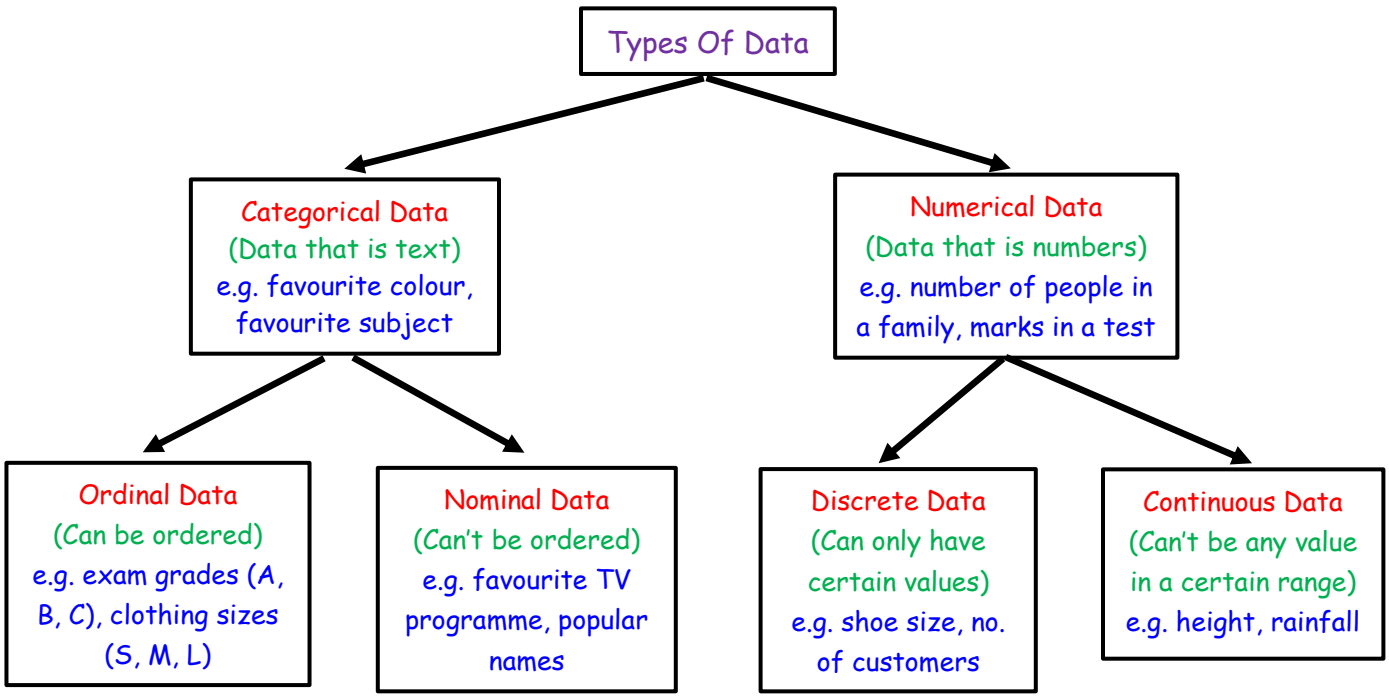
➢ Section 1A: Revision of Terminology from Junior Cycle



➢ Section 1B: More terminology from Junior Cycle and some new ones

1) Primary Data: Data that is collected by the person who's going to use it. E.g. Carry out a survey or a questionnaire
2) Secondary Data: Data that has been collected by someone else. E.g. Newspapers, Internet or census
3) Variable: The characteristic being recorded. E.g. if studying the size of households, then the variable is the number of people in each house
4) Univariate Data: When we study one variable at a time.
5) Bivariate Data: Comparing two variables together.
6) Observational Study: The researcher collects the information of interest without influencing events. E.g. a study of students favourite subjects where data is collected using a questionnaire.
7) Case-Control Study: Two groups are studied – one called the control and one called the cases. E.g. A medical study into the effectiveness of a new drug on a particular illness, where one group has the illness, and the other group doesn't have it.
8) Designed Experiment: Some treatment is applied to a group of subjects to see the effects of the treatment on the subjects.
9) Explanatory Variable: In a designed experiment, the explanatory variable is the treatment being applied. E.g. the drug being tested in a medical trial
10) Response Variable: In a designed experiment, the response variable is the effect of the treatment. E.g. the effect of the drug on the patient
11) Informed Consent: Subjects being studied must be told in advance of the nature of the study and any potential risks
12) Bias: If a sample isn't selected randomly then the sample is biased. E.g. surveying people coming out of the soccer match to get their opinions on supporting their local soccer team
13) Outlier: A data value that is way off other data values.
14) Descriptive Statistics: When large amounts of data is summarized or presented in graphs or charts.
15) Inferential Statistics: Trying to predict an outcome of an event from the responses of a smaller sample of the population.

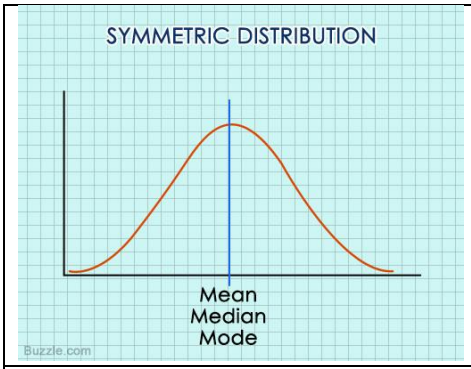
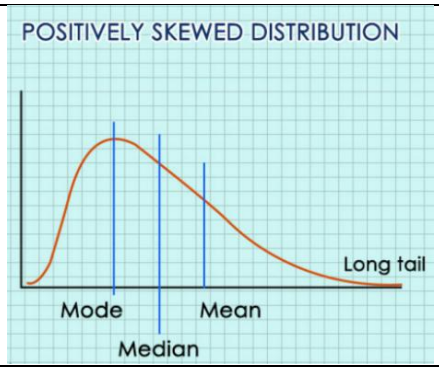
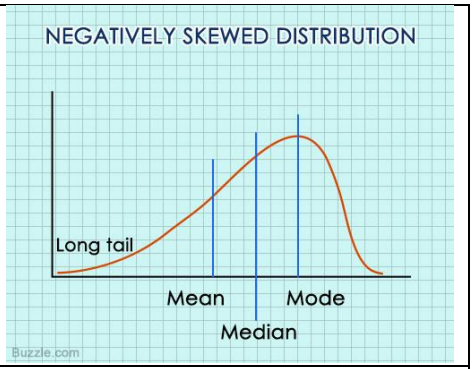
➢ Section 1C: Revision of methods of gathering data and questionnaires from Junior Cycle

Survey Type	Advantages	Disadvantages
Telephone Interview	- Questions can be explained - Easy enough to make random, using the telephone directory	- Not anonymous so responses may not be honest - Expensive compared to other types
Postal Questionnaire	- Cheap - Easy enough to make random, using the telephone directory	- Questions can't be explained - Not anonymous so responses may not be truthful - People don't always reply
Online Questionnaire	- Cheap - Responses can be anonymous which ensures more honesty	- Questions can't be explained to the respondent - People don't always reply - Sample is biased as only people online are surveyed
Face-to-face Interview	- Questions can be explained to the respondent	- Not anonymous so responses may not be honest - Expensive - Not random
Observation	- Cheap - Easy to carry out	- Not always suitable for some surveys - Questions can't be explained to the respondent

- Tips for designing questionnaires:

- Use clear and simple language - Avoid personal questions - Start with simpler questions at the start	- Allow for all possible responses - Be clear where answers should be recorded - No leading questions

➢ Section 1D: Describing Distributions

		
Example: If we surveyed the height of TY students in the country it would almost certainly be approximately symmetrical	**Example:** If we surveyed the ages at which people learn to drive in the country, it would probably be positively skewed as most people learn how to drive when they are young	**Example:** If we surveyed the heights of players in the NBA, it would almost certainly be negatively skewed as the majority of them are very tall

➢ Section 1E: Terminology used in Sample Surveys

1) Population: The entire group being studied.
2) Sample: A group that is selected from the population to represent the population.
3) Census: A survey of the whole population.
4) Parameter: A numerical measurement describing some characteristic of a population.
5) Statistic: A numerical measurement describing some characteristic of a sample.

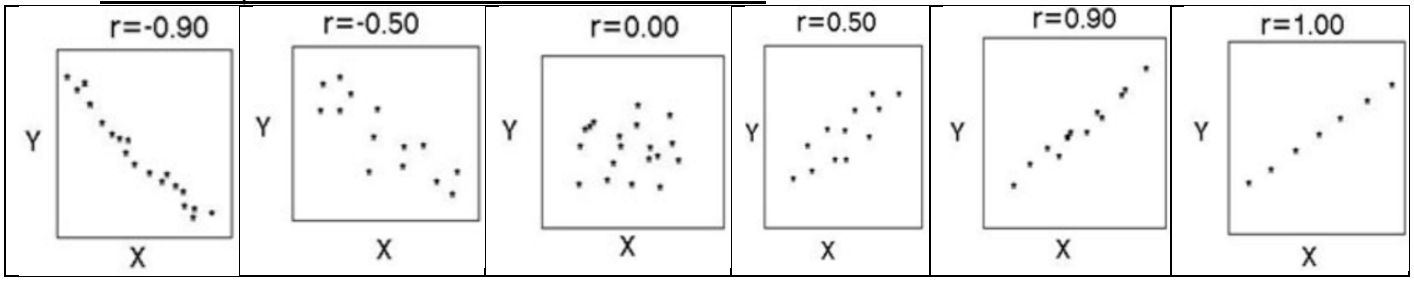
Section 1F: Methods of Sampling

1) Simple Random Sample: A sample of a certain size selected in such a way that each sample of that size has an equal chance of being selected. E.g. put names of the population being studied in a hat, and draw out the names. Or assign everyone a number and use a random number generator to pick them.
2) Stratified Random Sample: The population is divided into two or more subgroups with similar characteristics and then a proportional sample is drawn from each subgroup. E.g. to get the attitudes of students in a particular school to underage drinking where the students are split into 1st Year students, 2nd Year e.t.c. and then a sample is selected from each group. If there are twice as many 2nd years as 1st years, then the sample of 2nd years should be twice as big as the 1st year sample.
3) Cluster Sample: The population is divided into clusters and then the clusters are selected randomly. E.g. if a political party wanted to get the opinions of citizens leaving polling stations on a particular election day, the polling stations would be clusters of the population, and then a number of polling stations are selected and **everyone** coming out of that station is surveyed.
4) Quota Sampling: The person selecting the sample is given a quota to fill and selects it in the most convenient way. E.g. if a company wanted the opinion of men under the age of 25 on a particular issue, so the person collecting the data might stop random people in the street who are under the age of 25. This is not random and can be open to mistakes.
5) Systematic Sampling: The person selecting the sample chooses every n^{th} person from the population. E.g. if Tesco wanted to survey their customers, they could select every 20^{th} person that enters their stores on a particular day and survey them. A disadvantage would be the easy introduction of bias depending on who the n^{th} people are e.g. if every 20^{th} person was a pensioner than the results are not representative of the population of Tesco shoppers.
6) Convenience Sampling: You survey those that are easiest for you to reach. E.g. Surveying people from your workplace or school. One disadvantage of this method is the selection isn't random.

Section 2: Scatter Plot Correlation

- The Correlation Coefficient (r) measures how strong a relationship is between two variables.
- It can have values between -1 and 1 i.e. $-1 \leq r \leq 1$
- If r = 1, then the correlation is said to be **strong** and **positive** and would be a straight line.
- If r = -1, then the correlation is said to be **strong** and **negative** and would also be a straight line.
- The further away the coefficient gets from 1 or -1, the **weaker** the correlation.

Some examples of different correlation coefficients:



Section 3: Revision of Measures of Centre from Junior Cycle

The Mean:

- The mean of a set of values is calculated by adding up all the values and then dividing by how many values there are.
- The symbol that is sometimes used for it is: $\bar{x}$ (pronounced "x bar")
  E.g. the mean of the values 2, 5, 3, 7, 5, 8 is:
  $$\bar{x} = \frac{2+5+3+7+5+8}{6} = \frac{30}{6} = 5$$
- The mean can only be used with numerical data.
  - Advantages: It uses ALL the data.
  - Disadvantages: It is affected by extreme values or outliers and can't be used for categorical data

The Mode:

- The mode of a set of values is the one that appears the most of often.
  e.g. the mode of 1, 2, 1, 3, 3, 2, 1, 2, 3, 2 is 2 as it appears 4 times
  - Advantages:
    - Only one that can be used if data is categorical but can also be used for numerical data.
    - It's easy to find and is not affected by outliers.
  - Disadvantages: There is not always a mode….for example, if no data value is repeated.

The Median:

- The median of a set of values is the MIDDLE value, provided the data has been put in ascending order (or ranked).
  - Advantages:
    - It is easy to calculate and is not affected by outliers.
  - Disadvantages:
    - None really!

Example 1: Data: 3, 5, 2, 5, 4, 3, 3, 1, 6, 5
- Rank the data first i.e. put it in order: 1, 2, 3, 3, 3, 4, 5, 5, 5, 6
- In this case we have 12 values so we find the median by getting the average of the middle two values i.e. the 6th and 7th values => Median = $\frac{3+4}{2} = 3.5$

Useful shortcut for larger amounts of data:

Another way to find the median is:

a) Count the number of values in the data list
b) Add 1 to the answer from (a).
c) Half your answer from (b)
d) Pick out the value that corresponds to your answer from (c).

Example 2: Ranked Data: 2, 2, 2, 2, 3, 3, 4, 4, 4, 5, 5, 5, 6

a) Number of values = 13
b) Add 1 to this = 14
c) Half your answer = 14/2 = 7
d) This means the median is the 7th value i.e. Median = 4

- Section 4A: Revision of Measures of Spread from Junior Cycle:

The Range:

- The range of a set of data is: **Max Value – Min Value**
- One problem with using the range is that it is affected by outliers.

- Section 4B: New Measures of Spread/Variation for Leaving Cert

Interquartile Range:

- The Lower Quartile (Q_1) of a data set is the value that is $\frac{1}{4}$ of the way through the data.
- The Upper Quartile (Q_3) of a data set is the value that is $\frac{3}{4}$ of the way through the data.
- The interquartile range of a data set is:

Upper Quartile – Lower Quartile

Example 1: Ranked data: 2, 3, 3, 4, 5, 5, 5, 6, 6, 7, 8, 8, 9

- There are 13 values so the middle value first of all is 5 as there are six values either side of it. This is the median, as before. It can be called Q_2 as well.
- The lower quartile Q_1 is midway between the 3 and the 4, so it is 3.5.
- The upper quartile Q_3 is between the 7 and the 8, so it is 7.5.
- So, the Interquartile Range = 7.5 - 3.5 = 4

Steps for finding Q_1, Q_3 and IQR:

a) Find the median of the data and use it to split the data into two halves.
b) Find the median of the lower half of the data.
c) Find the median of the upper half of the data.
d) Subtract the two answers from steps (b) and (c) to get the IQR.

Example 2: Ranked data: 1, 1, 3, 4, 6, 6, 7, 7, 7, 9, 9, 10, 11, 15, 15, 16

a) There are 16 values in total, so the median will be $\frac{16+1}{2} = 8.5^{th}$ value = 7
b) The lower half contains 8 numbers => the median will be $\frac{8+1}{2} = 4.5^{th}$ value, which will be $\frac{4+6}{2} = 5$
c) The upper half also contains 8 numbers so the median will be the 4.5^{th} value of that set also, which is $\frac{10+11}{2} = 10.5$
   - That means our IQ Range = 10.5 - 5 = 5.5

Standard Deviation:

- The standard deviation is another way of measuring spread.
  - Advantages: It uses ALL the data.
  - Disadvantages: It is affected by extreme values or outliers.

➢ Section 5: Calculator Work

Standard Deviation and Mean using Casio fx-83GT:

Calculator Use:

x	5	8	11	14	17	20
f	8	5	3	9	5	2

Step 1:

- Turn on STAT mode by pressing 'MODE' and then '2'.
- Press '1' then to select "1-VAR" mode.
- The screen should now look like the screen on the right.

Note: If the "FREQ" column is not visible, then follow the following steps:

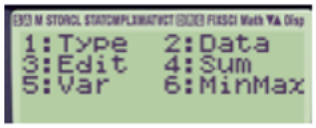
- Press 'SHIFT' and then 'MODE' to enter the screen shown on the right.
- Press the Down Arrow and then press '3' for "STAT".
- Now press '1' to turn the frequency setting ON.

Step 2:

- Enter the data from the table above by typing in the value and pressing '=' every time.
- Use the arrows to navigate between the columns.
- When all the data has been entered, the screen should look like the screen on the right.
- Now press 'AC' to clear the screen.

Step 3:

- Press 'SHIFT' and '1' to enter the STAT menu, which looks something like the screen on the right.
- Press the number corresponding to the "VAR" option.
- Now press the number that corresponds to the "σ x" and then press '=' to get the standard deviation of 4.85.



Note: We could also find the mean of the distribution by selecting $\bar{x}$ from the menu instead. Verify that the mean of this distribution is 11.375.

Correlation Coefficient using Casio fx-83GT:

Calculation of Correlation Coefficient:

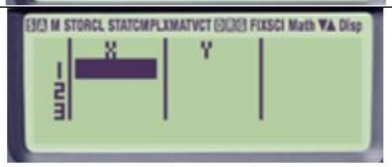
Variable 1	35	42	51	38	44	37	48	38	36
Variable 2	31	33	46	32	53	37	32	40	30

Step 1: Putting the calculator in the correct mode.

- Press 'MODE' and then '2' to enter STAT mode.
- The menu shown on the right should now be shown. Press '2' to select the "A+BX" mode

Step 2: Entering the data.

- The screen should now look like the screen on the right. There might be an extra "FREQ" column but that will have 1s in it that won't affect anything.
- The data for "Variable 1" in the table above goes into Column X and the data for "Variable 2" goes into Column Y.
- Type the first data value '35' and then press '='
- Enter the rest of the data for Variable 1 by typing in the value and pressing '=' each time.
- Repeat the steps above for Variable 2.
- The screen should now look like the screen on the right.
- The data is now entered so press 'AC' to clear the screen



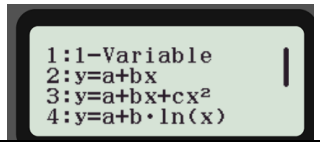
Step 3: Calculating the correlation coefficient 'r'.

- Press 'SHIFT' and '1' to enter the STAT menu. It looks like the screen on the right.
- Press '7' to access the REG menu.
- Now press '3' to access the correlation coefficient 'r'.
- The screen should now have an 'r' on the top of it. Press '=' to get the answer of **0.495**.

Standard Deviation and Mean using Casio fx-83GT X:

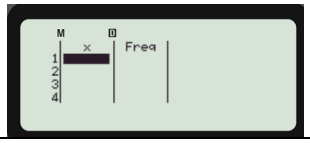
Step 1:

- Press "Menu" and "2" to enter Statistics mode on the calculator.
- You should now see the screen shown on the right.



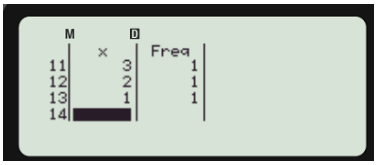
Step 2:

- Press "1" for Single Variable mode.
- You should now see the screen shown on the right.



Step 3a: For a single list of data: e.g. 2, 3, 1, 2, 3, 2, 4, 1, 2, 4, 3, 2, 1

- Enter the list of data above in the X column by typing in the value and then pressing "EXE" after each entry.
- The frequency values will be set to 1 by default, which is perfect.



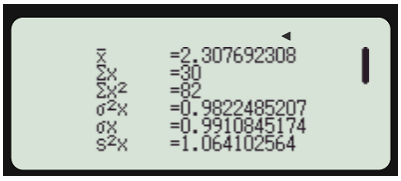
Step 3b: For a frequency distribution: E.g.

X	2	4	6	8	10
F	4	8	3	9	2

- Enter the list of data above in the X column by typing in the value and then pressing "EXE" after each entry.
- Then use the arrows to navigate back to the start of the "Freq" column and enter the numbers from the 2nd row of the table above.

Step 4:

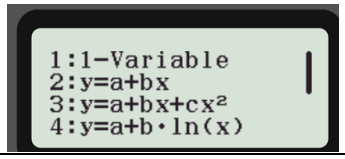
- Then press "OPTN" and then "2" for Single Variable calculations.
- The first figure $\bar{x}$ is the mean of the data i.e. $\bar{x} = 2.3$
- The second last figure σx is the standard deviation i.e. 0.99108



Correlation Coefficient using Casio fx-83GT X:

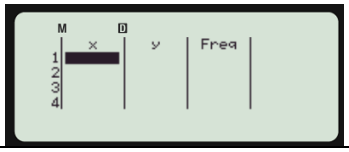
Step 1:

- Press "Menu" and "2" to enter Statistics mode on the calculator.
- You should now see the screen shown on the right.



Step 2:

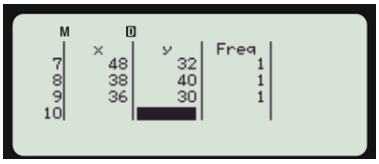
- Press "2" for Dual Variable mode.
- You should now see the screen shown on the right.



Step 3: For the data below:

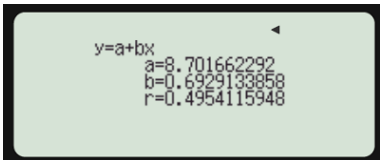
Variable 1	35	42	51	38	44	37	48	38	36
Variable 2	31	33	46	32	53	37	32	40	30

- Enter the data from Variable 1 in the X column by typing in the value and then pressing "EXE" after each entry.
- Then go back to the top of the Y column with the arrow buttons and enter the data from Variable 2 in the Y column in the same way.
- The frequency values will be set to 1 by default, which is perfect.



Step 4:

- Then press "OPTN" and then "4" for Regression Calculations.
- The value for "r" is the correlation coefficient. i.e. r = 0.4954
- The values of "a" and "b" are for the Line of Best Fit, so in this case the equation of the line would be: $y = 8.7x + 0.6929$



Standard Deviation and Mean using Casio fx-83GT CW:

Step 1:

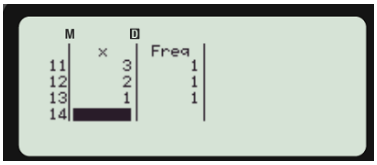
- On the Home screen, press the right arrow to select "Statistics" and press "OK".
- You should now see the screen shown on the right.

Step 2:

- Press "OK" again to select "1-Variable"
- You should now see the screen shown on the right.

Step 3a: For a single list of data: e.g. 2, 3, 1, 2, 3, 2, 4, 1, 2, 4, 3, 2, 1

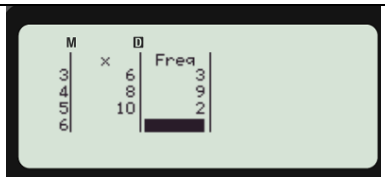
- Enter the list of data above in the X column by typing in the value and then pressing "EXE" after each entry.
- The frequency values will be set to 1 by default, which is perfect.



Step 3b: For a frequency distribution: E.g.

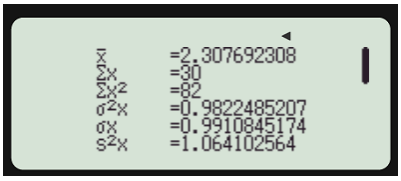
X	2	4	6	8	10
F	4	8	3	9	2

- Enter the list of data above in the X column by typing in the value and then pressing "EXE" after each entry.
- Then use the arrows to navigate back to the start of the "Freq" column and enter the numbers from the 2nd row of the table above.



Step 4:

- Then press "OK" and then "OK" again to select "1-Var Results".
- The first figure $\bar{x}$ is the mean of the data i.e. $\bar{x} = 2.3$
- The second last figure σx is the standard deviation i.e. 0.99108



Correlation Coefficient using Casio fx-83GT CW:

Step 1:

- On the Home screen, press the right arrow to select "Statistics" and press "OK".
- You should now see the screen shown on the right.

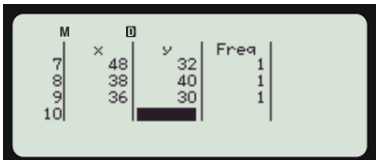
Step 2:

- Press the down arrow and "OK" again to select "2-Variable"
- You should now see the screen shown on the right.

Step 3: For the data below:

Variable 1	35	42	51	38	44	37	48	38	36
Variable 2	31	33	46	32	53	37	32	40	30

- Enter the data from Variable 1 in the X column by typing in the value and then pressing "EXE" after each entry.
- Then go back to the top of the Y column with the arrow buttons and enter the data from Variable 2 in the Y column in the same way.
- The frequency values will be set to 1 by default, which is perfect.



Step 4:

- Then press "OK" and use down arrow to select "Reg Results".
- Press "OK" again to select "y = a + bx" (Linear Regression)
- The value for "r" is the correlation coefficient. i.e. r = 0.4954
- The values of "a" and "b" are for the Line of Best Fit, so the equation of the line of best fit would be: $y = 8.7x + 0.6929$