners in her Grade 12 maths class. The heights are shown in the table below and are given in metres.

1,82	1,64	1,71	1,77	1,63	1,64	1,58
1,80	1,76	1,78	1,52	1,65	1,57	1,67
1,86	1,90	1,71	1,64	1,54	1,73	1,73
1,67	1,74	1,69	1,75	1,67	1,68	1,81
1,79	1,83	1,69	1,58	1,61	1,74	1,87

She drew a grouped frequency table using the class intervals $1{,}50 \leq x < 1{,}55$;
$1{,}55 \leq x < 1{,}60$; etc and then drew the following histogram:



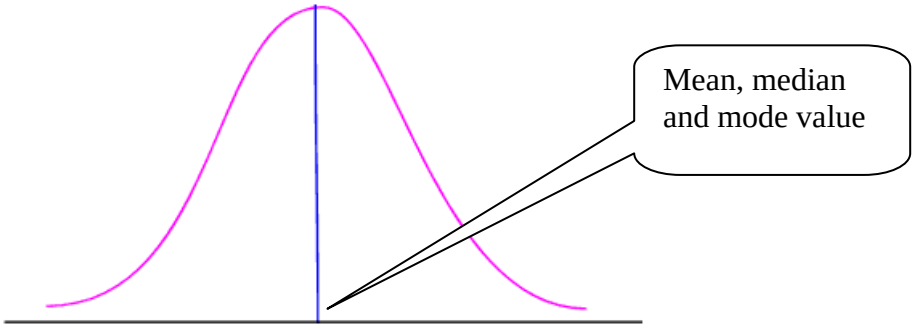
She then joined the midpoints of the bars and drew a frequency polygon as shown below:



Notice:
A frequency polygon like this shows the **distribution** of the data collected. Notice that the graph looks a little like a bell.

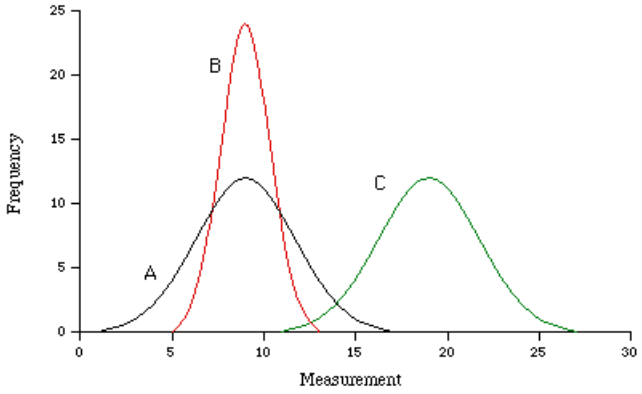
NORMAL CURVES

If you draw a frequency polygon for a set of data (and smooth the lines out) and the graph is **evenly distributed** and **symmetrical** you get what is called a **normal curve**. The normal curve shows a **normal distribution** of data. It is also sometimes called a Gaussian distribution. A normal distribution is a bell-shaped distribution of data where the mean, median and mode all coincide. A frequency polygon showing a normal distribution would look like this:



Notice:

- the frequency polygon has been smoothed out to give a 'rounded' curve
- the value under the highest point on the curve shows the mean, median and mode value
- the curve is symmetrical
- the curve appears to touch the horizontal axis but in fact it never does – the horizontal axis is an asymptote to the curve



In the above graphs, the most common measurement, 9, is the same in curves **A** and **B** but there is a greater range of values for **A** than for **B**. Curve **C** has the same distribution as **A** but the most common measurement is 18 which is twice that of curve **A**. All of these distributions are normal.

The normal distribution is one of the most important of all distributions because it describes the situation in which extremes of values (i.e. very large or very small values) seldom occur and most values are clustered around the mean. This means that normally distributed data is predictable and deductions about the data can easily be made. A great many distributions that occur naturally are normal distributions.

Examples of data that will give a normal distribution are:

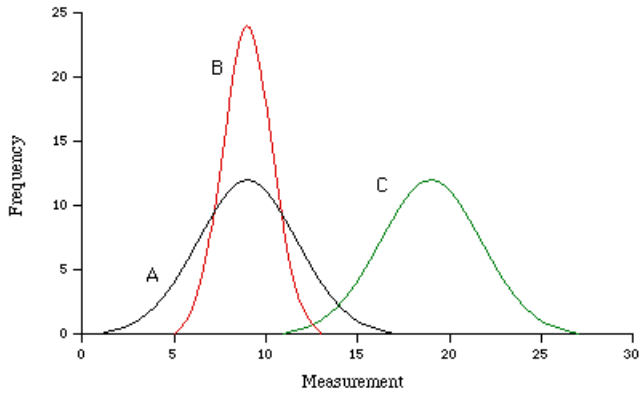
a) Heights and weights (very tall or short people or very fat or thin people are not common)
b) Time taken for professional athletes to run 100m
c) The precise volume of liquid in cool-drink bottles
d) Measures of reading ability
e) Measures of job satisfaction
f) The number of peas in a pod.

There are, however, many variables which are not normally distributed, but they are not 'abnormal' either! For this reason, the normal distribution is often referred to as the Gaussian Distribution, named after the German mathematician Gauss (1777-1855).

FEATURES OF NORMALLY DISTRIBUTED DATA:

Look again at this diagram showing different normal curves

Notice how the spread of the data is reflected by the width of the normal curve.



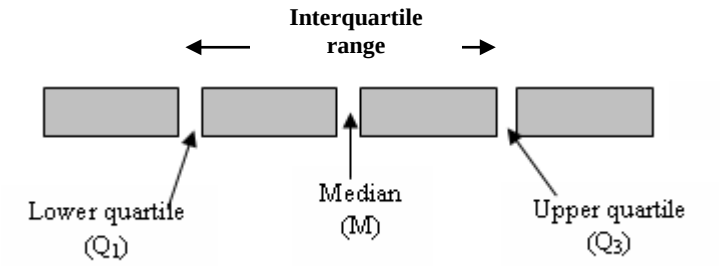
1) The median and the interquartile range

- The interquartile range is one of the measures of dispersion (spread) of a set of data.
- The **median** divides the distribution of a data set into two halves.

Each half can them be divided in half again;

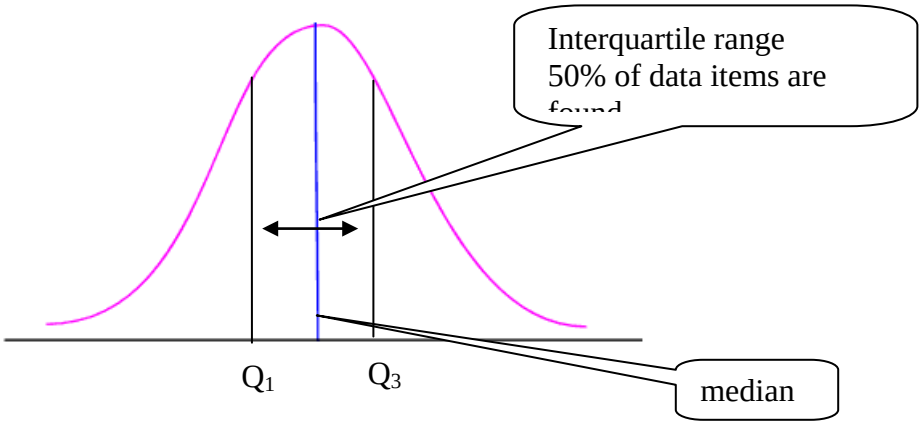
- the **lower quartile (Q_1)** is the median of the first half of the data set
- the **upper quartile (Q_3)** is the median of the second half of the data set.

The set of data is divided into 4 equal parts:



- The lower quartile (Q_1) is a quarter of the way through the distribution,
- The middle quartile which is the same as the median (M) is midway through the distribution.
- The upper quartile (Q_3) is three quarters of the way through the distribution.
- The **interquartile range** is where 50% of the data items lie.

The interquartile range can be seen in a normal distribution approximately like this:



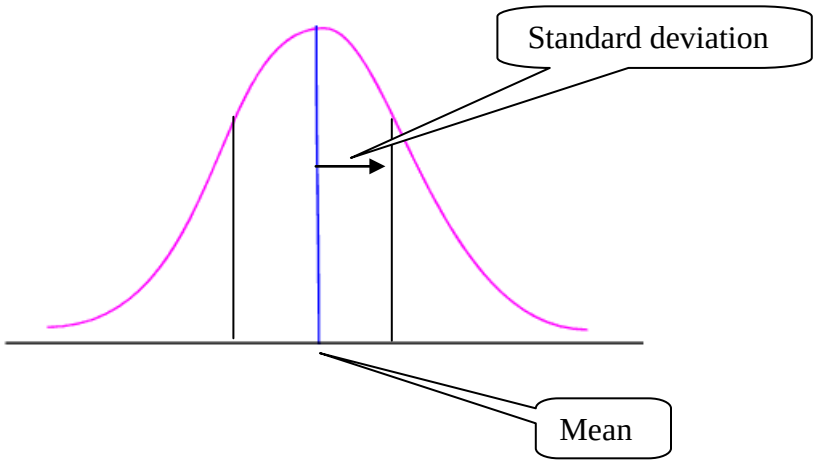
2) The mean and standard deviation

A more useful measure of spread associated with the normal distribution is the **standard deviation**.

- We use the formula **standard deviation** = σ = $\sqrt{\textbf{variance}}$ = $\sqrt{\frac{\sum(x-\bar{x})^2}{n}}$, where x is the value of the data item, $\bar{x}$ is the value of the mean, and n is the number of data items.

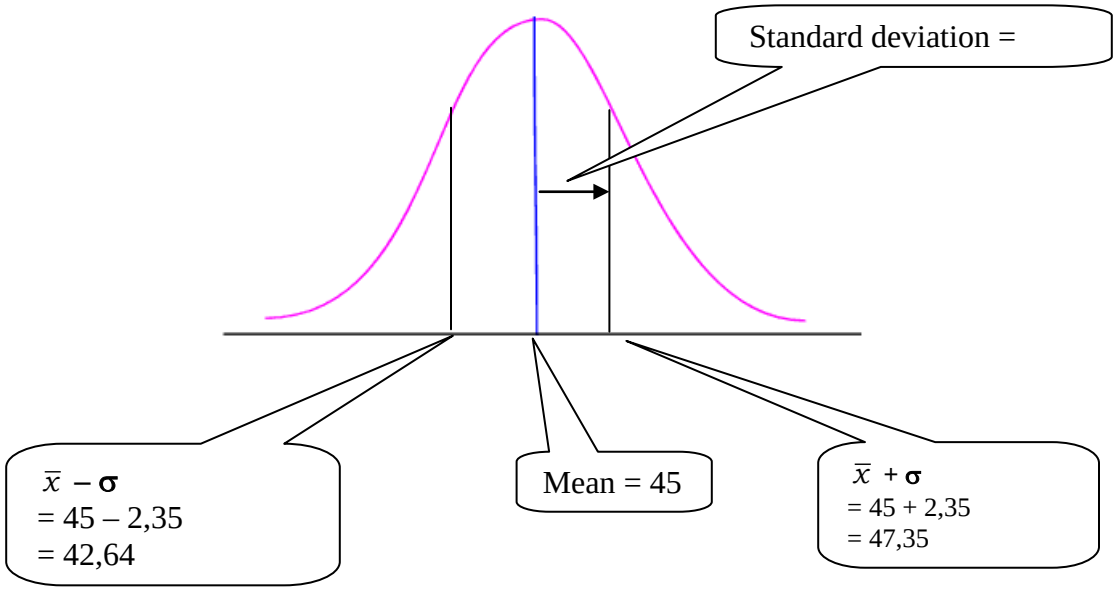
- When we have data listed in a frequency table, we use the formula **standard deviation** = σ = $\sqrt{\textbf{variance}}$ = $\sqrt{\frac{\sum f.(X-\bar{X})^2}{n}}$, where X is the value of the data item, $\bar{X}$ is the value of the mean, f is the frequency of the data item and n is the number of data items.

The standard deviation tells you the average **difference** between data items and the mean.



Example:
Suppose the mean of a data set $\bar{x}$ = 45 and the standard deviation = σ = 2,35. This means that the average difference between most of the data items and the mean = 2,35.

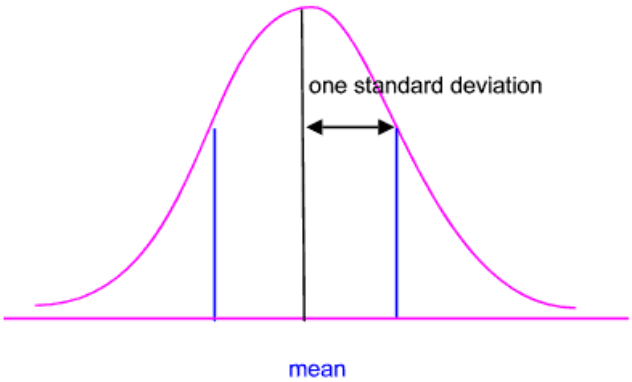
In other words most of the data lies with the values 45 – 2,35 = 42,64 and
45 + 2,35 = 47,35



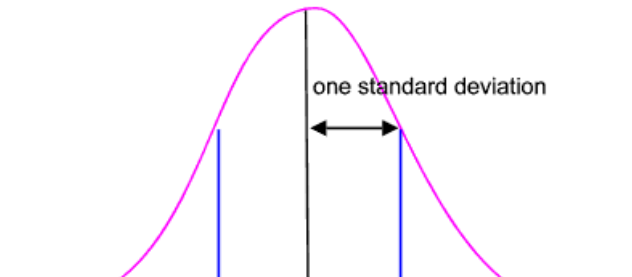
- Most of the data lies within 1 standard deviation of the mean i.e. within the data range $\bar{x} - \sigma$ and $\bar{x} + \sigma$

3) One Standard deviation

The spread on any normal curve may be large or small but in every case, most of the data falls within 1 standard deviation of the mean.

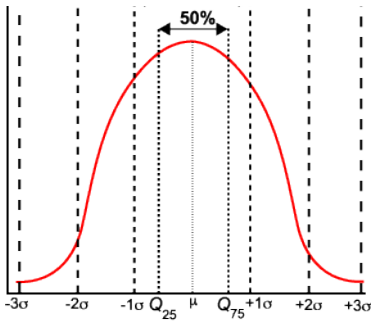


This means that most of the data values on the horizontal



axis lie within 1 standard deviation of the mean

The interquartile range and the standard deviations give a composite description of the normal curve as shown in this diagram:



4) The shape of the normal curve

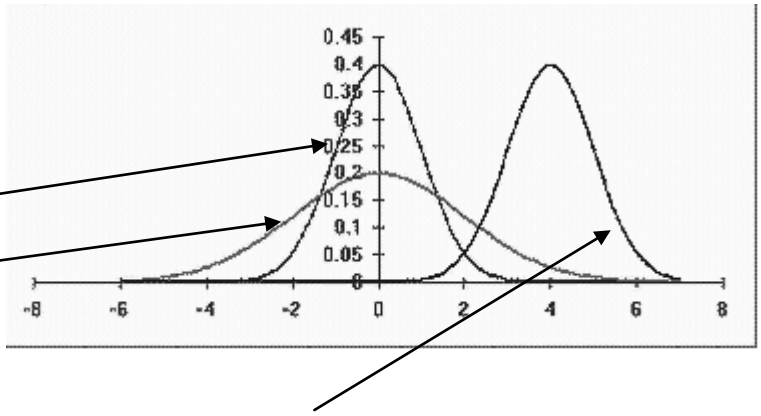
On a normal distribution, the mean and the standard deviation are important in determining the shape of the normal curve.

This diagram shows 3 different normal distributions.

Graph 1 has a mean = 0 and standard deviation = 1.

Graph 2 has mean = 0, but standard deviation = 2.

Notice how graph 2 is flatter and more stretched out than graph 1. The data items have a greater spread.



Graph 3 has the same standard deviation (1) as graph 1, and the mean = 4, so the curve is shifted along the horizontal axis.

Activity 1:

The office manager of a small office wants to get an idea of the number of phone calls made by the people working in the office during a typical day in one week in June. The number of calls on each day of the (5-day) week is recorded. They are as follows:

Monday – 15; Tuesday – 23; Wednesday – 19; Thursday – 31; Friday – 22

1) Calculate the mean number of phone calls made

2) Calculate the standard deviation (correct to 1 decimal place).

3) Calculate 1 Standard Deviation from the mean:

$\bar{x} \pm \sigma =$ .....or .....

The interval of values is $(\bar{x} - \sigma \; ; \bar{x} + \sigma)$ = ...

4) On how many days is the number of calls within one Standard Deviation of the mean?

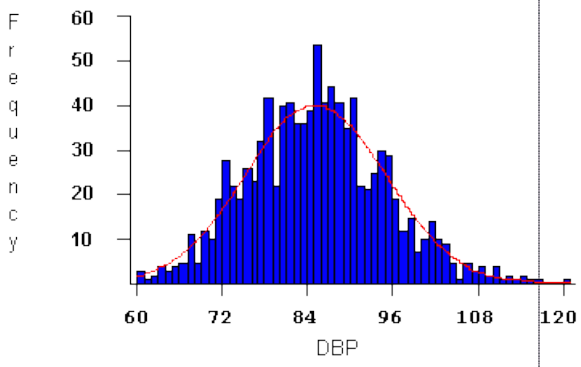
Number of days =

Percentage of days =

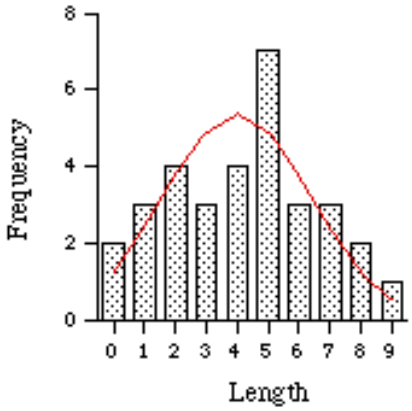
Therefore the phone calls on .....% of the days lies within 1 Standard Deviation of the mean.

5) Histograms and normal curves

Example 1: In research at a hospital the blood pressure of 1 000 recovering patients was taken The distribution of blood pressure was found to be approximated as a normal distribution with mean of 85 mm. and a standard deviation of 20 mm. The histogram of the observations and the normal curve is shown below.



Example 2: In ecological research the antennae lengths of wood lice indicates the ecological well being of natural forest land. The antennae lengths of 32 wood lice, with mean = 4mm and standard deviation of 2,37 mm, can be approximated to a normal distribution.



Activity 2:

1) The pilot study for the Census@School project gave the following data for the heights of 7 068 pupils from Grade 3 to Grade 11.

Height Less than (cm)	**Total number of Pupils (Cumulative frequency)**
106,11	8
121,54	119
136,97	1 233
152,40	3 441
167,83	5 854
183,26	6 959
198,69	7 067

The mean $\bar{x}$ = 152,4 cm and the Standard Deviation σ = 15,43 cm

a)

i) Calculate $\bar{x} + \sigma$
ii) Calculate $\bar{x} - \sigma$
iii) What is the total number of learners whose heights are less than $\bar{x} + \sigma$?
iv) What is the total number of learners whose heights are less than $\bar{x} - \sigma$?
v) Calculate the number of learners whose heights are within 1 standard deviation of the mean.
vi) Write this number as a percentage of the total number of learners

b)

i) Calculate the number of learners whose heights are within 2 standard deviations of the mean i.e. within the interval $(\bar{x} - 2\sigma \; ; \bar{x} + 2\sigma)$
ii) Write this as a percentage

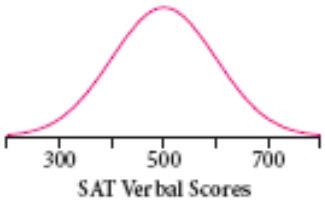
c)

i) Calculate the number of learners whose heights are within 3 standard deviations of the mean i.e. within the interval $(\bar{x} - 3\sigma \; ; \bar{x} + 3\sigma)$

ii) Write this as a percentage

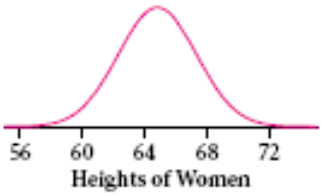
2) For each of the normal distributions below,
   a) estimate the mean and the standard deviation visually.
   b) Use you estimation to write a summary in the form "a typical score is roughly …. (mean), give or take….(standard deviation)"
   c) Check to see that this interval contains roughly 66% of the data items

   i) Verbal scores for SAT tests
   (The SAT test is the standardized test for college admissions in the USA)



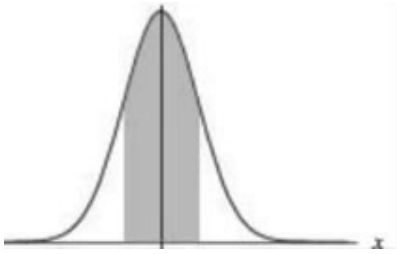
   ii) ACT (The ACT is the college-entrance achievement test in the USA)

   iii) Heights of women attending first year university

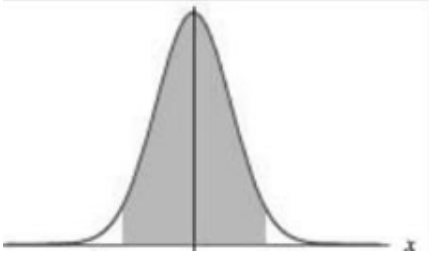


6) A general normal distribution

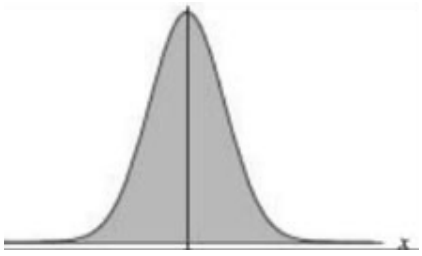
In question 1 in the above activity you worked out the percentage of data items within 1,2 and 3 standard deviations. This can be summarised in the diagrams below:



68% of the population falls within 1 standard deviation of the mean.



95% of the population falls within 2 standard deviations of the mean.



99,7% of the population falls within 3 standard deviations of the mean.

Activity 3:

Taken from Classroom Maths Grade 12

The arm lengths of 500 females and 500 males were measured. Measurements were taken from shoulder to fingertips when the arm was held out at shoulder height. The results were summarised in a table as follows:

Arm length (mm)	**Number of females**	**Number of males**	**Number of adults (total of females and males)**
620 –	3	0	
640 –	11	0	
660 –	41	0	
680 –	92	0	
700 –	132	2	
720 –	120	9	
740 –	69	27	
760 –	25	71	
780 –	6	114	
800 –	1	122	
820 –	0	89	
840 –	0	46	
860 –	0	15	
880 –	0	4	
900 –	0	1	
TOTALS	500	500	

Work with the members of your group.

- Person 1 should work with the female data
- Person 2 should work with the male data
- Person 3 should work with the adult data (by first finding the sum of the female and male data).

1) Fill in only your set of data on the table on the next page.

2) Work with your set of grouped data and calculate
   a) The mean
   b) The median
   c) The standard deviation

3) a) Do the mean and median of your set of data have approximately the same value?

b) Does approximately 99,7% of the data lie within three standard deviations of the mean?

4) Draw a histogram to illustrate your data.

5) Compare the histograms and comment on similarities and differences. Which set of data, if any, is normally distributed?

Solution to Activity 3

1)

Arm length (mm)			
620 –			
640 –			
660 –			
680 –			
700 –			
720 –			
740 –			
760 –			
780 –			
800 –			
820 –			
840 –			
860 –			
880 –			
900 –			
TOTALS			

2)

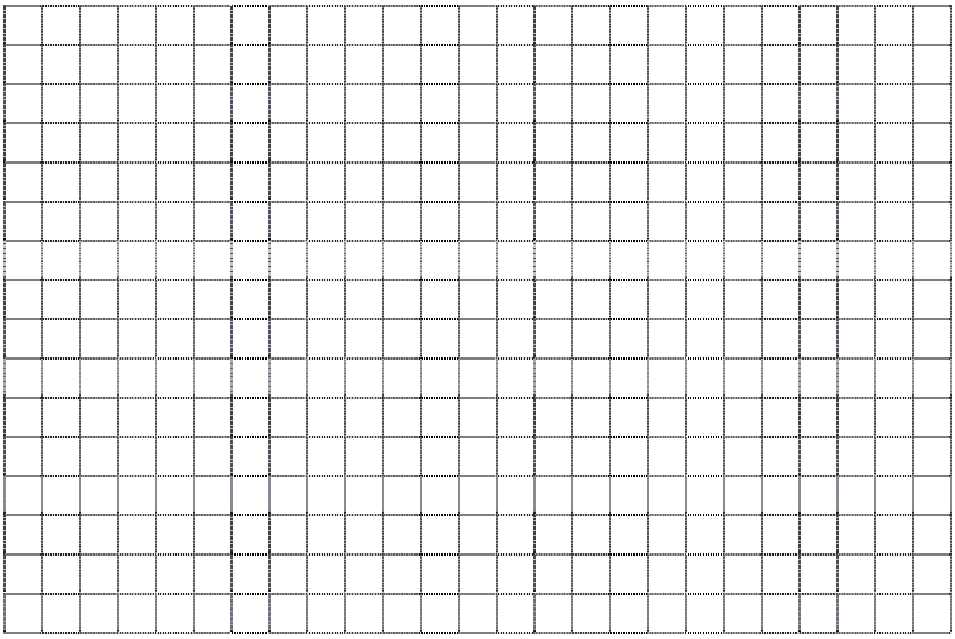
a) The mean
b) The median
c) The standard deviation

3)

a)

b)

4) Histogram of data



5) Comparison of the histograms:

Similarities

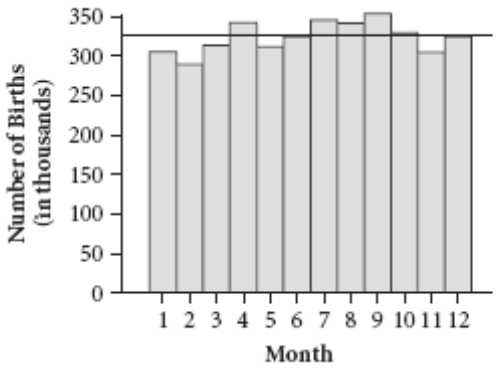
Differences

Normally distributed?

DISTRIBUTIONS of DATA

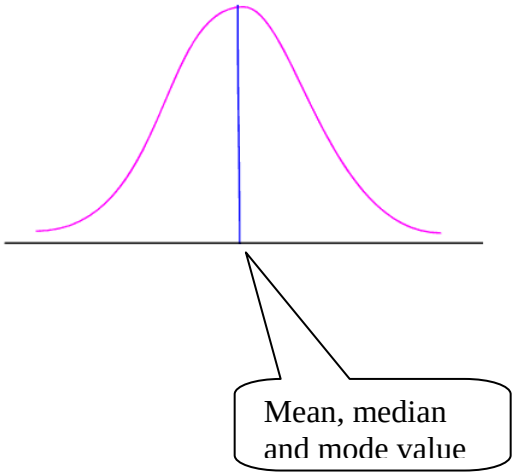
1) Data could be distributed **uniformly.** A uniform distribution shows a rectangular shape. Each data item has the same likelihood of occurring.

   e.g. the histogram shows the births in one year in Nigeria in 1997. There is little change from month to month. We can say that 'the distribution of births is roughly uniform.'

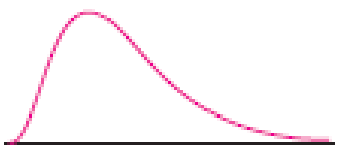


2) Data could be **normally** distributed. A normal distribution shows a symmetrical shape

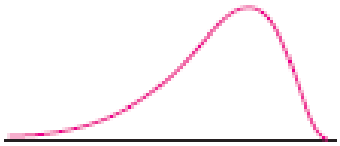
   A normal curve:
   - Is bell-shaped
   - Is symmetrical
   - Shows the mean, median and mode value under the highest point on the curve.
   - Appears to touch the horizontal axis but in fact it never does. The horizontal axis is an asymptote to the curve



3) Data could be **skewed.** Distributions are not symmetric or uniform; they show bunching to one end and/or a long tail at the other end. The direction of the tail tells whether the distribution is **skewed right** (a long tail towards high values) or **skewed left** (a long tail towards low values).



Skewed right



Skewed left

Activity 4:

- Work with the rest of the members of your group to answer the following:

1) Sketch the shape of the distribution you would expect from the following:

a) the height of all learners in Grade 10 in your school	
b) the height of the riders in the Durban-July horse race	
c) Grade 12 exam results at your school	

2) Describe each of the following distributions as skewed left, skewed right, approximately normal or uniform.

a) The incomes of the richest 100 people in the world	

b) The length of time the learners in your class took to complete a 40 minutes class test	
c) The age of people who died in South Africa last year	
d) IQs of a large sample of people chosen at random.	
e) Salaries of employees at a large corporation.	
f) The marks of learners on an easy examination.	
3) Sketch the following distributions:	
a) A uniform distribution showing the data you would get from tossing a fair dice 1 000 times	
b) A roughly normal distribution with mean 15 and standard deviation 5	
c) A distribution that is skewed left, with a median of 15 and the middle 50% of its values lying between 5 and 20	
d) A distribution skewed right with a median of 100 and an interquartile range of 200	

SKEWED DISTRIBUTIONS

An **outlier** is an unusual data item that stands apart from the rest of the distribution. Sometimes outliers are mistakes; sometimes they are values that are unexpected – for whatever reason (e.g. an extremely tall boy in the Grade 10 class); and sometimes they are an indication of unusual behaviour within the set of data.

Skewed data is sometimes described as **positively or negatively skewed** as shown in the diagram below.

Negatively skewed

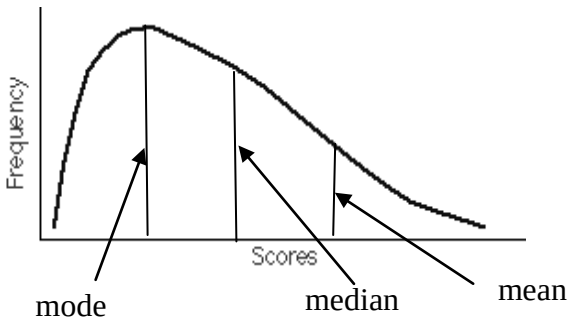
Symmetrical

Positively skewed

1) Positively skewed

When the peak is displaced to the left of the centre, the distribution is described as being **positively skewed.**

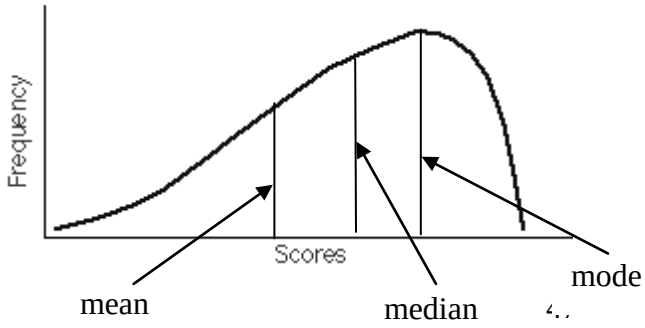
The distribution is said to be **skewed to the right.** It illustrates that there are a few very high values in the set of data. Since there are only a few high numbers, in general the mean is higher than the median.



2) Negatively skewed

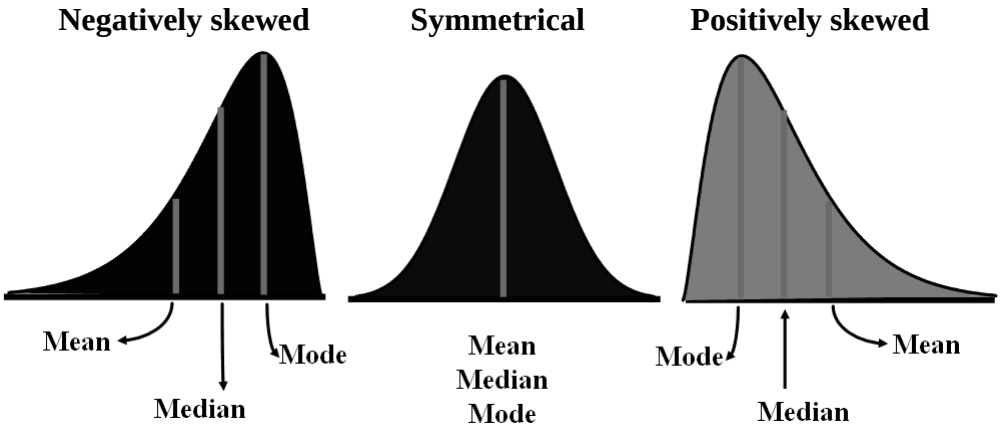
When the peak is displaced to the right of the centre, the distribution is described as being **negatively skewed**.

The distribution is said to be



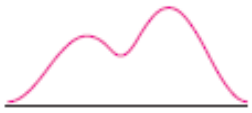
skewed to the left. It shows a data set containing a few numbers that are much lower than most of the other numbers. In general, the mean is lower than the median.

In all distribution curves the **mode** is the highest point of the curve. (Remember the highest point of the curve is the midpoint of the bar with the highest frequency).



Note:

- If a data set is approximately **symmetrical**, then the values of the mean and the median will be almost equal. These values will be close to the mode, if there is one. (i.e. mean – median ≈ 0)
- In **positively skewed data** (i.e. it is not symmetrical and there is a long tail of high values) the mean is usually greater than the mode or the median.
  (i.e. mean – median > 0)
- In **negatively skewed data** (there is a long tail of low values) the mean is likely to be the lowest of the averages. (i.e. mean – median < 0)
- Sometimes the distribution curve migl
  Have two peaks showing bimodal data



MEASURES OF SKEWNESS

In statistics there are a number of measures of skewness (how skewed) the data is. The simplest is **Pearson's coefficient of skewness**. There are two simple equations depending on whether you know the median or the mode of the set of data.

1) If you know the mode: $\mathbf{S} = \dfrac{\text{mean - mode}}{\text{standard deviation}}$

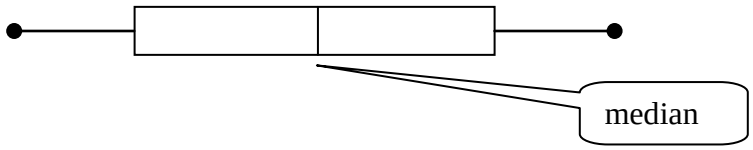
2) If you do not know the mode or there is more than one mode:
$\mathbf{S} = \dfrac{3(\text{mean - median})}{\text{standard deviation}}$

- If **S** is very close to 0, the data set is symmetrical
- If **S** > 0, then the data is skewed right, or is **positively skewed**.
- If **S** < 0, then the data is skewed left, or is **negatively skewed**.

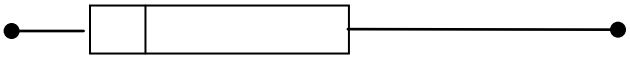
SKEWNESS AND BOX AND WHISKER PLOTS

The 'centre' and/or the spread of skewed distributions are not as clear-cut as in normal data. To make the distribution easier to understand, **quartiles** are usually used to describe the spread of skewed data. The median is the measure of the 'centre' of the distribution and the quartiles indicate the limits of the middle 50% of the data. Box and whisker plots are useful representations of data showing the spread around the median.

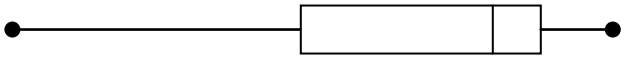
- In a **symmetrical** set of data the box and whisker plot is symmetrical about the median



- In data that is **positively skewed** the data has a long tail of items of very high value. This means the median is to the **left** of the box and there is a long whisker of high values to the right.

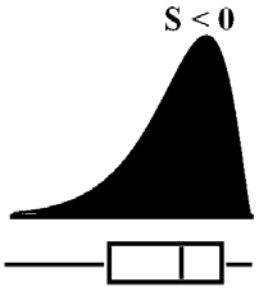


- In data that is **negatively skewed** the data has a long tail of items of very low value. This means the median is to the **right** of the box and there is a long whisker of high values to the left.

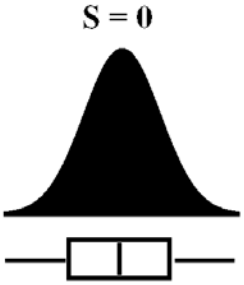


Any box and whisker plot can be superimposed on a frequency polygon to show skewness like this:

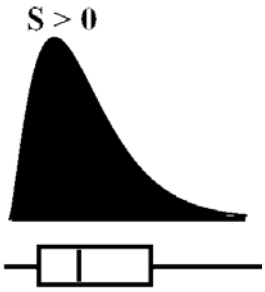
Negatively skewed



Symmetrical



Positively skewed



Activity 5:

1) The following data set represents the ages, to the nearest year, of 27 university students in a statistics class.

17	21	23	19	27	18	20	21	28	31
18	21	24	30	25	19	22	27	35	18
29	22	20	30	28	21	23			

a) Determine the mean, median and mode for the data set.

b) Determine the standard deviation of the data.

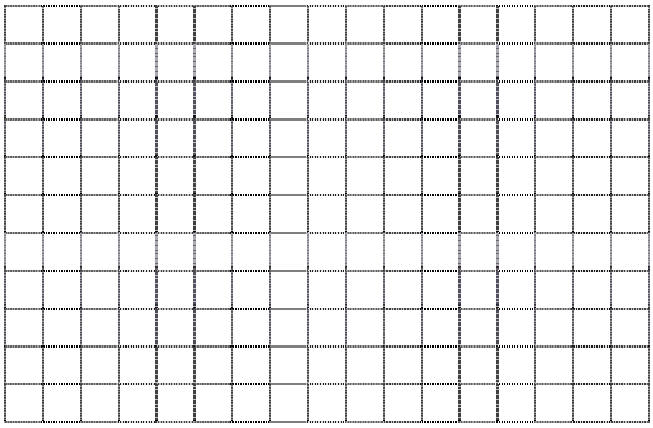
c) Determine Pearson's coefficient of skewness for the data. Is the data positively skewed, negatively skewed or symmetrical?

d) Determine the five-number summary and then draw a box and whisker diagram for the data. Does the diagram reflect your answer in (c) above?

Activity 5 (continued)

e) Using five equal class intervals construct a frequency table for this data.

f) Draw a frequency polygon to illustrate the data.



g) Describe the shape of the frequency polygon.

h) What relationship would you expect to find between the location of the median and the location of the mean? Why?

i) On the graph show the approximate positions of the mean, median and the mode.

Activity 5 (continued)

2) After an oil spill off the Cape coast, local beaches are checked for oiled water birds. To simplify the collection of the data, the beaches are divided into 100 m stretches and the number of oiled birds recorded separately for each stretch. Fifty of the recorded counts are summarised below:

0	1	5	2	19	47	21	8	7	4
0	0	0	0	0	0	1	3	15	11
2	2	0	0	0	0	0	1	0	0
1	4	6	6	0	1	2	2	0	0
7	0	0	3	1	1	0	1	4	0

a) Determine the mean, median, mode and standard deviation of the data

b) Determine the Pearson's coefficient of skewness using the mode

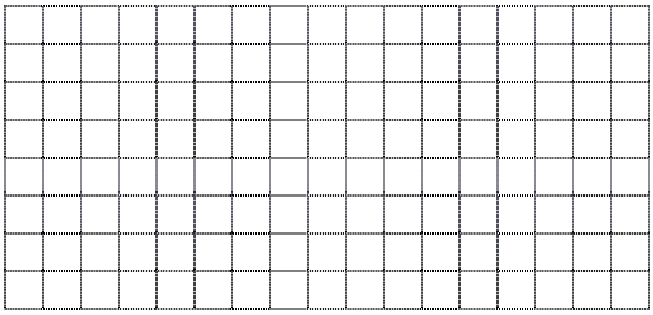
c) Determine the Pearson's coefficient of skewness using the median

d) Is the data negatively or positively skewed?

Activity 5 (continued)

e) Using five equal class intervals, draw up a frequency table for this data

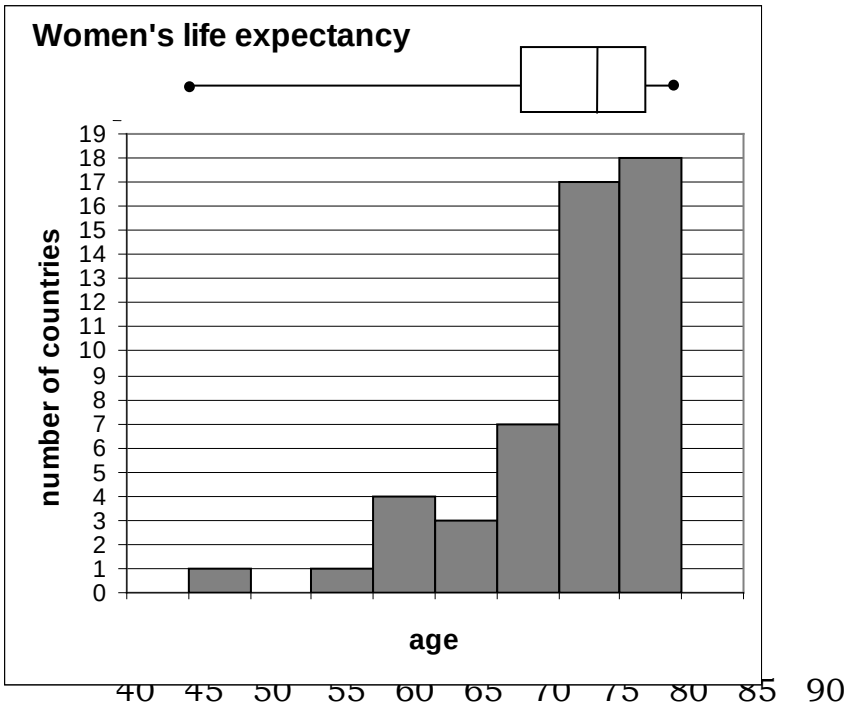
f) Draw a frequency polygon to represent this data visually.



g) Does the diagram confirm the skewness you calculated in (b) and (c)?

Activity 5 (continued)

3) This box and whisker diagram and histogram illustrate the life expectancy in 1999 of women in several countries in the world.



[Murdock, J. et al. (2002) Discovering Algebra, Key Curriculum Press, page 61]

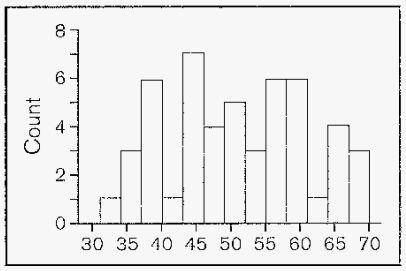
a) Describe the skewness of the data

b) The right whisker of the box plot is very short. What does this tell you about the life expectancy of women?

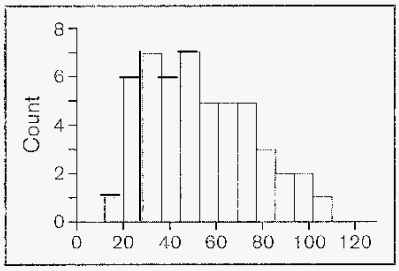
Activity 6:

The following activity is taken from Scheaffer et al (2004) *Activity based Statistics* (2nd edition) Key College Publishing, New York

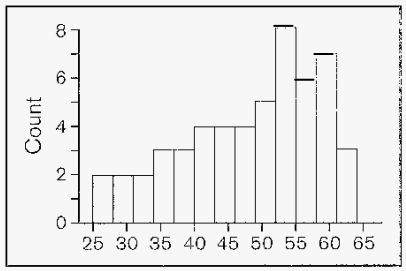
1) Consider the following histograms and the table of summary statistics. Each of the variables 1 – 6 (in the table) correspond to one of the histograms. Match the histograms and the variable in the last column of the table.



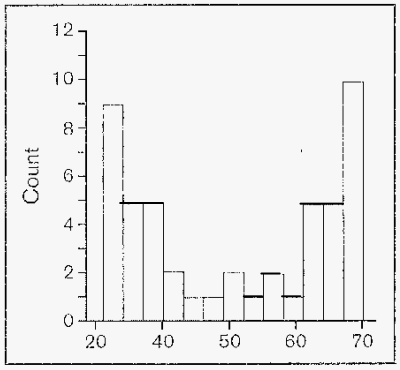
Histogram a



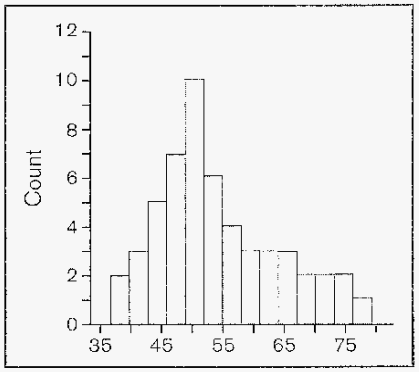
Histogram b



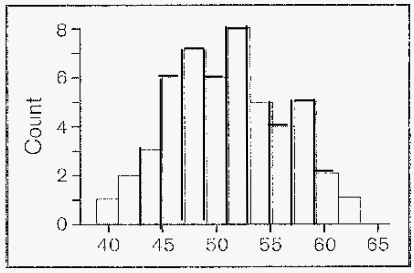
Histogram c



Histogram d



Histogram e

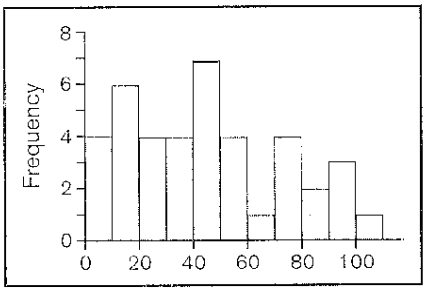


Histogram f

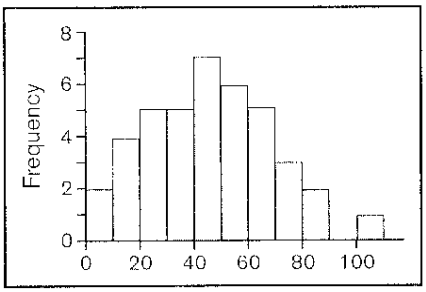
variable	mean	median	Standard deviation	Histogram number
1	60	50	10	
2	50	50	15	
3	53	50	10	

4	53	50	20	
5	47	50	10	
6	50	50	5	

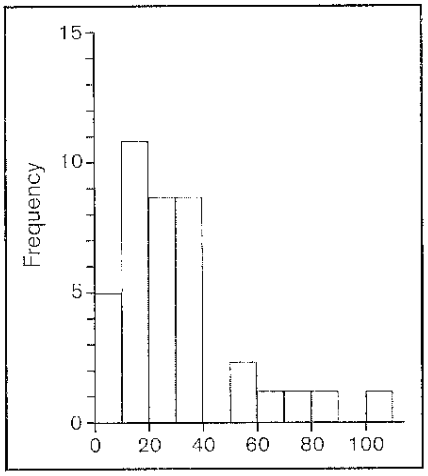
2) Each box plot corresponds to one of the histograms. Match the histograms and the box plots and explain why you made the choice you did.



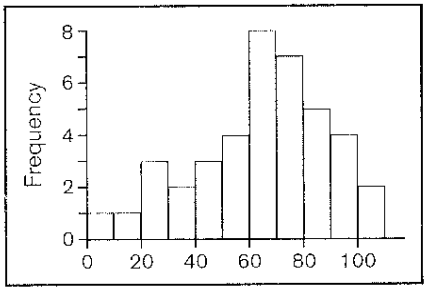
Histogram a



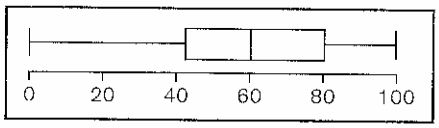
Histogram b



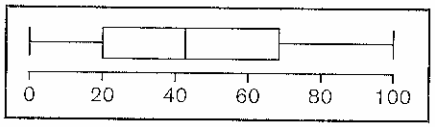
Histogram c



Histogram d



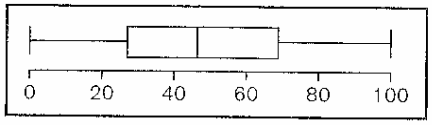
Box plot 1



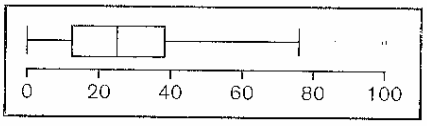
Box plot 2

Histogram: …… Histogram: …….

Reason:



Box plot 3



Box plot 4

Histogram: ……. Histogram: …….

Reason: