



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

200
1822 - 2022

Correlation Coefficients

Man Geni Addysg Uwch yng Nghymru
Where Higher Education Began in Wales

www.uwtsd.ac.uk

Introduction

- Today we'll look at
 - Correlation Coefficient
 - Pearson Correlation Coefficient
 - Interpreting Correlation Coefficients
- Until now we've been looking at data that can be described using a single variable
- Today we will start to look at the relationship between two variables

Is there a relationship

- between your IQ and the wealth of your parents?
- Between your computer skills and your GPA?
- Between your anxiety level and your perceived social attractiveness?

Need for correlation

- **Relationship Strength:**

- Covariance and correlation quantify the relationship between two variables.

- They indicate whether changes in one variable correspond to changes in another.

Ex: study time Versus Exam scores

income Versus spending etc.,

- **Prediction and Forecasting:**

- A strong positive or negative correlation helps in predicting one variable based on the other.

Ex: In sales forecasting, understanding the correlation between advertising expenditure and sales helps predict future sales figures based on ad spending.

- **Quality Control:**

- In manufacturing and quality control, understanding the correlation between **manufacturing parameters** and **product quality** helps in **optimizing production processes** and ensuring consistent quality.

- **Policy Decisions:**

- For policymakers, understanding correlations between different socio-economic factors helps in making informed decisions. For instance, correlations between **education levels** and **employment rates** can inform policies related to education and job creation

Correlation Coefficients

- measures the strength and direction of the relationship between two variables.
- The correlation coefficient r_{xy}

Values from -1.0 to +1.0

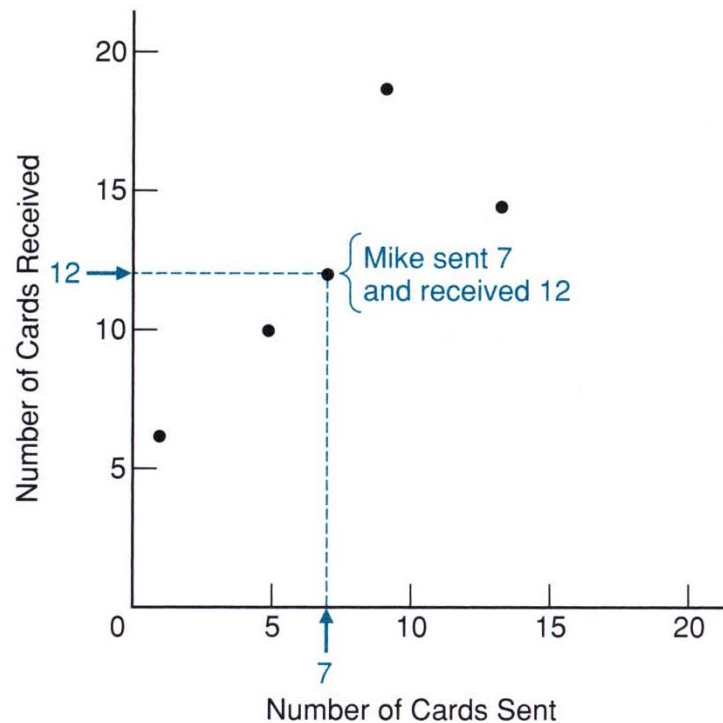
A correlation of 1 indicates a perfect positive linear relationship,

-1 indicates a perfect negative linear relationship,

When $r_{xy} = 0$ no relationship exists between x and y

Strength of relationship $\propto$ absolute value of r_{xy}

An investigator suspects that a relationship exists between the number of greeting cards sent and the number of greeting cards received by individuals

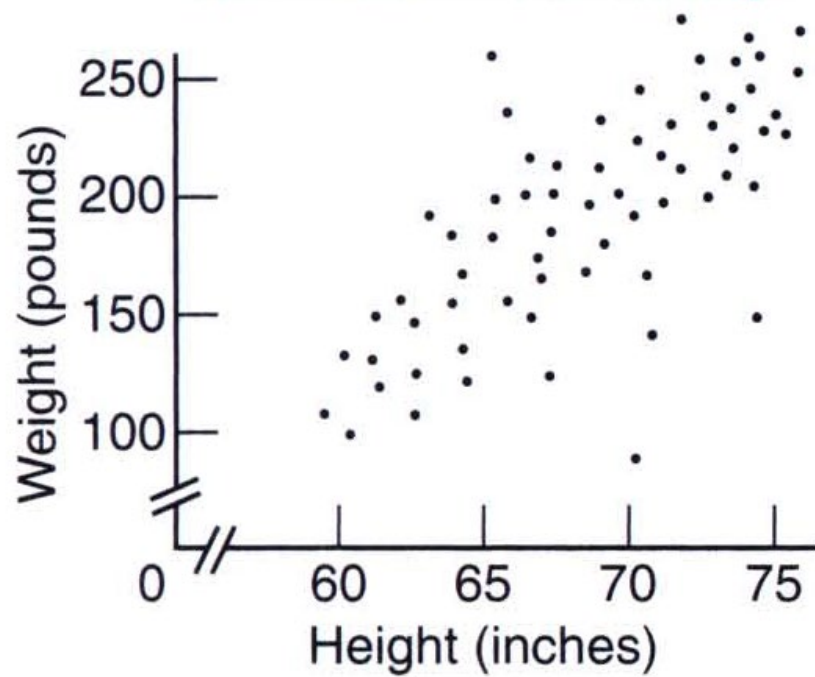


Scatterplot for greeting card exchange.

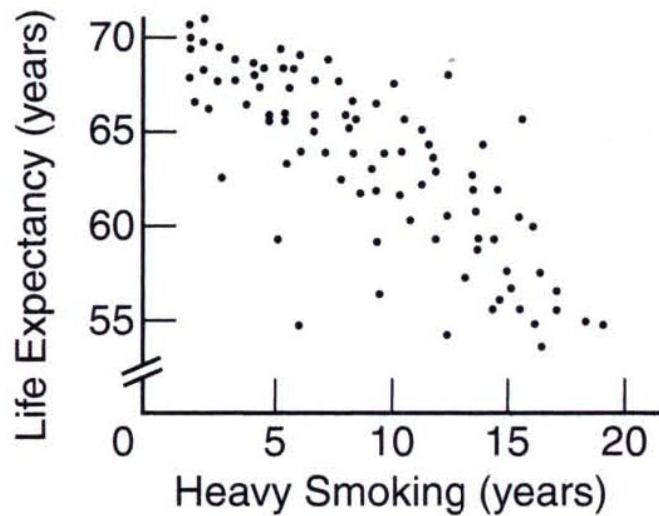
NUMBER OF CARDS

Friend	Sent	Received
Andrea	5	10
Mike	7	12
Doris	13	14
Steve	9	18
John	1	6

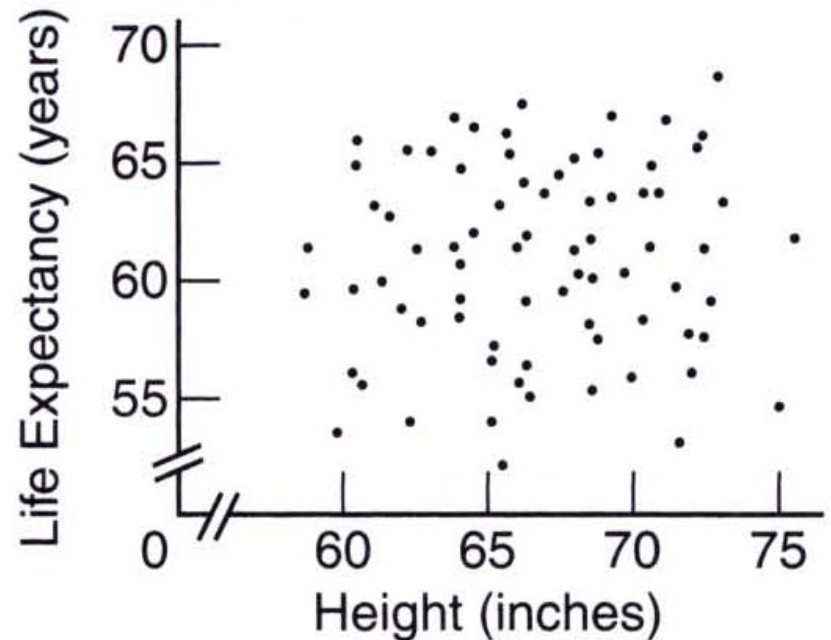
A. Positive Relationship



B. Negative Relationship

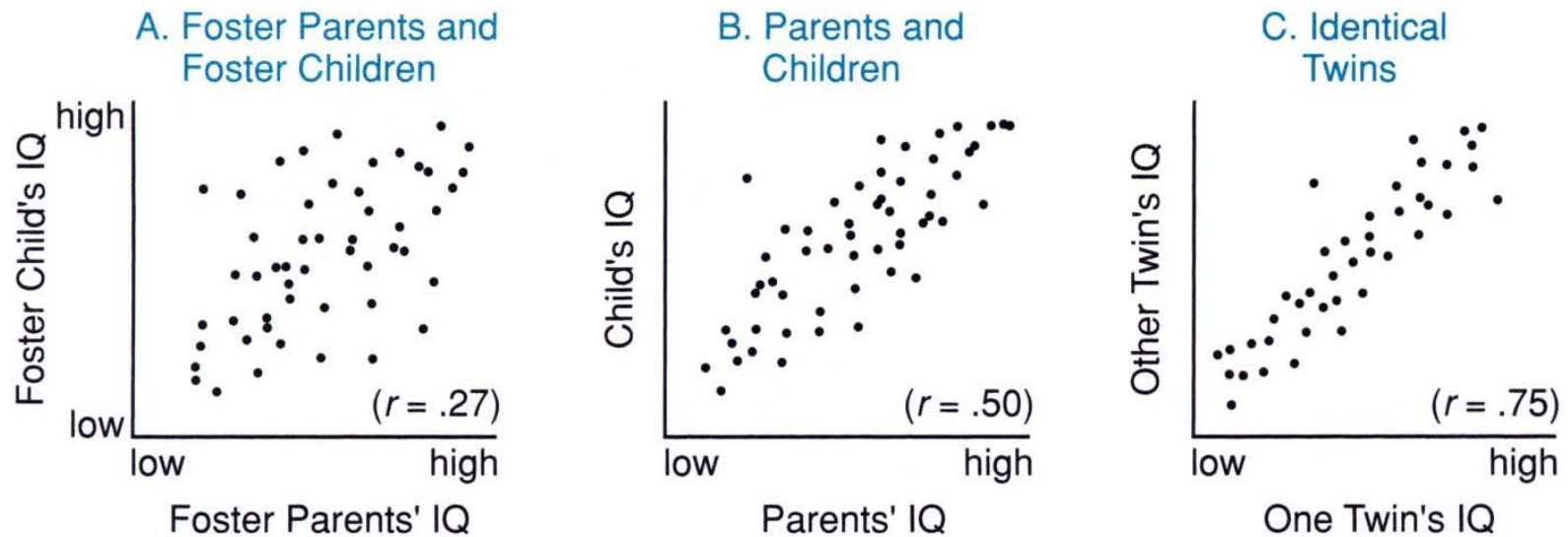


C. Little or No Relationship



Strong or Weak Relationship?

Genetic
overlap



*Three positive relationships. (Scatterplots simulated from a 50-year literature survey.)
Source: L. Erlenmeyer-Kimling and L. F. Jarvik. "Genetics and Intelligence: A Review."
Science, 142, 1477–1479.)*

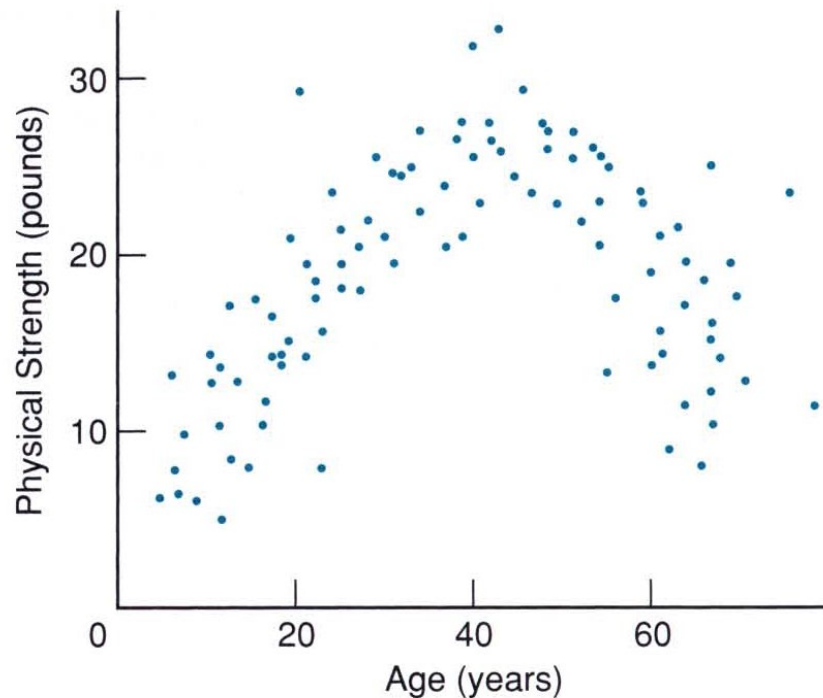
Perfect Relationship

Linear Relationship

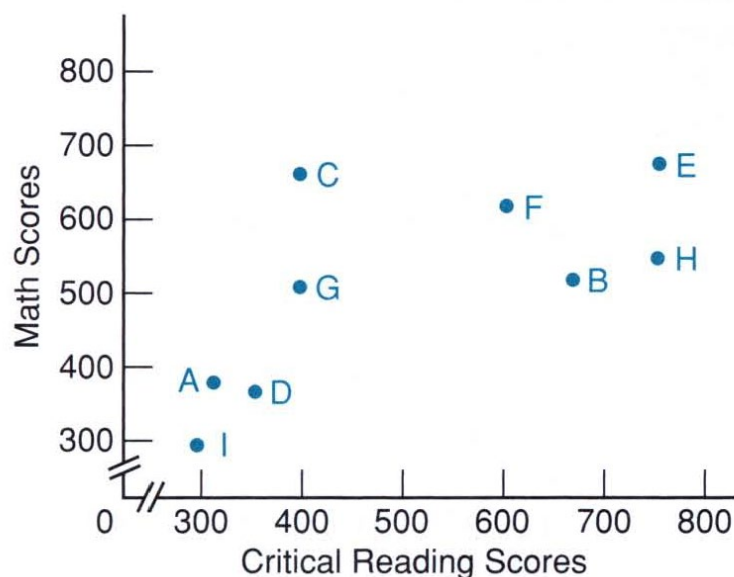
- A relationship that can be described best with a straight line.
- A dot cluster that equals (rather than merely approximates) a straight line reflects a perfect relationship between two variables.
- In practice, perfect relationships are most unlikely.

Curvilinear Relationship

- A relationship that can be described best with a curved line.



Critical reading and math scores on the SAT test for students A, B, C, D, E, F, G, and H are shown in the following scatterplot:



- (a) Which student(s) scored about the same on both tests?
- (b) Which student(s) scored higher on the critical reading test than on the math test?
- (c) Which student(s) will be eligible for an honors program that requires minimum scores of 700 in critical reading and 500 in math?
- (d) Is there a negative relationship between the critical reading and math scores?

Answers:

(a) I, D, F

(b) B, H, E

{c} E, H

{d} No. The relationship is positive.

Correlation Coefficient

- A number between -1 and 1 that describes the relationship between pairs of variables.

Named in honor of the British scientist Karl Pearson, the **Pearson correlation coefficient**, r , can equal any value between -1.00 and +1.00 .

- It assesses the strength and direction of the linear association between variables, assuming that the relationship between the variables is linear and that the data is normally distributed.
- Furthermore, the following two properties apply:
 1. *The sign of r indicates the type of linear relationship, whether positive or negative.*
 2. *The numerical value of r , without regard to sign, indicates the strength of the linear relationship.*

Pearson Correlation Coefficient

Methods

- Deviation Score Formula
- Computing Formula

Example

Person	X	Y	X - $\bar{X}$	Y - $\bar{Y}$	(X - $\bar{X}$)(Y - $\bar{Y}$)
1	10	15	-5	-5	25
2	12	18	-3	-2	6
3	15	20	0	0	0
4	18	22	3	2	6
5	20	25	5	5	25
Total	75	100	0	0	62
N	5				
$\sum(X - \bar{X})(Y - \bar{Y})$	62				
$\sum(X - \bar{X})^2$	35				
$\sum(Y - \bar{Y})^2$		50			
SD_x	2.915				
SD_y		2.236			
r_{xy}	0.714				

Deviation Score Formula

$$r_{xy} = \frac{Cov_{xy}}{SD_x SD_y}$$

Where

$$Cov_{xy} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{N - 1}$$

$(X - \bar{X})$, the deviation scores for X

$(Y - \bar{Y})$, the deviation scores for Y

$\sum(X - \bar{X})(Y - \bar{Y})$ = the sum of the products of the paired deviation scores

N = the number of pairs of scores

SD_x = the standard deviation of X

SD_y = the standard deviation of Y

Example

			$x = (X - \bar{X})$	$y = (Y - \bar{Y})$	
Persons	X	Y	x	y	xy
Alice	37	75	-1.8	-4.8	8.64
Bob	41	78	2.2	-1.8	-3.96
Charlotte	48	88	9.2	8.2	75.44
David	32	80	-6.8	0.2	-1.36
Eve	36	78	-2.8	-1.8	5.04
N=5	$\sum X = 194$	$\sum Y = 399$			$\sum xy = 83.8$
	$\bar{X} = 38.8$	$\bar{Y} = 79.8$			

$$\bar{X} = \frac{\sum X}{N} = 194/5$$

$$= 38.8$$

Example

Deviation Score Formula

$$r_{xy} = \frac{cov_{xy}}{SD_x SD_y}$$

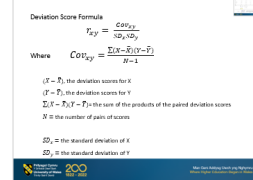
Where $cov_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n - 1}$

($\bar{x} = \bar{X}$, the deviation scores for X)
 ($\bar{y} = \bar{Y}$, the deviation scores for Y)
 $\sum (x - \bar{x})(y - \bar{y})$ the sum of the products of the paired deviation scores
 n is the number of pairs of scores
 SD_x is the standard deviation of X
 SD_y is the standard deviation of Y

			$x = (X - \bar{X})$	$y = (Y - \bar{Y})$	
Persons	X	Y	x	y	xy
Alice	37	75	-1.8	-4.8	8.64
Bob	41	78	2.2	-1.8	-3.96
Charlotte	48	88	9.2	8.2	75.44
David	32	80	-6.8	0.2	-1.36
Eve	36	78	-2.8	-1.8	5.04
N=5	$\sum X = 194$	$\sum Y = 399$			$\sum xy = 83.8$
	$\bar{X} = 38.8$	$\bar{Y} = 79.8$			
	$SD_x = 6.1$				

$$SD_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} = \sqrt{\frac{(-1.8)^2 + (2.2)^2 + (9.2)^2 + (-6.8)^2 + (-2.8)^2}{4}} = 6.05$$

Example



Persons			$x = (X - \bar{X})$	$y = (Y - \bar{Y})$	xy
	X	Y	x	y	
Alice	37	75	-1.8	-4.8	8.64
Bob	41	78	2.2	-1.8	-3.96
Charlotte	48	88	9.2	8.2	75.44
David	32	80	-6.8	0.2	-1.36
Eve	36	78	-2.8	-1.8	5.04
N=5	$\sum X = 194$	$\sum Y = 399$			$\sum xy = 83.8$
	$\bar{X} = 38.8$	$\bar{Y} = 79.8$			
		$SD_y = 4.9$			

$$SD_y = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}} = \sqrt{\frac{(-4.8)^2 + (-1.8)^2 + (8.2)^2 + (0.2)^2 + (-1.8)^2}{5 - 1}} = 4.9$$

Example

Persons	X	Y	x	y	xy
Alice	37	75	-1.8	-4.8	8.64
Bob	41	78	2.2	-1.8	-3.96
Charlotte	48	88	9.2	8.2	75.44
David	32	80	-6.8	0.2	-1.36
Eve	36	78	-2.8	-1.8	5.04

$$N=5 \quad \sum X = 194 \quad \sum Y = 399 \quad \sum xy = 83.8$$

$$\bar{X} = 38.8 \quad \bar{Y} = 79.8$$

$$SD_x = 6.1 \quad SD_y = 4.9$$

$$Cov_{xy} = \frac{\sum xy}{N-1} = \frac{83.8}{5-1} = 20.95$$

$$r_{xy} = \frac{Cov_{xy}}{SD_x SD_y} = \frac{20.95}{(6.1)(4.9)} = \frac{20.95}{29.89} = +0.70$$

$$r_{xy} = \frac{Cov_{xy}}{SD_x SD_y}$$

Correlation Coefficients Interpretation

Recall:

- perfect correlation between two variables would result in $r = +1.00$ or -1.00
- No correlation between X and Y, then $r_{xy} = 0$

$r_{xy} = +0.70$ $\rightarrow$ Imply?

we can say that

variables X and Y are related but that the relationship is not perfect.

This means that, on average, high scores on X are related to high scores on Y and vice versa.

Correlation Coefficients Interpretation

Method 2:

$$r_{xy}^2$$

$$r_{xy} = 0.7$$

$$r_{xy}^2 = 0.49 = 49\%$$

Implies → Almost half of the differences in Y are due to differences in X

Coefficient of Determination

Suppose a correlation of 0.6 exists between the number of risk factors for skin cancer (X) a patient has and the likelihood of the patient suffering from skin cancer (Y)

$$r_{xy} = 0.6$$

What does it imply?

$$r_{xy}^2 = 0.36$$

36% of the risk is due to these factors, but 64% is due to other

unknown factors

Home Activity

Supply a verbal description for each of the following correlations.

- (a) an r of $-.84$ between total mileage and automobile resale value
- (b) an r of $-.35$ between the number of days absent from school and performance on a math achievement test
- (c) an r of $.03$ between anxiety level and college GPA
- (d) an r of $.56$ between age of schoolchildren and reading comprehension

Exercise

Couples who attend a clinic for first pregnancies are asked to estimate (independently of each other) the ideal number of children. Given that X and Y represent the estimates of females and males, respectively, the results are as follows:

COUPLE	X	Y
A	1	2
B	3	4
C	2	3
D	3	2
E	1	0
F	2	3

Exercise

Couples who attend a clinic for first pregnancies are asked to estimate (independently of each other) the ideal number of children. Given that X and Y represent the estimates of females and males, respectively, the results are as follows:

COUPLE	X	Y
A	1	2
B	3	4
C	2	3
D	3	2
E	1	0
F	2	3

$$r = 0.65$$

r -value of 0.65 indicates a moderate to strong positive linear relationship between the estimates of the ideal number of children provided independently by females and males attending the clinic for first pregnancies.

- **Strength of Relationship:** The correlation coefficient (r) of 0.65 suggests a moderate to strong positive linear relationship between the estimates given by the females and males.
- **Positive Correlation:** The positive value of the correlation coefficient indicates that as one set of estimates (either from females or males) tends to increase, the other set also tends to increase. In this scenario, as the ideal number of children estimated by females tends to be higher, males also tend to estimate a higher ideal number of children.
- **Direction of Relationship:** The positive sign indicates that both estimates move in the same direction; when one estimate increases, the other estimate tends to increase as well.

Exercise

Given the following data, what are the correlation and the coefficient of determination between the following?

Student	IQ (X)	EQ (Y)	Exam Scores (Z)
1	140	14	42
2	130	20	44
3	120	29	35
4	119	6	30
5	115	20	23
6	114	27	27
7	114	29	25
8	113	30	20
9	112	35	16
10	111	40	12

Exercise

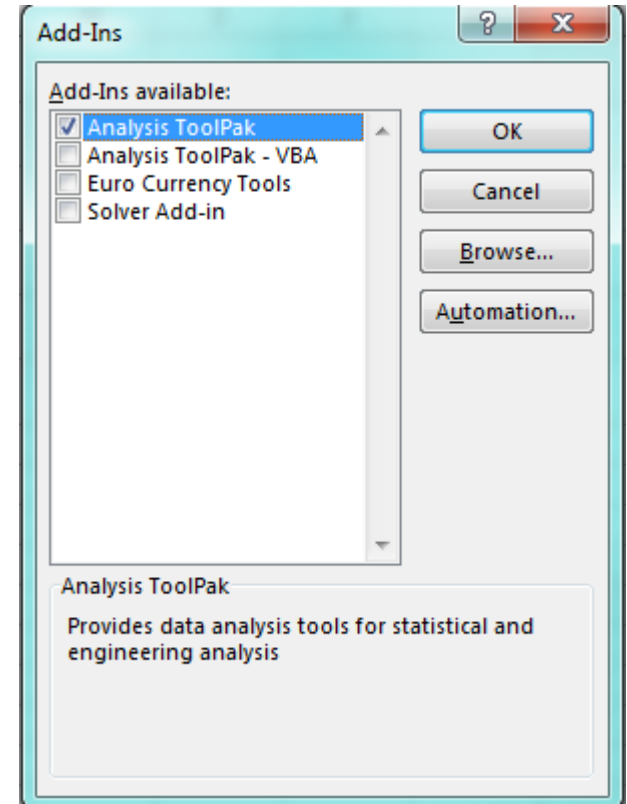
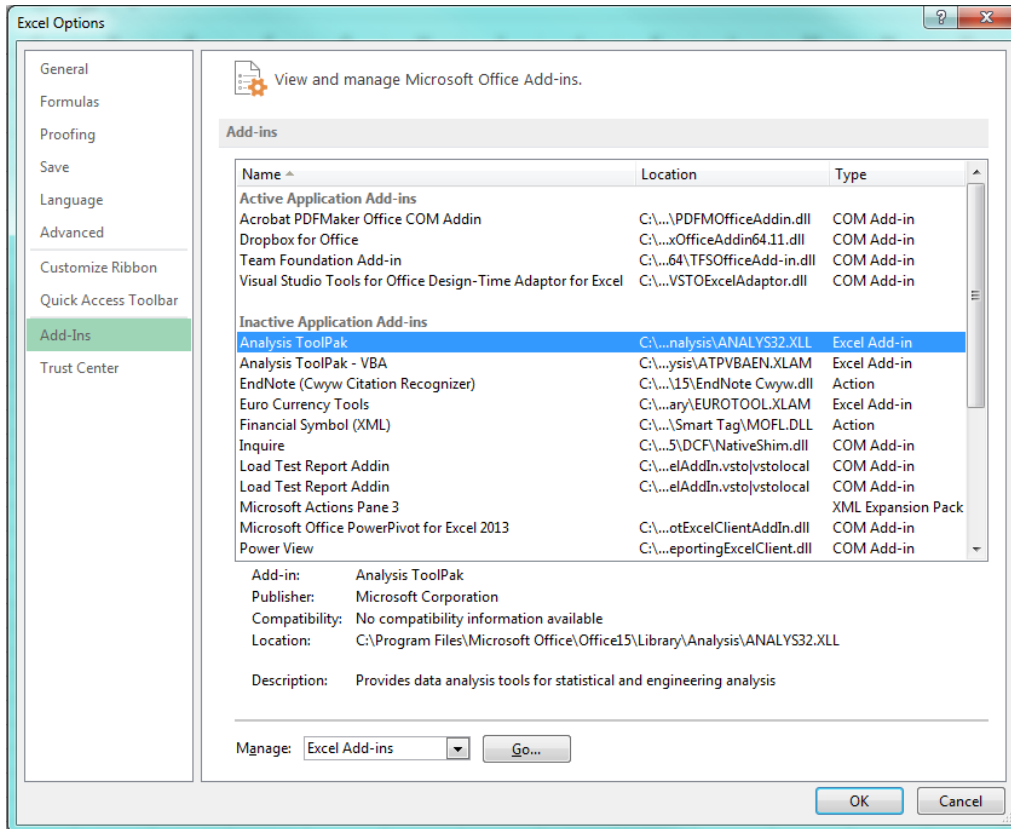
Given the following data, what are the correlation and the coefficient of determination between the following?

- a) Scores on tests of intelligence (IQ) and emotional intelligence (EQ)
- b) IQ scores and exam scores
- c) EQ test scores and exam scores

In addition, how would you interpret the size of each of these correlations?

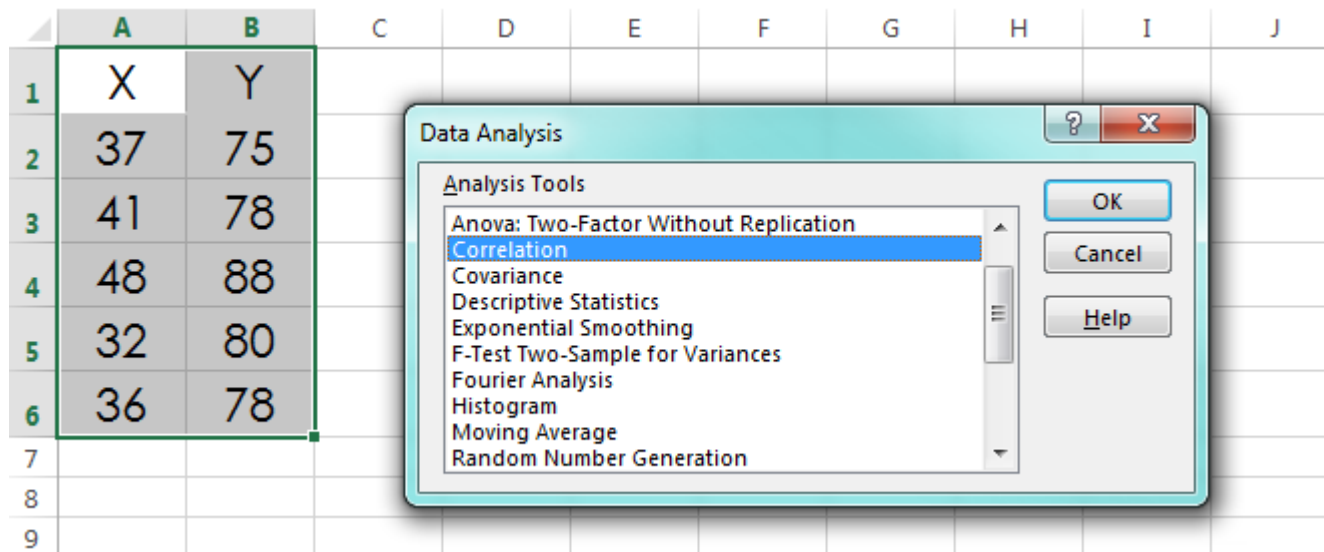
Correlations in Excel

- To do any complex calculations in Excel requires the Analysis ToolPak
 - This is not part of the standard Excel build and has to be added in through the Excel 'Options' dialogue



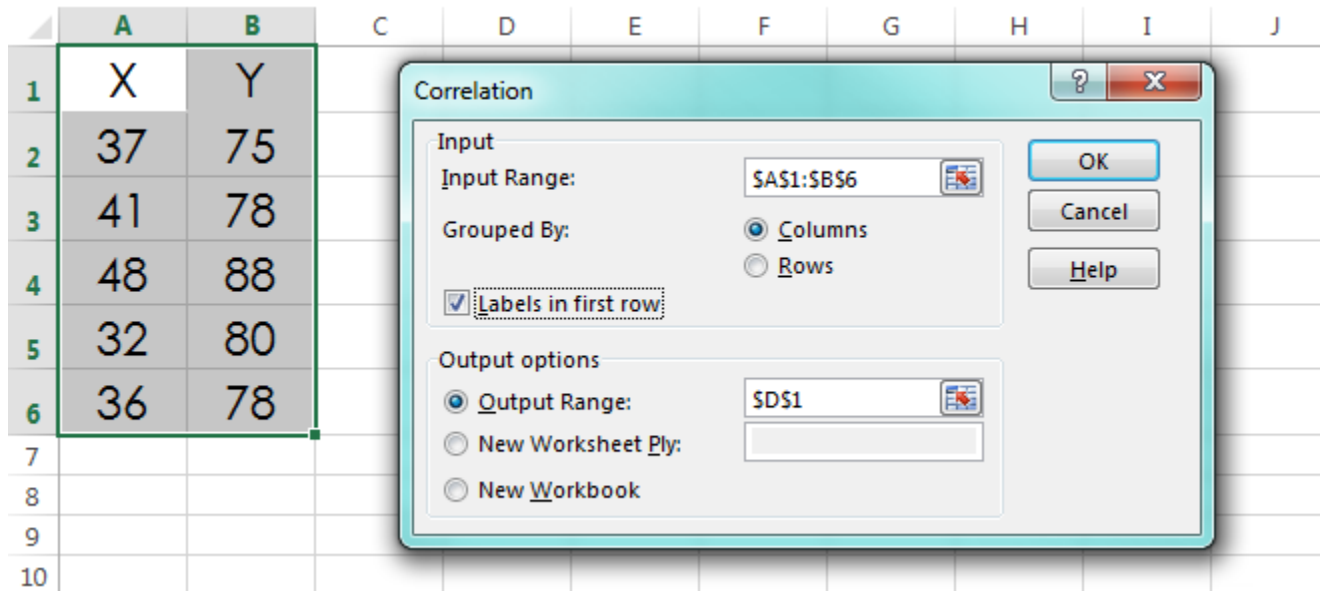
Correlations in Excel

- Once the Analysis ToolPak has been added it will appear at the end of the Data tab of the ribbon
- Select 'Correlation' from the data analysis tool



Correlations in Excel

- Input the data range for analysis, input a location for the data to be output, and tick the labels box if appropriate



Correlations in excel

	A	B	C	D	E	F	G
1	X	Y			X	Y	
2	37	75		X	1		
3	41	78		Y	0.702981	1	
4	48	88					
5	32	80					
6	36	78					

On completion, Excel calculates the Pearson's correlation for the data

Summary

Topics covered

- Correlation Coefficient
- Pearson Correlation Coefficient
- Interpreting Correlation Coefficients

Next week we will cover

- Inferential Statistics
- Regression Equations

References

- Chapter 8 - *Statistics for the Terrified*.