General Maths Unit 3

Core Topic: Data Analysis
Chapter 2 Part 2
Investigation Associations between two Variables

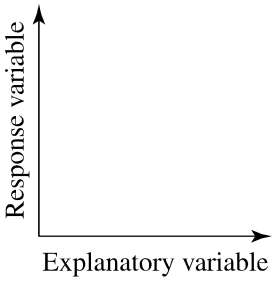
Area of Study 1 – Data Analysis

- response and explanatory variables and their role in investigating associations between variables
- scatterplots and their use in identifying and qualitatively describing the association between two numerical variables in terms of direction (positive/negative), form (linear/non-linear) and strength (strong/moderate/weak)
- answering statistical questions that require a knowledge of the associations between pairs of variables
- Pearson correlation coefficient, *r*, its calculation and interpretation
- Cause and effect; the difference between observation and experimentation when collecting data and the need for experimentation to definitively determine cause and effect
- non causal explanations for an observed association including common response, confounding, and coincidence; discussion and communication of these explanations in a particular situation in a systematic and concise manner.

Investigating Associations Between Two Numerical Variables (Ex 2d)

The first step in investigating the association between two numerical variables is to construct a **scatterplot.**

- The horizontal axis (x) is always the explanatory variable.
- The vertical axis (y) is always the response variable.



Quick Revision on Explanatory and Response Variables…

For each of the following pairs of variables, identify the explanatory (independent) variable and the response (dependent) variable. If it is not possible to identify this, then write "not appropriate".

a) The number of visitors at a local swimming pool and the daily temperature.

b) The blood group of a person and his or her favourite TV channel.

****Know how to plot scatterplots by hand and using CAS**

Edrolo link: https://edrolo.com.au/s/2678924/

Example – in class task

Use your CAS calculator to display the following information using a scatterplot

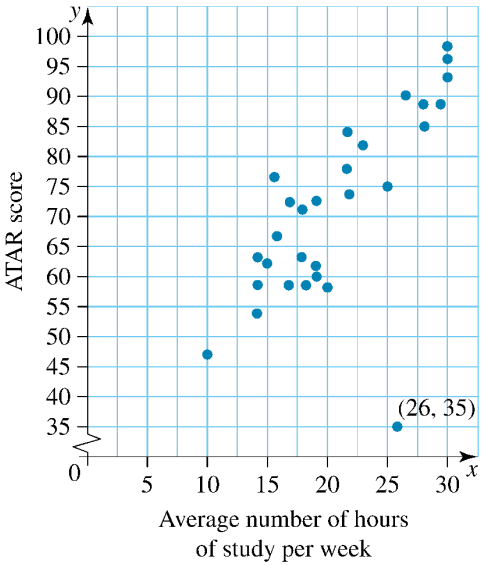
Average hours of study	ATAR score
18	59
16	67
22	74
27	90
15	62
28	89
18	71
19	60
22	84
30	98
14	54
17	72
14	63
19	72
20	58

Average hours of study	ATAR score
10	47
28	85
25	75
18	63
19	61
17	59
16	76
14	59
29	89
30	93
30	96
23	82
26	35
22	78

Method

1. In lists and spreadsheets name the first column hours and the second column atar and input the data
2. Add a new page in "Data and Statistics"
3. All your dots will appear
4. Go to the horizontal axis and add the title hours
5. Do the same to the vertical axis and have atar shown

Your scatterplot should look like this……



Ex 2D

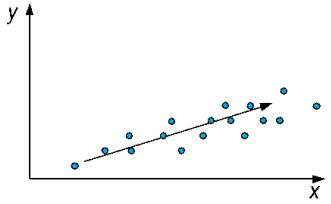
How to interpret a Scatterplot (Ex 2e)

Scatterplots can be used to describe 3 aspects of the relationship between the variables.

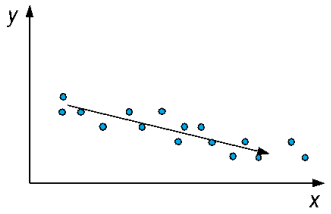
- Direction and outliers(positive/negative)
- Form (Linear/Non Linear)
- Strength (Strong, Moderate, Weak)

Interpretation of Scatterplots

1 Direction



Positive association
The points generally go up as x increases, similar to a straight line with positive gradient.
"As the independent variable (x) increases, the dependent variable (y) also increases."

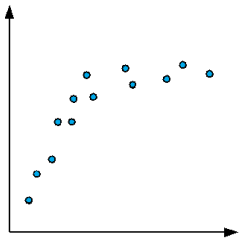


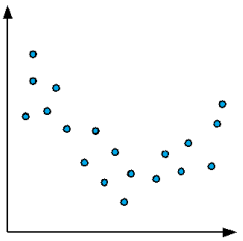
Negative association
The points generally go down as 'x' increases, similar to a straight line with negative gradient.
"As the independent variable (x) increases, the dependent variable (y) decreases."

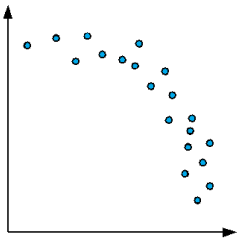
2 Form

In the scatterplots above, the points are generally in a straight line. The relationship between the variables is said to be linear.

These scatterplots show relationships which are not linear.





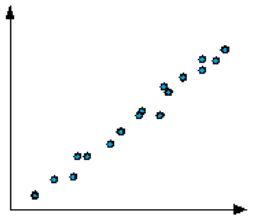


3 Strength

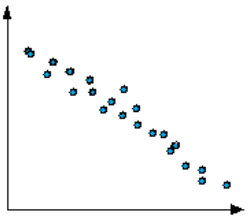
If the points form a well-ordered pattern then the strength of the association is said to be **strong**.

For example:

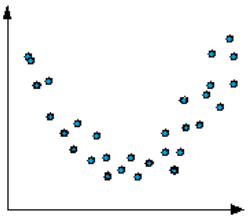
Strong positive



Strong negative



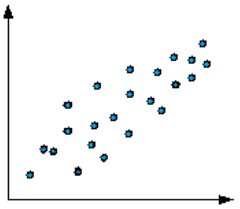
Strong non-linear



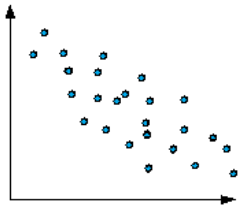
If the points form a pattern which is less well defined, then the strength is said to be **moderate**.

For example:

Moderate positive



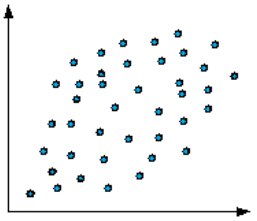
Moderate negative



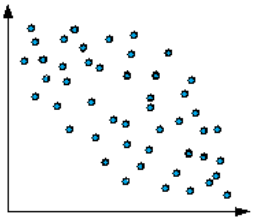
If the points are scattered but a general pattern is still discernable then the association is said to be **weak**.

For example:

Weak positive

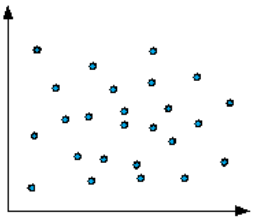


Weak negative



If the points appear to be randomly scattered then there is **no association** between the variables.

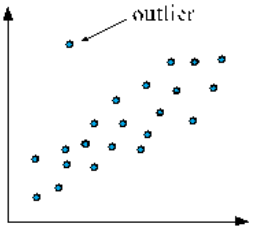
An example of this is shown opposite.



4 Outliers

Outliers stand out from the general body of data.

The example opposite shows a "moderate positive association with one outlier".

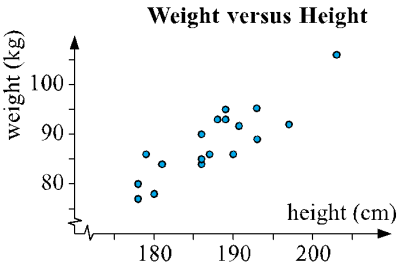


Example

after careful consideration.

We can interpret the *Weight* versus *Height* scatterplot from earlier as follows:

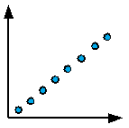
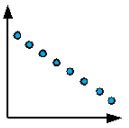
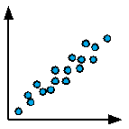
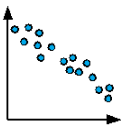
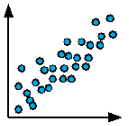
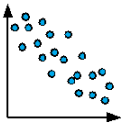
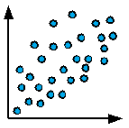
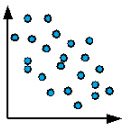
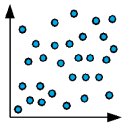
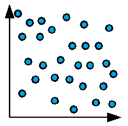
"There is a moderate positive association between the variables *height* and *weight*. This means that as height increases, weight increases. The relationship appears linear and there are no obvious outliers."



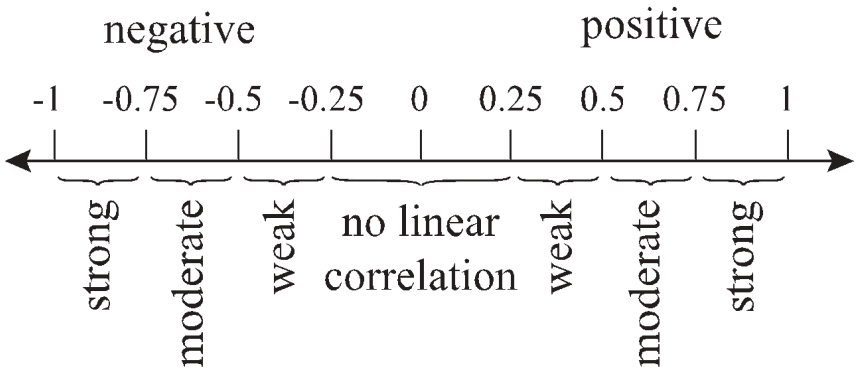
Ex 2E

Strength of a linear relationship - Pearson's product-moment correlation coefficient (Ex 2F)

Pearson's product-moment correlation coefficient (r) is used to measure the strength of a linear correlation between two variables. It varies from -1 to 1 as shown below.

r	Description	r	Description
1	perfect positive correlation 	-1	perfect negative correlation 
0.75 to 1	strong positive correlation 	-1 to -0.75	strong negative correlation 
0.50 to 0.75	moderate positive correlation 	-0.75 to -0.50	moderate negative correlation 
0.25 to 0.50	weak positive correlation 	-0.50 to -0.25	weak negative correlation 
0 to 0:25	almost no correlation 	¡ 0:25 to 0	almost no correlation 

or……



When using Pearson's product-moment correlation coefficient

- Positive values always represent relationships with a positive gradient

- Negative values always represent relationships with a negative gradient

Note that….

- It is designed for linear data only
- It should be used with caution if outliers are present

Warning!!!

If you use the value of the correlation coefficient as a measure of the strength of an association, you are implicitly implying that:

1. The variables are numeric
2. The association is linear
3. There are no outliers in the data

The correlation coefficient can give a misleading indication of the strength of the linear association if there are outliers present.

Calculating the correlation coefficient (r)

Pearson's product-moment correlation coefficient (r)

We generally use CAS to calculate this, however this is the formula used:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

where n is the number of pairs of data in the set

s_x is the standard deviation of the x-values

s_y is the standard deviation of the y-values

$\bar{x}$ is the mean of the x-values

$\bar{y}$ is the mean of the y-values.

There are two important limitations on the use of r:

1. Since r measures the strength of a linear relationship it is not appropriate to use for non-linear data.
2. Outliers can bias r, therefore if outliers are present r is not a reliable measure of the strength of the relationship.

Therefore…It is a good idea to draw up a quick scatterplot so that you can see how accurate r will be, that is, you will test to see the data is linear in form and that outliers don't exist!

- The calculation of r is often done using a CAS calculator

- Even if we have two variables with a high degree of correlation, e.g. r=0.95, we **cannot** say that the value of one variable is **caused** by the value of the other variable.
- If r = 1 or -1 there is a perfect linear relationship

Using CAS

The heights of 21 footballers were recorded against the number of marks they took in a game of football. The data are shown in the following table.

a) Construct a scatterplot of the data

b) Comment on the correlation between the heights of the players and the number of marks that they take, and estimate a value of r. (Use direction, strength, form, outliers)

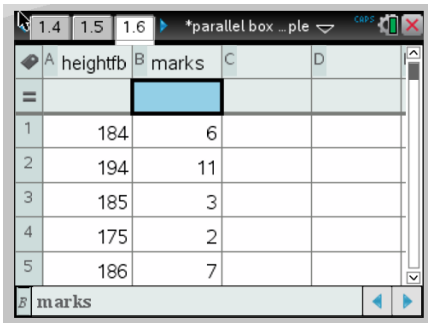
c) Calculate r and use it to comment on the relationship between the heights of the players and the number of marks they take in a game

Height (cm)	184	194	185	175	186	183	174	200	188	184	188	182
No of marks	6	11	3	2	7	5	4	10	9	7	6	7

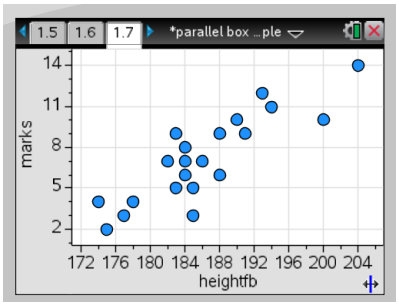
Height (cm)	185	183	191	177	184	178	190	193	204
No of marks	5	9	9	3	8	4	10	12	14

Using CAS

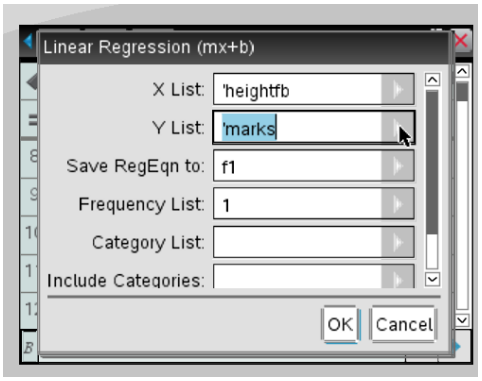
1. Set up a Lists & Spreadsheet page with the following data

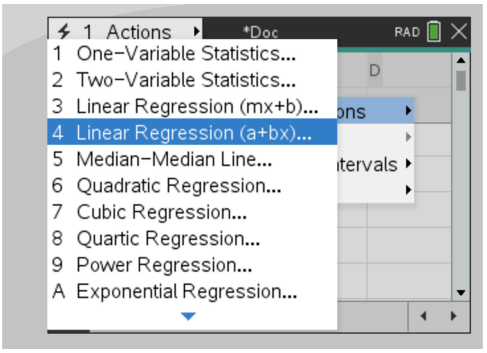


2. Construct a scatterplot and estimate the value of r and also check that the relationship is linear with no outliers.



3. The calculate the value of r and r^2go to
   - MENU
   - 4. Statistics
   - 1 Stat Calculations
   - 3. Linear Regression (ax+b)
   - Press OK



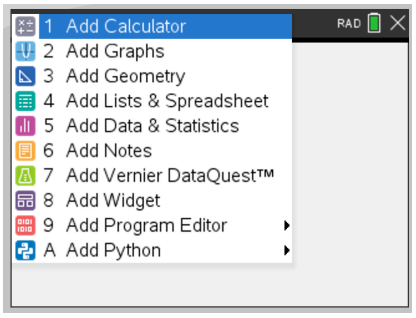


Scroll down and you will see that r = 0.859311 and r^2=0.738415

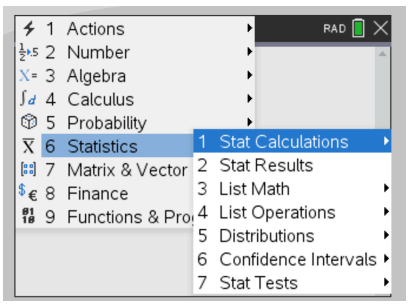
OR……

To calculate r and the other linear regression analysis statistics….

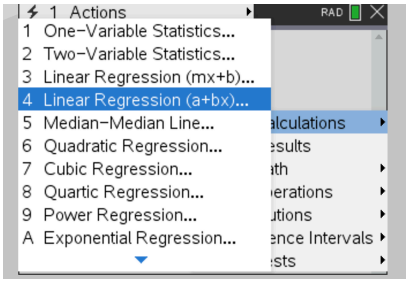
1. Add a page and go to "add calculator"



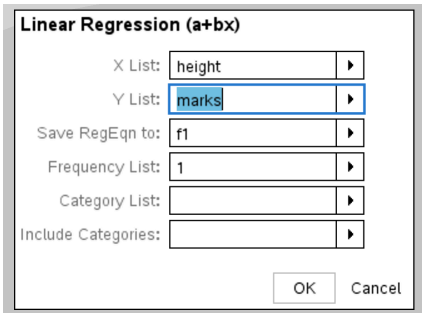
2. Go to Menu – Statistics- Stat Calculations

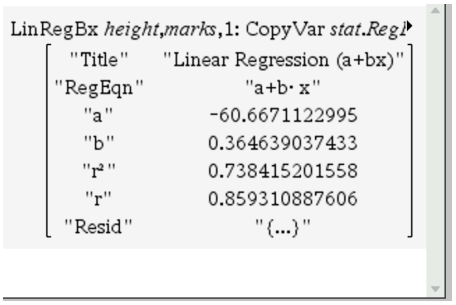


3. Go to Linear Regression a + bx



4. Fill out x list and y list again





Ex 2f

5. Results will be shown

Edrolo link: https://edrolo.com.au/s/2678925/

The Coefficient of Determination (r^2) (Ex 2G)

- The Coefficient of Determination is equal to r^2
- It describes the **influence** the explanatory variable (x) has on the response variable (y).
- It is usually expressed as a percentage. (To change the value of r^2 to a percentage, multiply by 100)
- It provides a measure of how well the linear rule linking the two variables (x and y) predicts the value of y when we are given the value of x.

In a bivariate set of numerical data, the coefficient of determination gives us a means of measuring the influence that one variable has over the other variable.

Coefficient of determination = r^2 = (Pearson's correlation coefficient)^2

Sentence to explain……

The coefficient of variation tells us that ____% of the variation in ***y value*** *can be explained by the variation in* ***x value.***

Or…..

The proportion of variation in ***y value*** *can be explained by the variation in* ***x value*** *is ____%.*

r = 0.86 means that there is a strong, positive, linear association between the height of a player and the number of marks he takes in the gams. That is, the taller the player, the more marks we might expect him to take.

r^2=0.738415 means that 74% of the variation in the number of marks taken can be explained by the variation in height.

Example

A set of data giving the number of police patrols on duty and the number of fatalities for the region was recorded and a correlation coefficient of r = -0.8 was found.

a) Calculate the coefficient of determination and interpret its value.

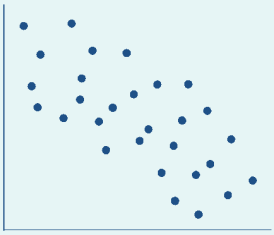
Example

Calculating the correlation coefficient from the coefficient of determination

Example 10 **Calculating the correlation coefficient from the coefficient of determination**

For the relationship described by this scatterplot, the coefficient of determination = 0.5210.

Determine the value of the correlation coefficient, *r*.



Edrolo link: https://edrolo.com.au/s/2678926/

Ex 2G

Correlation and causation (Ex 2H)

- We found above that r=0.86, therefore we may be thinking that there is a strong association between the height of a footballer and the number of marks they take. We cannot say though that the height of a footballer causes him to take lots of marks!! There are many other factors that might come into it (skill, teammates accuracy in kicking, strength of the opposition etc.)
- Therefore we can establish a **high correlation** between the two variables but we cannot say one variable **causes** the other variable.
- **Causation** states that one event is the result of the occurrence of the other event.
- An example of **cause and effect relationship** is that an alarm could go off **(cause)** and a person wakes up **(effect)**
- High correlation does not imply causation e.g. a person smoking could have high correlation with alcoholism, but it does not necessarily cause alcoholism.

Correlation does not imply causality

A correlation tells you about the strength of the association between the variables, but no more. It tells you nothing about the source or cause of the association.

Non-causal explanations

Three points to consider…. (the three "c"s)

- A strong correlation between two variables does not necessarily mean that an association exists. E.g. a study could show that a strong correlation exists between house size and life expectancy. This does not mean that a big house leads to a long life!! A **common response variable** (income of the owner) provides a direct link to both variables and is more likely to be the underlying cause of the correlation.

- There may be hidden reasons for the strong correlation between two variables. E.g. a lack of exercise may have a strong correlation to heart failure. Hidden variables such as poor nutrition and lifestyle may have a stronger influence. These are called **confounding variables.**

- Finally an association between two variables may be **coincidental!** The larger the data set the less chance of coincidence.

Exam Questions

2011

Question 11

For a group of 15-year-old students who regularly played computer games, the correlation between the time spent playing computer games and fitness level was found to be $r = -0.56$.

On the basis of this information it can be concluded that

A. 56% of these students were not very fit.

B. these students would become fitter if they spent less time playing computer games.

C. these students would become fitter if they spent more time playing computer games.

D. the students in the group who spent a short amount of time playing computer games tended to be fitter.

E. the students in the group who spent a large amount of time playing computer games tended to be fitter.

2013

Question 7

For a city, the correlation coefficient between

- population density and distance from the centre of the city is $r = -0.563$
- house size and distance from the centre of the city is $r = 0.357$.

Given this information, which one of the following statements is true?

A. Around 31.7% of the variation observed in house size in the city can be explained by the variation in distance from the centre of the city.

B. Population density tends to increase as the distance from the centre of the city increases.

C. House sizes tend to be larger as the distance from the centre of the city decreases.

D. The slope of a least squares regression line relating population density to distance from the centre of the city is positive.

E. Population density is more strongly associated with distance from the centre of the city than is house size.

Question 8

The table below shows the hourly rate of pay earned by 10 employees in a company in 1990 and in 2010.

Employee	Hourly rate of pay ($)	
	1990	2010
Ben	9.53	17.02
Lani	9.15	16.71
Freya	8.88	15.10
Jill	8.60	15.93
David	7.67	14.40
Hong	7.96	13.32
Stuart	6.42	15.40
Mei Lien	11.86	19.79
Tim	14.64	23.38
Simon	15.31	25.11

The value of the correlation coefficient, r, for this set of data is closest to

A. 0.74

B. 0.86

C. 0.92

D. 0.93

E. 0.96

Ex 2H

Which Graph???? (Ex 2i)

Type of variables		*Graph*
Response variable	*Explanatory variable*	
Categorical	Categorical	Segmented bar chart, side-by-side (parallel) bar chart
Numerical	Categorical	Parallel box plots, parallel dot plots
Numerical	Categorical (two categories only)	Back-to-back stem plot, parallel dot or box plots
Numerical	Numerical	Scatterplot

Ex 2I