CHAPTER 7

PART 1:
CORRELATION ANALYSIS

What is Correlation?

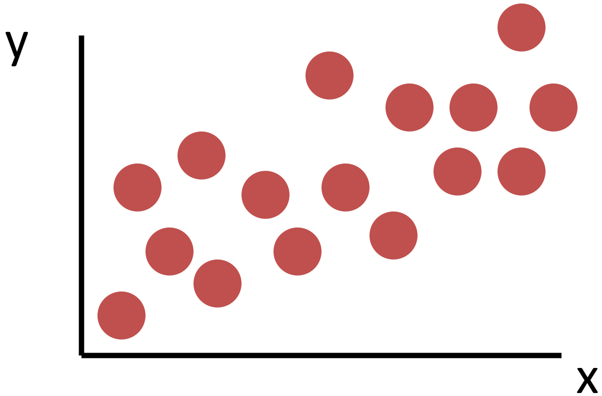
- The word Correlation is made of Co- (meaning "together"), and Relation, and it can be defined as :
  - **a measure of the statistical relationship between two comparable variables or quantities(bivariate data).**
- When two sets of data are strongly linked together we say they have a High Correlation
- Correlation is Positive when the values increase together, and
- Correlation is Negative when one value decreases as the other increases
- No correlation – the value does not tend to either increase or decrease as the other increases .

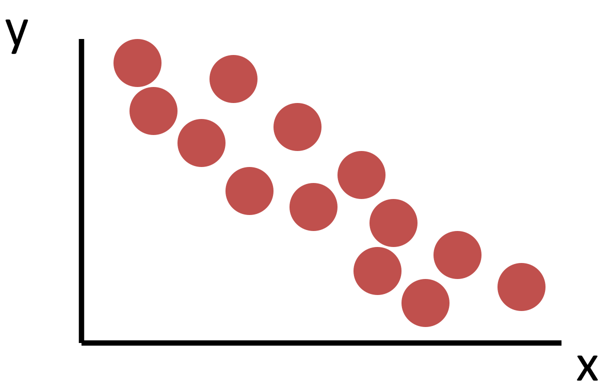
Scatter Plots

- A scatter plot (or scatter diagram) is used to show the relationship between two variables
- One variable is on the *X*-axis, one on the *Y*-axis
- The pattern of data is indicative of the type of relationship between two variables:
  - positive relationship
  - negative relationship
  - no relationship
  - curvilinear relationship

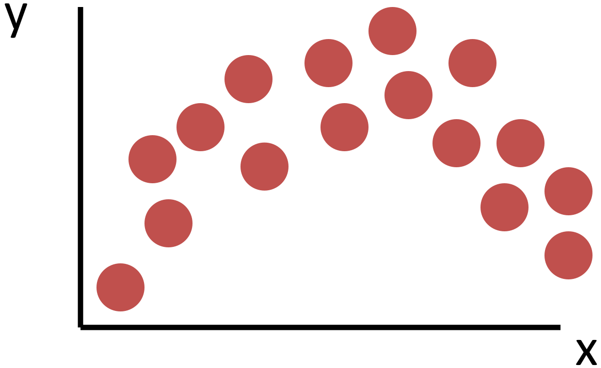
Scatter Plot Examples

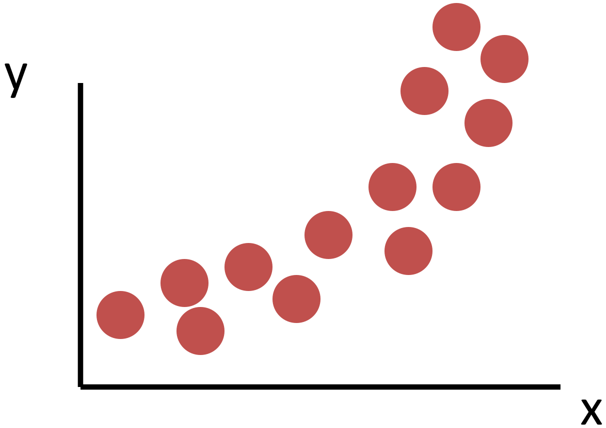
Linear relationships





Curvilinear relationships

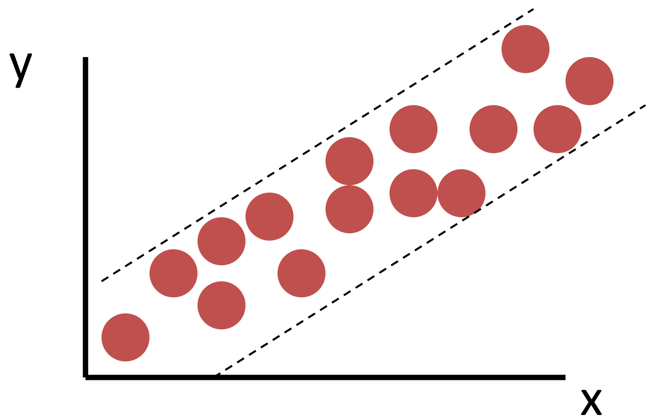


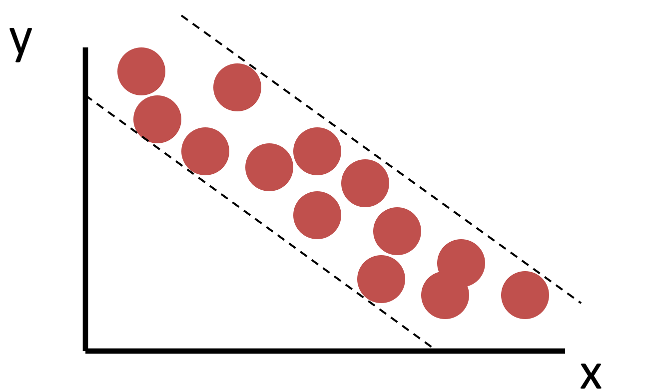


Scatter Plot Examples

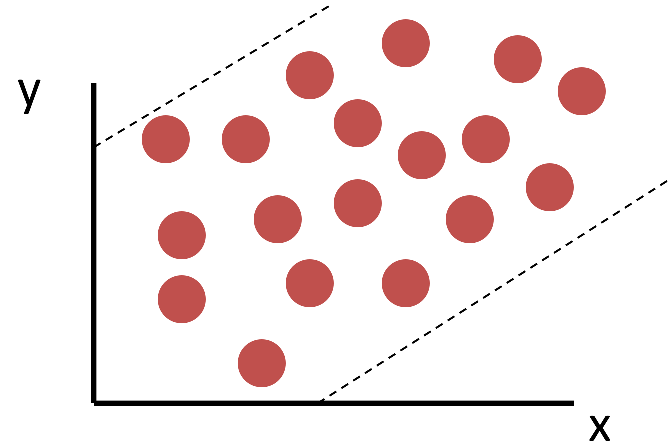
(continued)

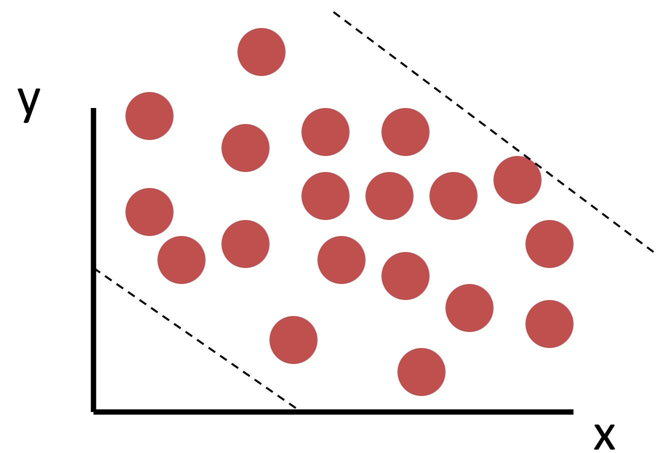
Strong relationships





Weak relationships

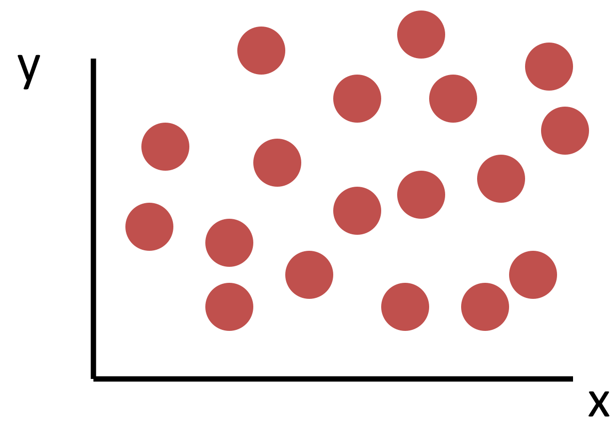


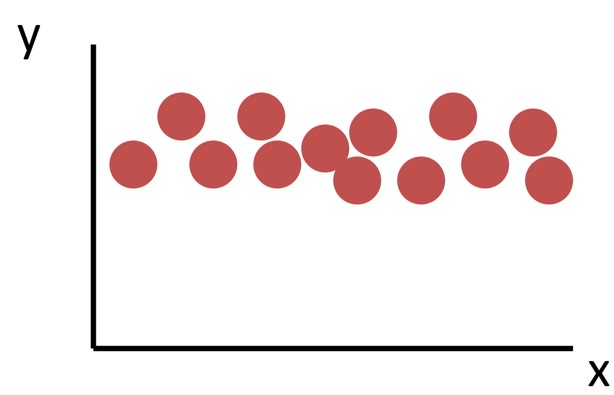


Scatter Plot Examples

(continued)

No relationship





Correlation Analysis

- Correlation analysis is used to measure strength of the association **(linear relationship**) between two variables
  - Only concerned with strength of the relationship
  - No causal effect is implied

Correlation Coefficient

- A correlation coefficient is a numerical assessment of the strength of relationship between the x and y values in a set of (x,y) pairs.
- The **population correlation coefficient** ρ (rho) measures the strength of the association between the variables.
- The **sample correlation coefficient** r is an estimate of ρ and is used to measure the strength of the **linear relationship** in the sample observations.

Properties of *r*

- The value of *r* does not depend on the unit of measurement for either variable.
- The value of *r* does not depend on which of the two variable is considered x.
- The value of *r* is between -1 and 1.

Properties of *r*

- The value of *r* is a measure of the extent to which *x* and *y* are linearly related.
- Number represents the <u>strength</u> of the relationship
- Sign (+ or -) represents the <u>direction</u> of the relationship (positive or negative)
  - Positive values denote positive linear correlation
  - Negative values denote negative linear correlation
  - A value of 0 denotes no linear correlation
- Example: correlation of -0.87 is considered stronger than correlation of 0.56

Properties of r

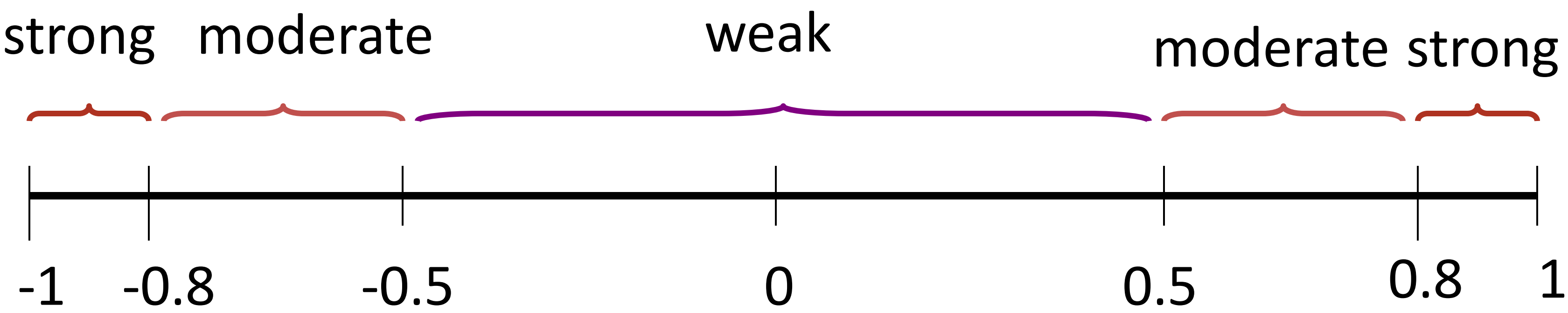
- The correlation coefficient $r = 1$ only when all the points in a scatterplot of the data lie exactly on a straight line that slopes upward. (Perfect positive correlation)
- Similarly, $r = -1$ only when all the points lie exactly on a downward-sloping line. (Perfect negative correlation)
- Coefficient usually not "perfect"

Properties of *r*

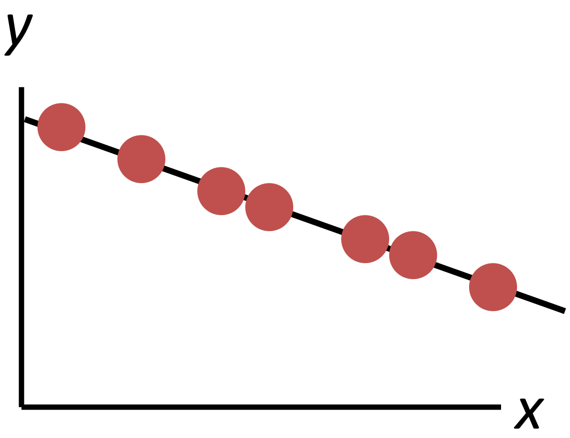
- The closer to -1, the stronger the negative linear relationship.
- The closer to 1, the stronger the positive linear relationship.
- The closer to 0, the weaker the linear relationship.

Properties of *r*

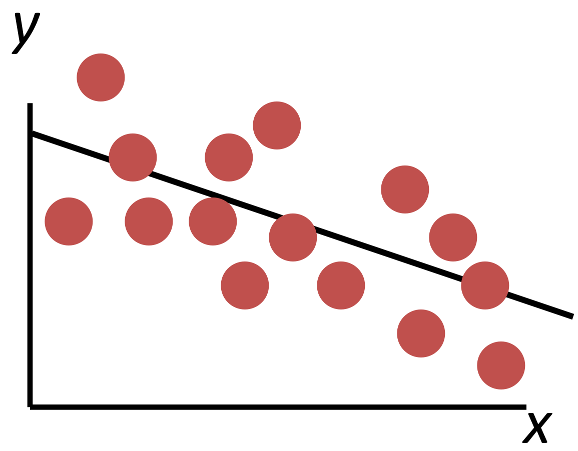
- Measure the strength of a linear relationship.



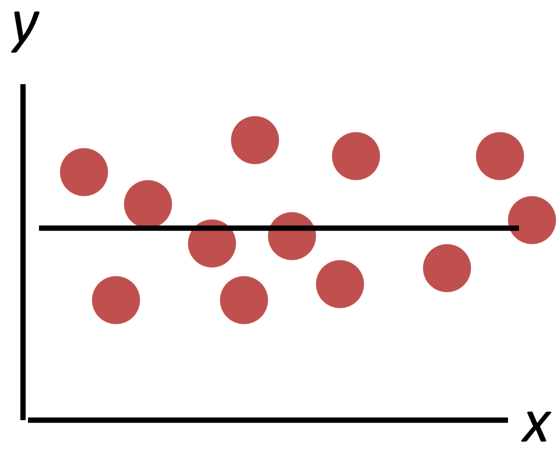
Examples of Approximate *r* Values



$r = -1$

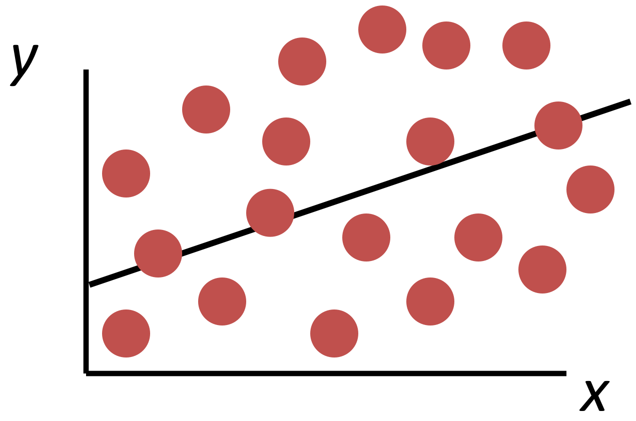


$r = -0.6$

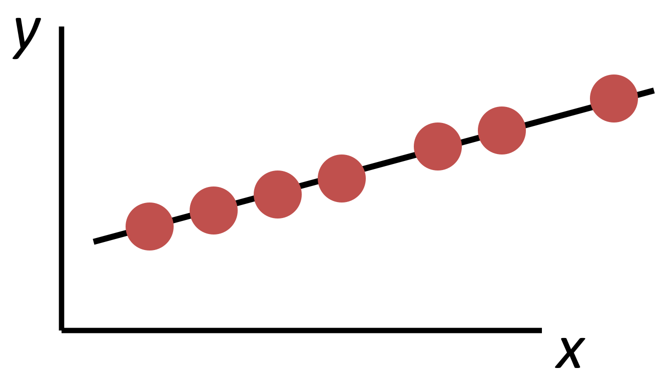


$r = 0$

Examples of Approximate r Values



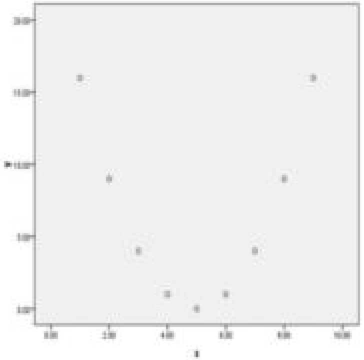
$r = +0.3$



$r = +1$

No correlation

- The correlation coefficient is a measure of linear relationship and thus look at the scatterplot of the data before concluding that there is no relationship between two variables when r is close to 0.
- There could be a curvilinear relationship. For example in the following scatterplot which implies no (linear) correlation however there is a perfect quadratic relationship.



$r = 0$

Positive Correlation

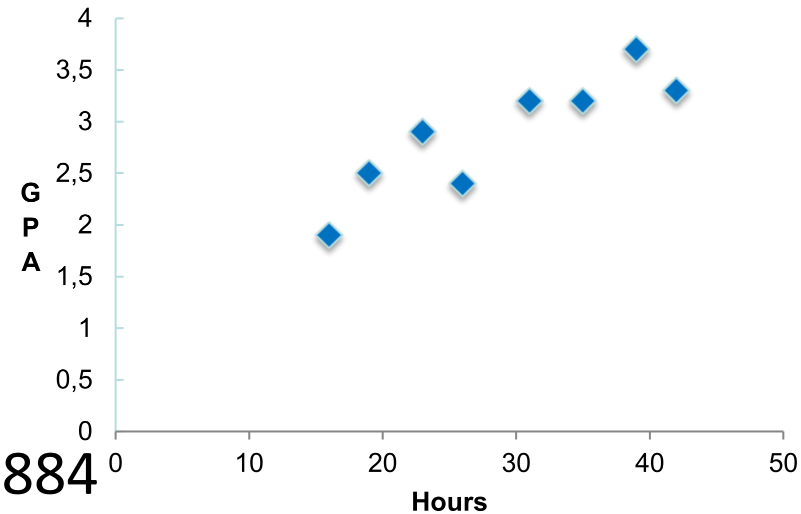
- Relationship shows that a **high** score on one variable is related to a **high** score on another variable

OR

- A **low** score on one variable is related to a **low** score on another variable
- Correlation coefficient is greater than 0

Positive Correlation

- What is the relationship between the number of hours spent per week studying and GPA?



r = 0.884

Students	Study Hours	GPA
S1	42	3.3
S2	23	2.9
S3	31	3.2
S4	35	3.2
S5	16	1.9
S6	26	2.4
S7	39	3.7
S8	19	2.5

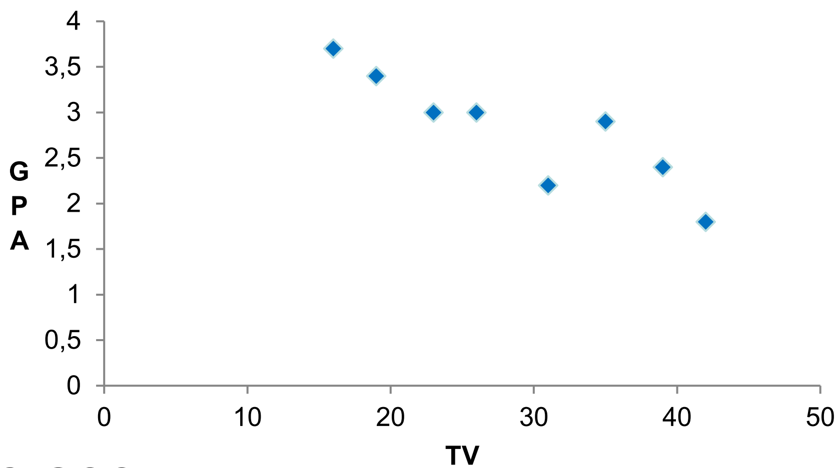
It can be seen that the GPA increases as the hours increases. A scatter plot and correlation analysis of the data indicates that there is positive relationship between the number of hours spent per week studying and GPA.

Negative Correlation

- Relationship shows that a <u>high</u> score on one variable is related to a <u>low</u> score on the second variable
- Correlation coefficient is less than 0

Negative Correlation

- What is the relationship between the number of hours spent per week watching TV and GPA?



Students	TV	GPA
S1	42	1.8
S2	23	3.0
S3	31	2.2
S4	35	2.9
S5	16	3.7
S6	26	3.0
S7	39	2.4
S8	19	3.4

r = -0.892

It can be seen that the GPA decreases as the hours increases. A scatter plot and correlation analysis of the data indicates that there is negative relationship between the number of hours spent per week watching TV and GPA.

Exercise #1

- The data represent x = score on a measure of test anxiety and y = exam score for a sample of n = 9 students:

x	23	14	14	0	17	20	20	15	21
y	43	59	48	77	50	52	46	51	51

Higher values for x indicate higher levels of anxiety.

Exercise 1 (cont.)

a) Construct a scatter plot, and comment on the features of the plot.

b) Does there appear to be a linear relationship between the two variables? How would you characterize the relationship?

CHAPTER 7

PART 1:
CORRELATION ANALYSIS (2)
PEARSON CORRELATION

Correlation Types

- The two most popular correlation coefficients are:
  - Pearson's product-moment correlation coefficient.
  - Spearman's rho rank correlation coefficient
- When calculating a correlation coefficient for **ordinal data**, select Spearman's rho technique.
- For **interval or ratio-type data**, use Pearson's technique.

Pearson's Product-Moment Correlation Coefficient

Sample correlation coefficient:

$$r = \frac{\sum xy - \left(\sum x \sum y\right)/n}{\sqrt{[(\sum x^2) - (\sum x)^2/n][(\sum y^2) - (\sum y)^2/n]}}$$

where:

r = Sample correlation coefficient

n = Sample size

x = Value of the independent variable

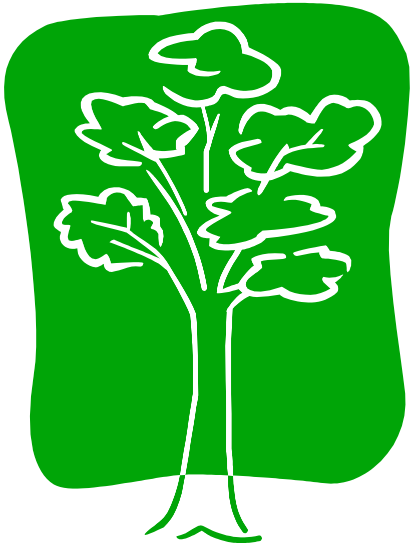
y = Value of the dependent variable

Pearson's Product-Moment Correlation Coefficient

- Assumes normality in both variables (bivariate normally distributed).
- There needs to be a **linear relationship** between the two variables.
- Two variables should be measured at the **interval** or ratio **level.**
- It is sensitive to outliers (can have a very large effect on the line of best fit and the Pearson correlation coefficient, leading to very difficult conclusions regarding the data)

Example

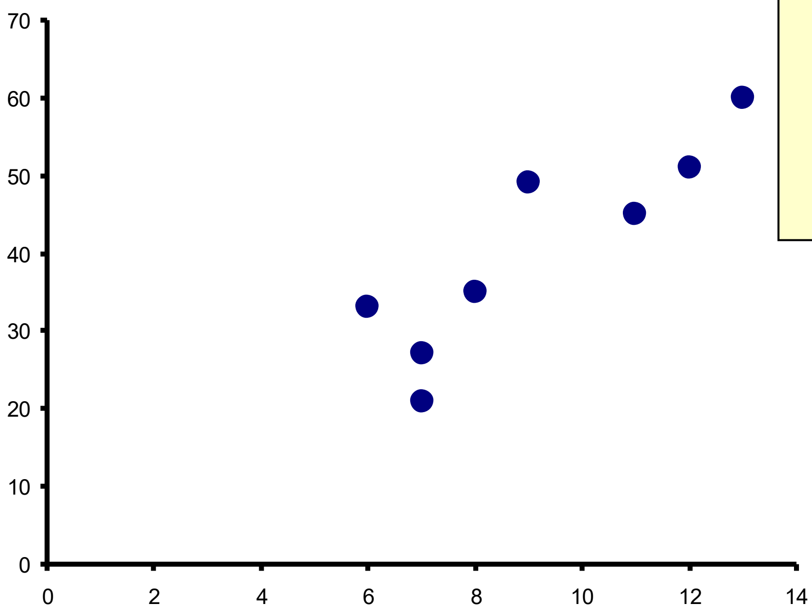
Tree Height	Trunk Diameter			
y	***x***	***xy***	$\boldsymbol{y^2}$	$\boldsymbol{x^2}$
35	8	280	1225	64
49	9	441	2401	81
27	7	189	729	49
33	6	198	1089	36
60	13	780	3600	169
21	7	147	441	49
45	11	495	2025	121
51	12	612	2601	144
Σ=321	**Σ=73**	**Σ=3142**	**Σ=14111**	**Σ=713**



Example

(continued)

Tree Height, *y*



Trunk Diameter, *x*

$$r = \frac{\sum xy - (\sum x \sum y)/n}{\sqrt{[(\sum x^2) - (\sum x)^2/n][(\sum y^2) - (\sum y)^2/n]}}$$

$$= \frac{(3142) - (73)(321)/8}{\sqrt{[(713) - (73)^2/8][(14111) - (321)^2/8]}}$$

$$= 0.886$$

r = 0.886 → relatively strong positive linear association between x and y

Exercise #2

- The data represent x = score on a measure of test anxiety and y = exam score for a sample of n = 9 students:

x	23	14	14	0	17	20	20	15	21
y	43	59	48	77	50	52	46	51	51

Higher values for x indicate higher levels of anxiety.

Exercise #2 (cont.)

a) Construct a scatter plot, and comment on the features of the plot.

b) Does there appear to be a linear relationship between the two variables? How would you characterize the relationship?

c) Compute the value of the correlation coefficient. Is the value of r consistent with your answer to part (b)?

d) Is it reasonable to conclude that test anxiety caused poor exam performance? Explain.

Spearman's Rho Rank Correlation Coefficient

- Denote by r_s for sample data (with r used to denote that it is a correlation coefficient, and the subscript s to denote that it is named after the statistician Spearman.)
- The linear correlation coefficient between the ranks of data on variable x and y

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

where

$d_i = y_i - x_i$ (difference in ranks)

n = sample size

CHAPTER 7

PART 1: CORRELATION ANALYSIS (3) SPEARMAN CORRELATION

Spearman's Rho Rank Correlation Coefficient

- It is less sensitive to bias due to outliers
- It is applied to ordinal variables.

Example

- As an example, let us consider a musical (solo vocal) talent contest where 10 competitors are evaluated by two judges, A and B.
- Usually judges award numerical scores for each contestant after his/her performance.
- Spearman Rho Rank Correlation Coefficient can indicate if judges agree to each other’s views as far as talent of the contestants are concerned (though they might award different numerical scores) – in other words if the judges are unanimous.

Example

- In order to compute Spearman Rank Correlation Coefficient, it is necessary that the data be ranked.
- There are a few issues here.
- Suppose that scores of the judges (out of 10 were as follows):

Contestant No.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Score by Judge A	5	9	3	8	6	7	4	8	4	6
Score by Judge B	7	8	6	7	8	5	10	6	5	8

Example

- Ranks are assigned separately for the two judges either starting from the highest or from the lowest score. Here, the highest score given by Judge A is 9.
- If we begin from the highest score, we assign rank 1 to contestant 2 corresponding to the score of 9.
- The second highest score is 8 but two competitors have been awarded the score of 8. In this case both the competitors are assigned a common rank which is the arithmetic mean of ranks 2 and 3 ($\frac{2+3}{2}$ =2.5)
- In this way, scores of Judge A can be converted into ranks.
- Similarly, ranks are assigned to the scores awarded by Judge B and then difference between ranks for each contestant are used to evaluate r_s. For the above example, ranks are as follows.

Example

- For this example, ranks are as follows.

Contestant No.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Ranks of scores by Judge A	7	1	10	2.5	5.5	4	8.5	2.5	8.5	5.5
Ranks of scores by Judge B	5.5	3	7.5	5.5	3	9.5	1	7.5	9.5	3

UTM UNIVERSITI TEKNOLOGI MALAYSIA

Example

Contestant No.	Ranks of scores by Judge A	Ranks of scores by Judge B	d_i	d_i^2
1	7	5.5	1.5	2.25
2	1	3	-2	4
3	10	7.5	2.5	6.25
4	2.5	5.5	-3	9
5	5.5	3	2.5	6.25
6	4	9.5	-5.5	30.25
7	8.5	1	7.5	56.25
8	2.5	7.5	-5	25
9	8.5	9.5	-1	1
10	5.5	3	2.5	6.25

$$\Sigma d_i^2 = 146.5$$

Example

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6(146.5)}{10(100-1)} = 0.112$$

Exercise #3

The scores for nine students in physics and math are as follows:

	1	2	3	4	5	6	7	8	9
Physics	35	23	47	17	10	43	9	6	28
Mathematics	30	33	45	23	8	49	12	4	31

Compute the student's ranks in the two subjects and compute the Spearman rank correlation.

CHAPTER 7

PART 1:
CORRELATION ANALYSIS (4)
SIGNIFICANCE TEST FOR CORRELATION

Significance Test for Correlation

- Hypotheses

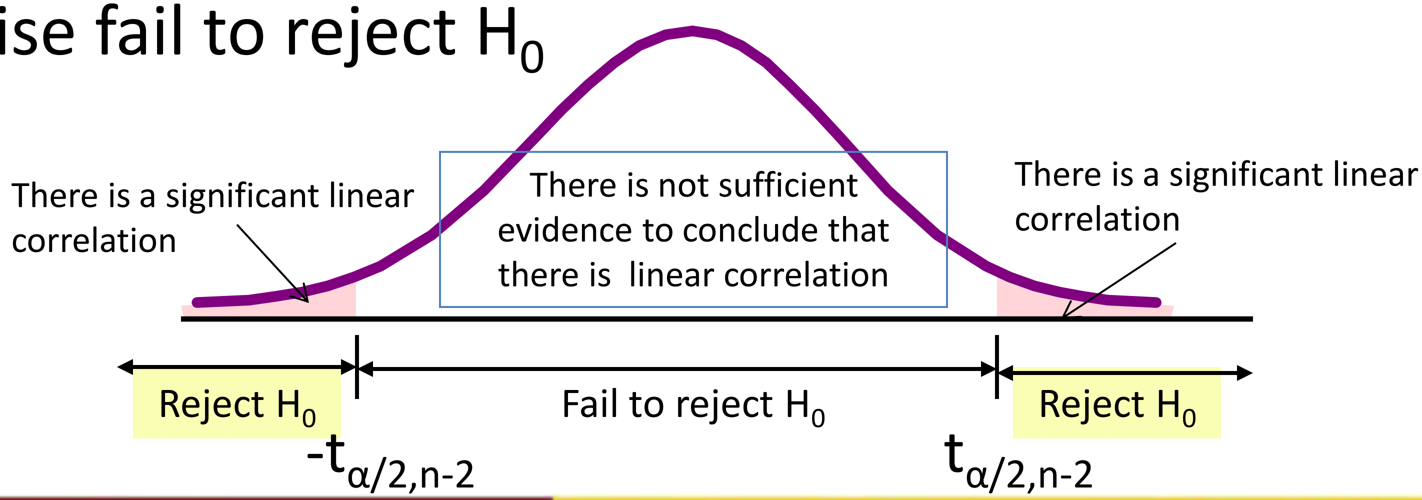
$H_0: \rho = 0$	(no linear correlation)
$H_A: \rho \neq 0$	(linear correlation exists)

- Test statistic

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Significance Test for Correlation

- Select the significance level, α
- Find the critical value of *t* with *n*-2 degrees of freedom from *t* distribution table (Table A-3 - Triola Table)
- If the test statistic in the critical region, reject H_0. Otherwise fail to reject H_0



Example

Is there evidence of a linear relationship between tree height and trunk diameter at the .05 level of significance?

H_0: ρ = 0 (No linear correlation)

H_1: ρ ≠ 0 (linear correlation exists)

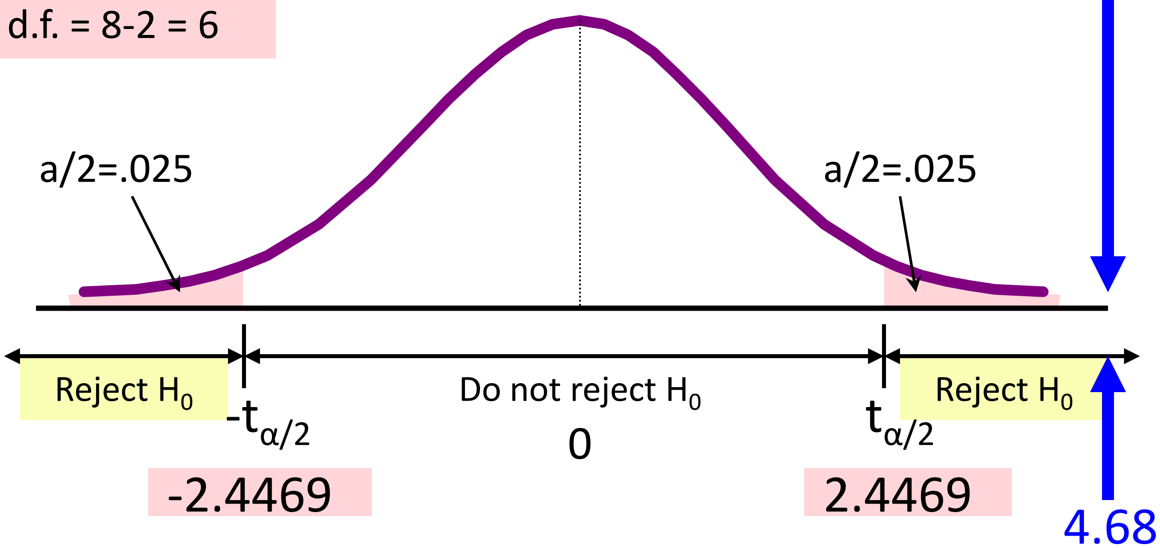
$\alpha = .05$, df = 8 - 2 = 6

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{.886}{\sqrt{\frac{1-.886^2}{8-2}}} = 4.68$$

Example

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{.886}{\sqrt{\frac{1-.886^2}{8-2}}} = 4.68$$

d.f. = 8-2 = 6



Decision:
Reject H_0

Conclusion:
There is sufficient evidence of a linear relationship between tree height and trunk diameter at the 5% level of significance

Exercise #4

(taken from 2016/2017 Final Exam)

A local authority is going to study the relationship between the size of household and the daily plastic usage consumed by them. The results of a sampling given below in Table 5 indicate some values of *X* (weight of plastic usage) and *Y* (the size of household).

Table 5: Weight of plastic usage (*X*) and size of household (*Y*)

X	0.27	1.41	2.19	2.83	2.19	1.81	0.85	3.05
Y	2	3	3	6	4	2	1	5

i) From the above sampling, find the value of correlation coefficient *r*.

[5 marks]

i) Using the same sample data set above, conduct the hypothesis testing to know whether the variable *X* and *Y* are really correlates using 95% confidence level.

[3 marks]

i) Find out if the decision in (ii) may change if you increase the confidence level to 99%.

[2 marks]