SECTION 1: STATISTICS

1.1. Revision of grade 11:

- The **five-number summary** is written in the following order: minimum, lower quartile, median, third quartile, maximum. These numbers are usually represented by a **box-whiskers diagram**.
- *Position* of median = $\frac{n+1}{2}$
  Position of UQ = $\frac{3(n+1)}{4}$
  Position of LQ = $\frac{n+1}{4}$
  Except for ogives - just use n as the numerator

  ∴ These do not calculate the final answer. The actual median, UQ and LQ values lie **at** these positions. These values must be given at the answer.

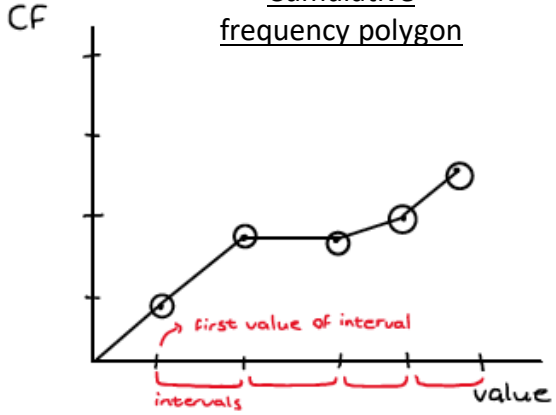
  Note: If asked for the "median interval" – This is the interval that the position calculated by $\frac{n+1}{2}$ lies within.

  The **mean** is the average value of a data set. It is calculated by: $\bar{x} = \frac{\sum fx}{n}$ (Sum of the data values in a set / Number of data values in a set)

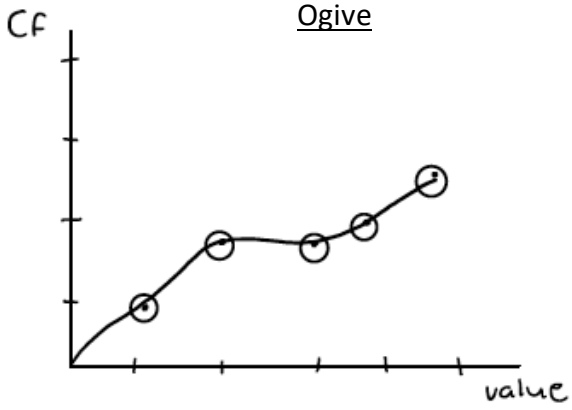
- The **standard deviation** measures the concentration of data values around the mean i.e. the average distance between data set values and the mean. It is calculated by: $\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}$ **Note:** this formula is given as σ^2 (variance) on the formula sheet

- The above two formulas can also be calculated with your calculator in MODE, 3: STAT. You then select option 1: 1 – VAR and enter your data set. Then press AC, SHIFT, and 1 and option 4: VAR. Then you can choose which one of the above values you want to be calculated. **Note:** σ will be represented as σx or σn depending on which model of calculator used.
- An **ogive** or **cumulative frequency polygon** for data values given in intervals is drawn as follows:
  1. Plot points with an x-value as the *beginning* of the interval and with a y-value as the *cumulative frequency* (frequency of interval and all lower intervals added together) of that interval
  2. Connecting these points for every interval with straight lines gives a cumulative frequency polygon. Connecting them with a smooth curve gives an ogive.

Cumulative frequency polygon

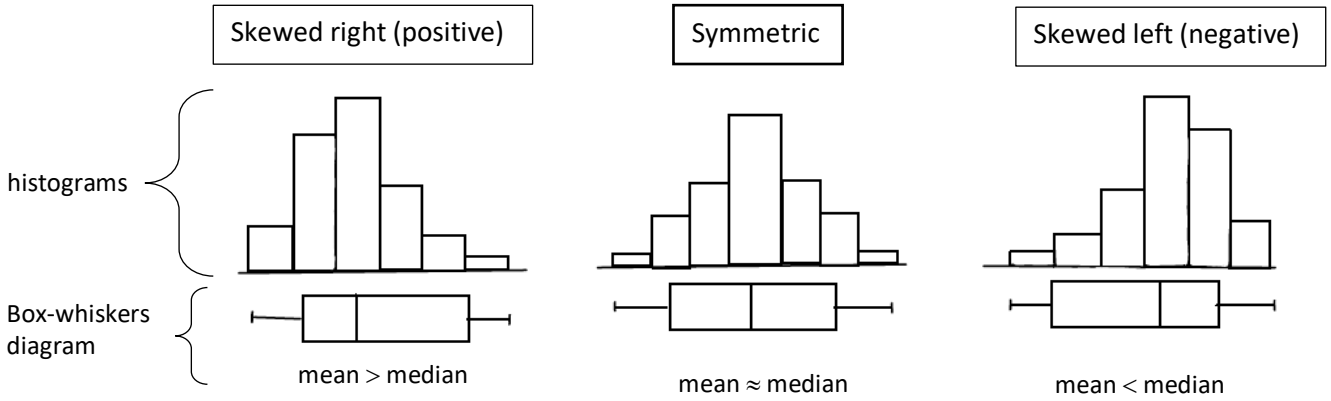


Ogive



- H**istograms** consist of blocks drawn without spaces in between them to represent the frequency (number of units) in each interval of data values. The interval is written below each block on the x-axis and the frequency is represented on the y-axis. **Note:** when calculating the mean of the data from a histogram, fx is the midpoints of each interval and n is the frequency of each interval added together.
- **Note:** when converting between ogives and histograms – the frequency of each interval can be found by subtracting the cumulative frequency of the interval just below the desired interval from the cumulative frequency of the desired interval.

- **Symmetric and skewed data:**



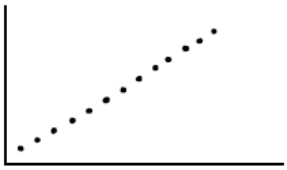
1.2. The line of regression or line of best fit:

- $\hat{y} = a + bx$
  This line is also known as the l**east squares regression** line
- **r** is the correlation coefficient between points of data i.e. how closely data points are situated to the regression line

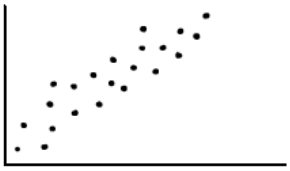
  Properties of the correlation coefficient **$-1 \leq r \leq 1$**:

r=1	Perfect positive linear correlation
$1 < r \leq 0.7$	Strong positive linear correlation
$0.3 \leq r < 0.7$	Weak positive linear correlation
$-0.3 < r < 0.3$	No significant correlation
r = 0	No correlation
$-0.7 \leq r < -0.3$	Weak negative linear correlation
$-1 < r < -0.7$	Strong negative linear correlation
r = -1	Perfect negative linear correlation

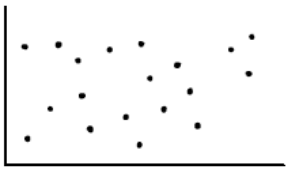
Some scatterplots showing the properties above:
(note: these do not have to start at zero)



Perfect (maximum) positive correlation



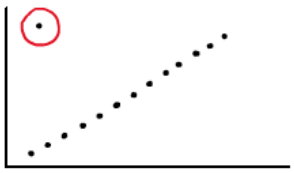
Strong positive correlation



Zero correlation

Perfect (maximum) negative correlation

Weak negative correlation



Strong positive correlation with outlier

1.3. How to find a, b and r:

Finding these values is quite a simple process and makes use of your calculator. You will be given either a scatterplot or a table with x and y values. You then click **MODE 3: STAT** on your calculator and then option **2: A+BX**. You then plug in your x and y values as given to you (note: x value in the calculator must be the independent variable of the data). You then press **AC,** press **SHIFT** then **1** and finally select option **5: REG.** You can then select the A, B or r option depending on which value you want to be given (note: if you want to find all 3 you don't have to do the whole process again, you can just press **AC** again and go from there).

If you are asked to now use your regression line to predict x or y values:

Interpolation: the x-value being used to predict a y-value or the x-value that is predicted by a given y-value is within the given data values (domain). This means that the prediction is accurate and valid.
Extrapolation: the x-value being used to predict a y-value or the x-value that is predicted by a given y-value is outside the given data values (domain). This means that the prediction is inaccurate and not as valid (or not valid at all).

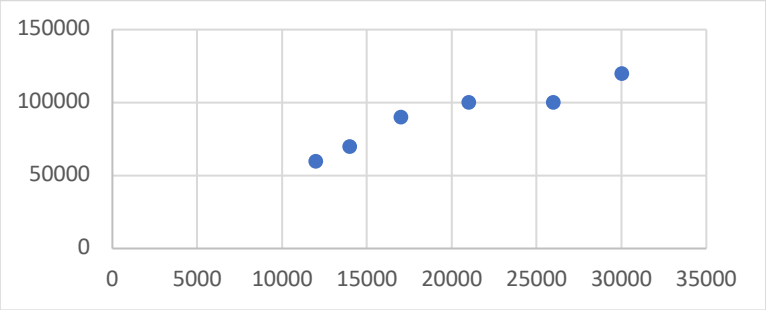
1.4. Example:

The owner of a travel company collected the following table of data which shows the relationship between the company's annual profit and its annual expenses (in rand):

Annual expenditure (x)	12 000	14 000	17 000	21 000	26 000	30 000

Annual profit (y)	60 000	70 000	90 000	100 000	100 000	120 000

1.) Draw a scatterplot for the given data.

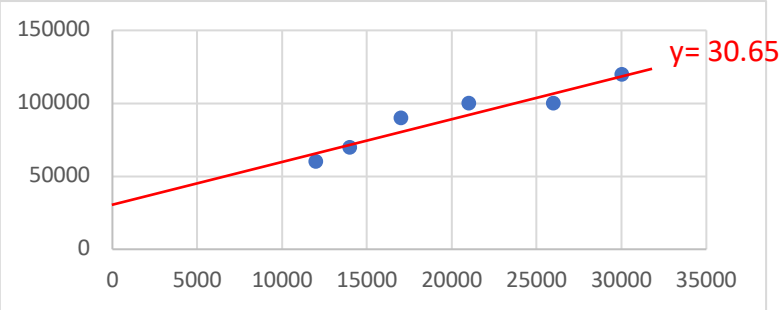


2.) Calculate the equation of the least squares line for the data

y = A + Bx
y= 30.65 + 2.97x

(using calculator as explained above)

3.) Draw the least squares line for the data



4.) **Predict the annual profit if the annual advertising expenditure is R25 000**

Use regression line: y = 30.65 +2.97(25 000)
= 74 280.65 rand

5.) Calculate the correlation coefficient

r = 0.95

(using calculator as explained above)

6.) What conclusion can you reach about the strength of the relationship between the annual profit and the annual expenditure?

$1 < r \leq 0.7$ therefore the variables have a strong positive linear correlation