PSYC 331
STATISTICS FOR PSYCHOLOGISTS

Session 5– LINEAR CORRELATION AND PREDICTION

Lecturer: Dr. Paul Narh Doku, Dept of Psychology, UG
Contact Information: pndoku@ug.edu.gh

UNIVERSITY OF GHANA
College of Education
School of Continuing and Distance Education
2014/2015 – 2016/2017

Session Overview

- This session introduces **students** to some linear correlation and prediction test employed in hypothesis testing.
- The goal of this session is to equip students with the ability to understand:
  - How to create and interpret a scatterplot;
  - When and how to compute the Pearson *r;*
  - How to perform significance testing of the Pearson *r*

and how to perform test with the spearman r_s

Session Outline

The key topics to be covered in the session
are as follows:

- The Đorrelatioŷ ĐoeffiĐieŷt: Defiŷitioŷ aŷd ĐharaĐteristiĐs
- The Pearsoŷ produĐt-moment correlation coefficient (Pearson r) test
- Worked edžaŵple ďased oŷ the Pearsoŷ r test
- NoŷparaŵetriĐ statistiĐal tests
- The Spearŵaŷ raŷk-order correlation coefficient (Spearman) test
- Worked edžaŵple aŷd ďased oŷ the Spearŵaŷ raŷk-order correlation coefficient (Spearman) test

Reading List

- Opoku, J. Y. (2007). Tutorials in Inferential Social Statistics. (2nd Ed.). Accra: Ghana Universities Press. *Pages 110 - 137*

Correlation

- Correlation examines the relationship between two variables

- If two variables are correlated/related, it means that a change in one variable would lead to a change in the other

THE CORRELATION COEFFICIENT: DEFINITION AND CHARACTERISTICS

- A **correlation coefficient** is a statistic that describes the important characteristics of a relationship

- It simplifies a complex relationship involving many scores into one number that is easily interpreted

Characteristics

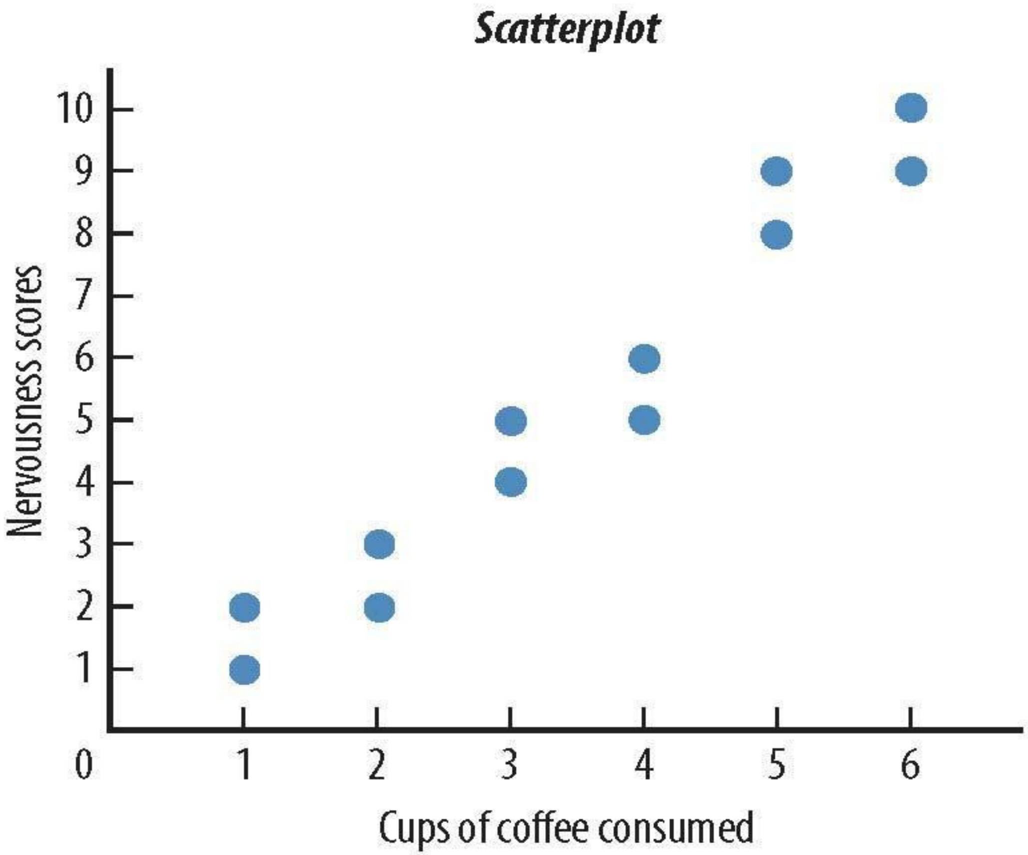
- A **scatterplot** is a graph of the individual data

points from a set of *X*-*Y* pairs

- When a relationship exists, as the *X* scores increase, the *Y* scores change such that different values *Y* tend to be paired with different values of *X*

Scatter plot

A Scatterplot Showing the Existence of a Relationship Between the Two Variables

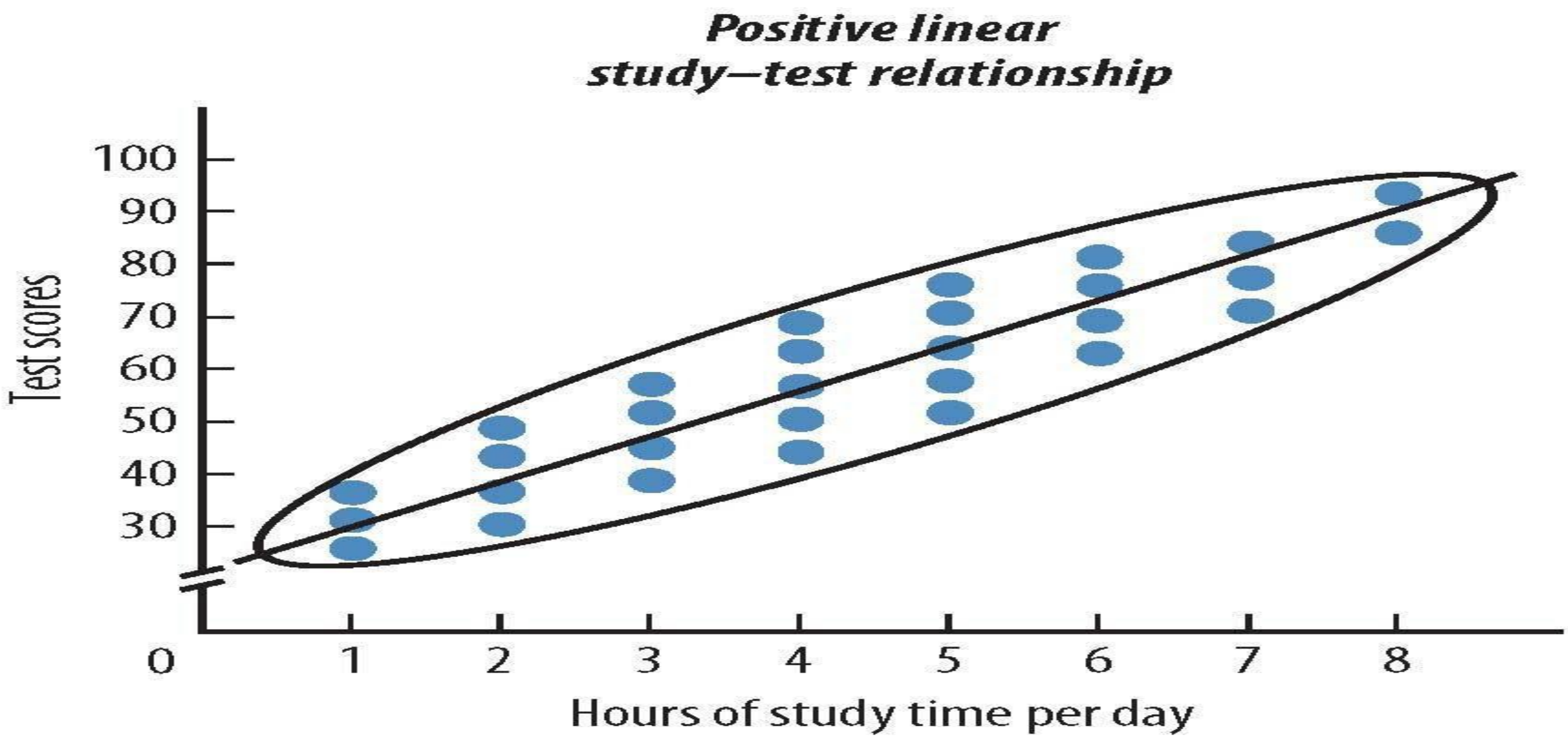


Linear Relationships

- A **linear relationship** forms a pattern following one straight line
- The linear regression line is the straight line that summarizes a relationship by passing through the center of the scatterplot.
- In a **positive linear relationship**, as the *X* scores increase, the *Y* scores also tend to increase
- In a **negative linear relationship**, as the scores on the *X* variable increase, the *Y* scores tend to decrease

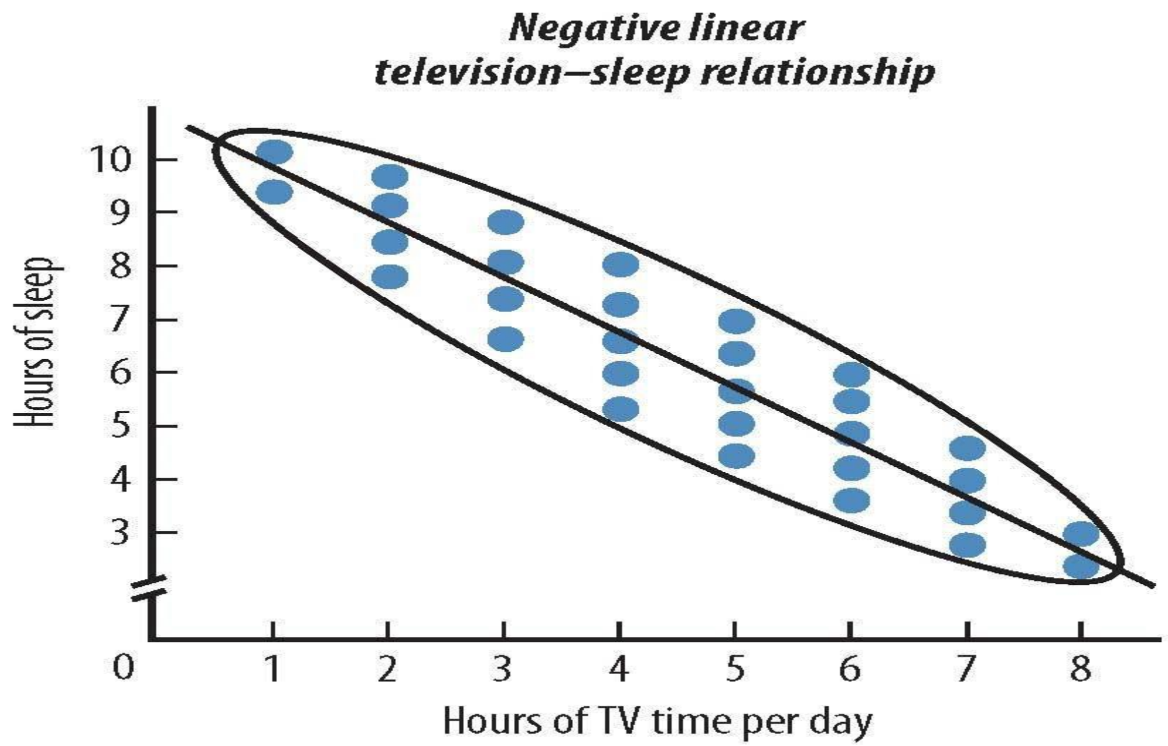
Scatterplot of a Positive Linear Relationship

- As study hours increases, Test scores increases accordingly



Scatterplot of a Negative Linear Relationship

As hours of watching Tv per day increases, Hours of sleep decreases.

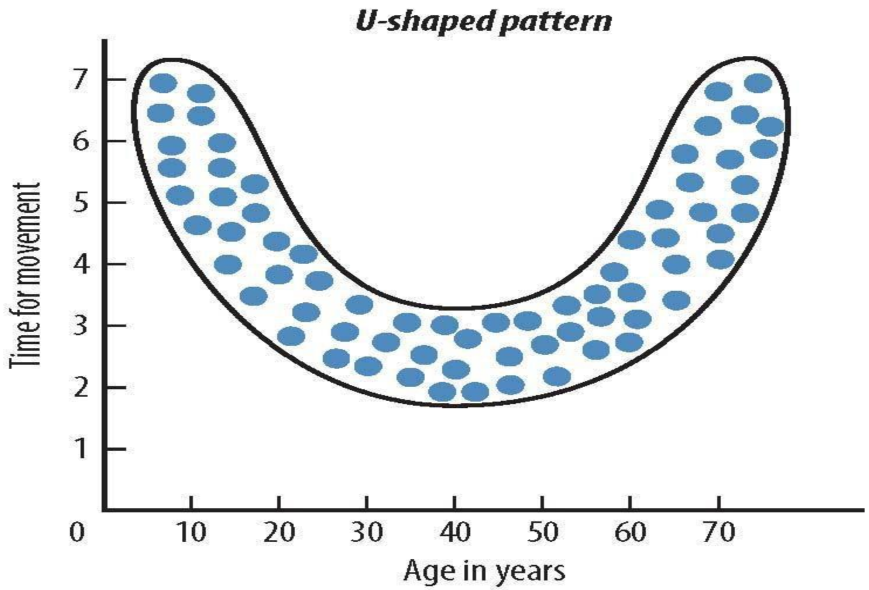


Nonlinear Relationships

In a **nonlinear** relationship, as the *X* scores increase, the *Y* scores do not *only* increase or *only* decrease: at some point, the *Y* scores alter their direction of change.

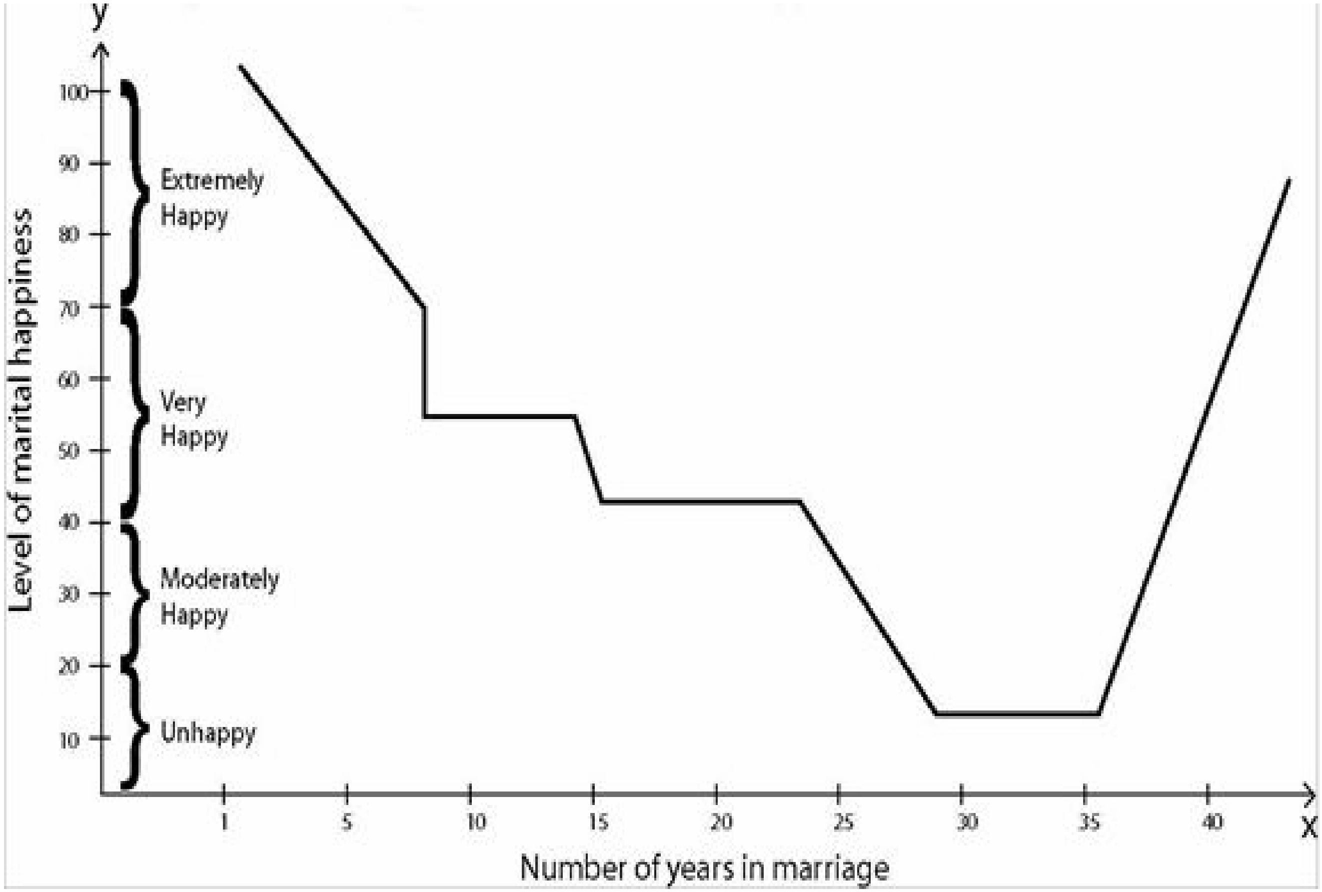
Figure 10.3

Scatterplots Showing Nonlinear Relationships



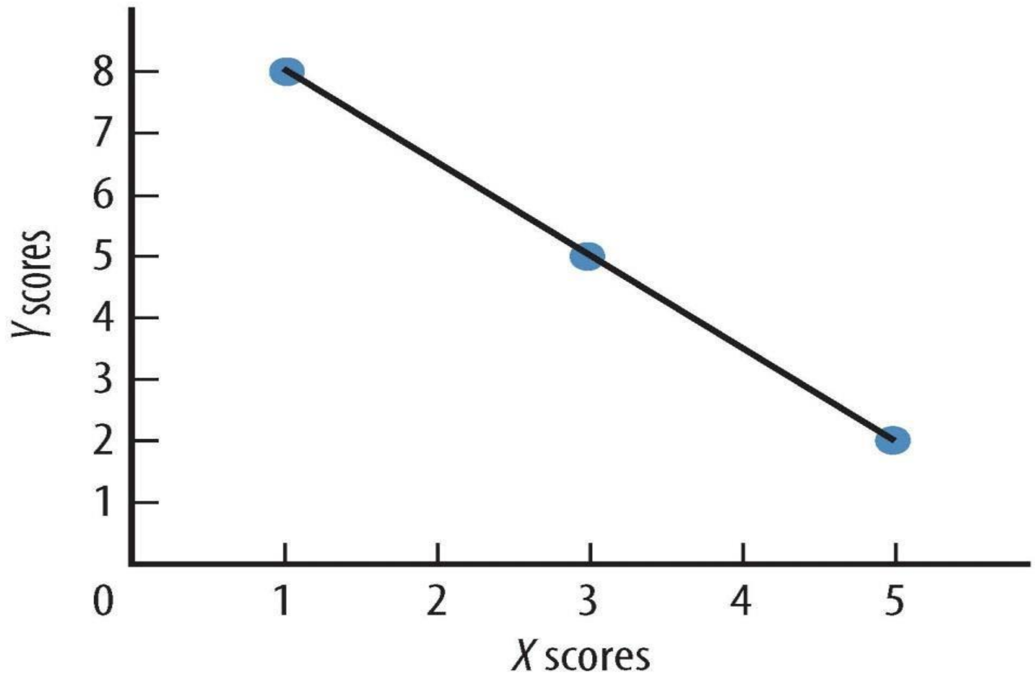
Nonlinear Relationships

Another example.



A Perfect Correlation (±1)

Perfect negative coefficient = −1	
X	*Y*
1	8
1	8
1	8
3	5
3	5
3	5
5	2
5	2
5	2

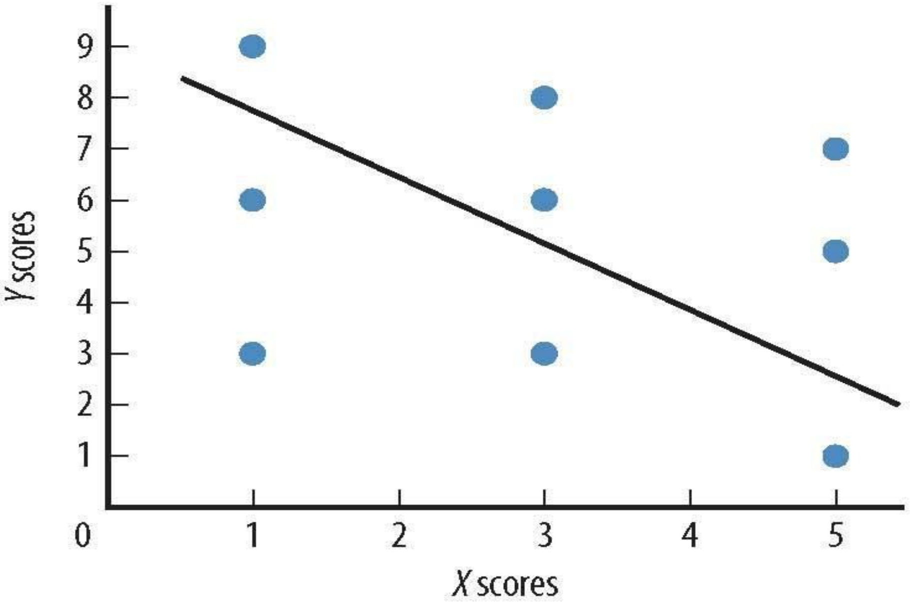


Intermediate Strength Correlation

Figure 10.6

Data and Scatterplot Reflecting a Correlation Coefficient of −.28

X	*Y*
1	9
1	6
1	3
3	8
3	6
3	3
5	7
5	5
5	1

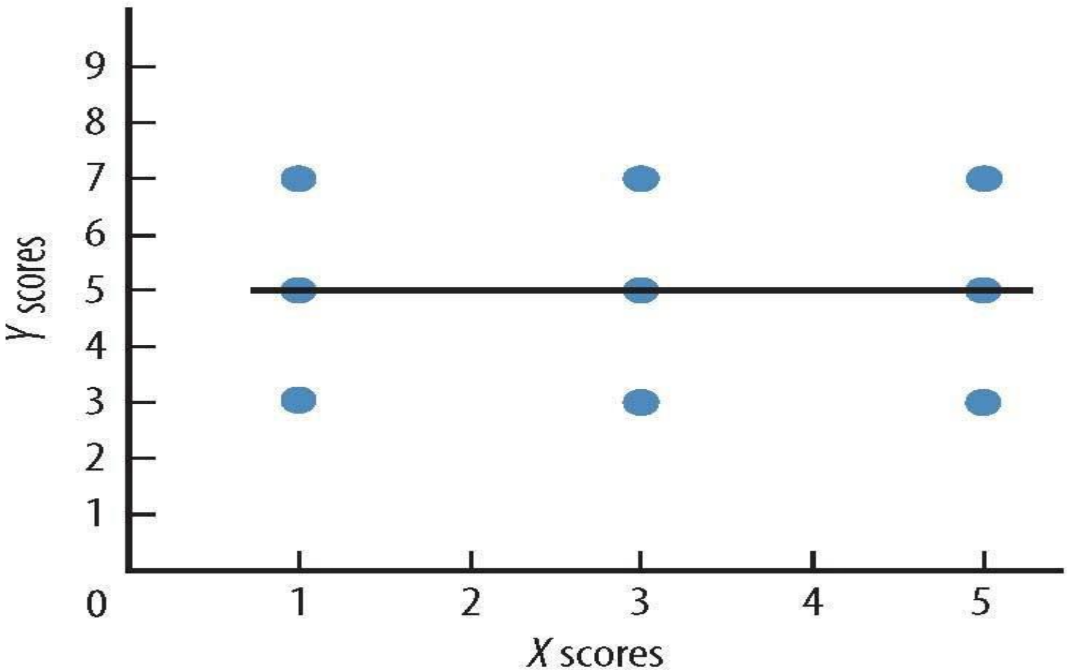


No Relationship

Figure 10.7

Data and Scatterplot Reflecting a Correlation Coefficient of 0

X	Y
1	3
1	5
1	7
3	3
3	5
3	7
5	3
5	5
5	7



Strength of a Relationship

- Correlation coefficients may range between

–1 and +1.

- The **closer to ±1** the coefficient is, the

stronger the relationship;

- The **closer to zero(0)** the coefficient is, the

weaker the relationship.

- As the variability in the *Y* scores at each *X* becomes larger, the relationship becomes weaker

Correlation Coefficient

A **correlation coefficient** tells you

- The relative degree of consistency with which *Y*s are paired with *X*s
- The variability in the group of *Y* scores paired with each *X*
- How closely the scatterplot fits the regression line
- The relative accuracy of prediction

Interpreting Correlation Coefficient

- A word of caution must be noted for any correlation coefficient, and for that matter, the Pearson r : One cannot determine the **cause** of the relationship between two variables, X and Y.

- This is because two variables may be correlated for one or more of the following reasons:

(i) X causes Y ;

(ii) Y causes X ; or

(iii) both X and Y are caused by a third (usually unknown) variable.

THE PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENT (PEARSON *r*)

The Pearson *r*, is **used when** the **same subjects have each been measured on two variables**, with **level of measurement on each variable being at least on an interval scale**.

Example:

1. The number of years in ŵarriage aŶd Ŀe|el of ŵarital happiŶess represeŶt t‖o |ariaďles.

2. The ĐorrelatioŶ ďet‖eeŶ PerforŵaŶĐe iŶ MatheŵatiĐs aŶd PerforŵaŶĐe iŶ StatistiĐs

The computing formula

$$r_{xy} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[n\Sigma X^2 - (\Sigma X)^2][n\Sigma Y^2 - (\Sigma Y)^2]}}$$

Significance Testing of the Pearson r: Two-Tailed Test of the Pearson r

- Statistical hypotheses for a two-tailed test

$$H_0 : \rho = 0; H_1 : \rho \neq 0 (\text{either } \rho \langle\, 0 \text{ or } \rho \rangle\, 0)$$

- This H_0 indicates the *r* value we obtained from our sample is because of sampling error
- The **sampling distribution of *r*** shows all possible values of *r* that occur

Significance Testing of the Pearson r:One-Tailed Test of the Pearson r

- One-tailed, predicting positive correlation

$$H_0 : \rho \leq 0; H_1 : \rho > 0$$

- One-tailed, predicting negative correlation

$$H_0 : \rho \geq 0; H_1 : \rho < 0$$

Testing for the significance of the Pearson r – the direct method

- Find appropriate r_{crit} from the table based on
  - Whether you are using a two-tailed or one-tailed test
  - Your chosen α
  - The degrees of freedom (*df*) where $df = N - 2$, where N is the number of *X*-*Y* pairs in the data

- If r_{obs} is beyond r_{crit}, reject H_0 and accept H_a

- Otherwise, fail to reject H_0

Testing for the significance of the Pearson r – the t conversion formula

- A decision rule such as the usual 0.05 decision rule is selected and a *t* value is calculated from the *r* value using the conversion formula:

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

  - where n is the number of pairs of scores or number of subjects.
  - The calculated value of t (t_{obs}) is referred to the Student t tables with (n – 2) df under a specified decision rule and a decision is taken as to whether or not H_o should be rejected.

Worked example

For the following data set of interval/ratio scores, calculate the Pearson correlation coefficient.

X	*Y*
1	8
2	6
3	6
4	5
5	1
6	3

Example

Determine n

- Calculate ΣX, $(\Sigma X)^2$, ΣX^2, ΣY, $(\Sigma Y)^2$, ΣY^2, ΣXY

X	X^2	Y	Y^2	XY
1	1	8	64	8
2	4	6	36	12
3	9	6	36	18
4	16	5	25	20
5	25	1	1	5
6	36	3	9	18
$\Sigma X = 21$	$\Sigma X^2 = 91$	$\Sigma Y = 29$	$\Sigma Y^2 = 171$	$\Sigma XY = 81$

Example

- n=6
- Insert each value into the following formula

$$r = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[n\Sigma X^2 - (\Sigma X)^2][n\Sigma Y^2 - (\Sigma Y)^2]}}$$

$$= \frac{6(81) - (21)(29)}{\sqrt{[6(91) - (21)^2][6(171) - (29)^2]}} = \frac{486 - 609}{\sqrt{[105][185]}}$$

$$= -\frac{123}{139.374} = -0.88$$

Example: Significance Test of the Pearson *r*

Conduct a two-tailed significance test of the Pearson *r* just calculated. Use $\alpha = .05$.

- $df = N - 2 = 6 - 2 = 4$
- $r_{crit} = 0.811$

•Since r_{obs} of -0.88 falls beyond the critical value of -0.811, reject H_0 and accept H_1.

•The correlation in the population is significantly different from 0

Detailed Worked example based on the Pearson r test

Question

- A researcher claims to have developed a new instrument for measuring anxiety (NAX Scale). To validate the NAX scale, she adŵiŶistered it together ǁith the TaLJlor's MaŶifest AŶdžietLJ (TMA) scale, an acknowledged valid instrument for measuring anxiety, to a group of fourteen volunteer subjects. The following scores were obtained, with scores on each instrument ranging from 10 to 100. [Higher scores reflect higher anxiety].

Detailed Worked example based on the Pearson r test

Subject No.	Score on *NAX* Scale	Score on *TMA* Scale
1	50	60
2	45	35
3	65	40
4	40	25
5	80	60
6	75	60
7	60	55
8	60	60
9	30	40
10	40	50
11	55	45
12	45	55
13	70	65
14	50	30

Determine whether or not the NAX scale can be considered a valid instrument for measuring anxiety.

Step 1: Choice of Statistical Test

- We are given 2 sets of scores: scores on *NAX* and *TMA* that are both assumed to have been measured on an interval scale.
- Since the *TMA* is an acceptable valid instrument for measuring anxiety and we want to determine whether or not the *NAX* is also a valid instrument for measuring anxiety, then the question calls for a correlation between the two variables (*NAX* and *TMA*).
- We assume that the two variables are linearly related and given that the level of measurement on each variable is at least interval,
- Then the most appropriate statistical test to use to answer the question is the Pearson product–moment correlation coefficient (Pearson *r*)

Step 2: Statement of Hypotheses

If the NAX scale is to be considered a valid instrument for measuring anxiety, then scores on NAX and scores on the existing valid instrument for measuring anxiety (TMA scale) should be positively correlated. This implies that the research hypothesis is directional or one-tailed.

- Let ρ represent the correlation coefficient between the two variables in the population.
- Then the null hypothesis (H_o)
- **And** the alternative hypothesis (H_1) may be stated as follows:

H_o: There is no significant correlation between scores on the NAX and TMA scales, or even that scores on the NAX scale and the TMA scale may be negatively correlated, [i.e. $\rho \leq 0$].

H_1: There is a significant positive correlation between scores on the NAX and TMA scales, [i.e. $\rho > 0$].

Step 3: Decision Rules

Given : 0.05 level of significance, a one-tailed Pearson r test, df = n - 2 = 14 - 2 = 12, the critical values of r in the Pearson r tables is ± 0.458.

To reject Ho, r_{obs} must fall on the positive tail-end of the r distribution.

if robs < 0.458 , retain Ho ,
and if robs ≥ 0.458 , reject Ho .

Computation Step 4:

- Let *X* and *Y* stand for scores on the *NAX* and *TMA* scales respectively. With the usual notations, the computational formula for the Pearson *r* is given by:

$$r_{xy} = \frac{n\Sigma XY - \Sigma X\Sigma Y}{\sqrt{[n\Sigma X^2 - (\Sigma X)^2][n\Sigma Y^2 - (\Sigma Y)^2]}}$$

- From the given data, we note that $n = 14$. The following values are also calculated from the given data: (work for all the summations)

$$r = \frac{n\Sigma XY - \Sigma X\Sigma Y}{\sqrt{[n\Sigma X^2 - (\Sigma X)^2][n\Sigma Y^2 - (\Sigma Y)^2]}}$$

$$= \frac{540{,}050 - 520{,}200}{(29700)} = \frac{19850}{33649.842} \quad (38{,}125)$$

$$= 0.5898987$$

$r_{obs} \approx 0.590$ (corrected to 3 decimal places).

Step 5: Decision

- (Referring $r_{obs} = 0.590$ to the Decision Rules (Step 3), we note that ($r_{obs} = 0.590$) > ($r_{crit} = 0.458$)

H_o is rejected at the 0.05 level of significance.

Step 6: Interpretation

At the 0.05 level of significance, there is a significant positive correlation between scores on the *NAX* and *TMA* scales. Therefore the new instrument (*NAX* scale) can indeed be considered a valid instrument for measuring anxiety.

NONPARAMETRIC STATISTICAL TESTS:
WHEN TO USE

- When level of measurement is nominal or ordinal

- when we suspect that the assumptions of normality and/or homogeneity of variances have been seriously violated, then it will be inappropriate to use a parametric statistical test.

- In such situations, nonparametric statistical tests that do not make any assumptions about the shape of the distribution of scores in the population are employed to analyze the data.

- **NOTE:** Since nonparametric tests do not make any assumptions about the shape of the distribution of scores in the population, they are sometimes referred to as **distribution-free tests**.

Refresh your memory on some Non-parametric Test

- The **Mann-Whitney *U* test** – the nonparametric equivalent of the independent *t* test
- The **Wilcoxon matched-pairs signed-ranks (*T* or *W*) test** – the nonparametric equivalent of the matched or correlated *t* test
- The **Kruskal-Wallis One-Way analysis of variance by ranks (*H*) test** – the nonparametric equivalent of the One-Way analysis of variance
- The **Spearman rank-order correlation coefficient (r_s)** – the nonparametric equivalent of the pearson r correlation test.

THE SPEARMAN RANK-ORDER CORRELATION COEFFICIENT (SPEARMAN)

The Spearman rank-order correlation coefficient (Spearman r_s) is the nonparametric equivalent of the Pearson used to establish a linear relationship between two variables, X and Y

When to use

- All assumptions underlying the Pearson *r* apply to the Spearman r_s except that, for the scale of measurement.

- Thus, the spearman correlation requires that measurement on both variables, *X* and *Y* should at least be on an **ordinal** (ranking) scale.

Correlation coefficient

- Like the Pearson *r*, the values of the Spearman r_s range between –1.00 and +1.00.
- That is, the smallest possible value of r_s is equal to –1.00 (perfect negative correlation), and the largest possible value is equal to +1.00 (perfect positive correlation).
- The sign of the coefficient (i.e. whether positive or negative) again indicates the direction of the relationship
- The (absolute) value of the coefficient also indicates the strength of the relationship.
- An r_s value of zero (0) means that there is no linear correlation

between the two variables in the population

The formula

$$r_s = 1 - \frac{6\Sigma D_i^{\,2}}{n(n^2 - 1)}$$

worked example

- **Question**

In a nationwide examination in Mathematics, candidates could obtain integral grades ranging from **1** to **16**, where lower values reflect better performance. Sixteen (16) candidates who obtained grades ranging from **1** to **16** were admitted to a course in Statistics at a University. At the end of the first semester, the candidates were examined in Statistics and their performance recorded in percentage. The standings of the sixteen candidates in the two examinations were as follows

example

Candidate No.	Standing (Rank) in Maths	Standing (Grade) in Stats (%)
1	2	70
2	7	50
3	1	80
4	6	60
5	15	50
6	14	30
7	8	70
8	9	60
9	12	40
10	3	70
11	13	40
12	4	50
13	10	60
14	5	40
15	11	40
16	16	30

- If you were the teacher of the Statistics course, would you recommend that good standing in Mathematics be made a pre-requisite for enrolment into the Statistics course?

Step 1: Choice of Statistical Test

- We have sixteen candidates who have been measured on two variables, Performance in Mathematics, and Performance in Statistics.
- The question requires that we establish a relationship between the two variables.
- Performance in Mathematics is in ranks (ordinal scale) while Performance in Statistics is in percentages, an assumed interval scale of measurement.
- Since measurement on one variable (Performance in Mathematics) is on an ordinal scale while measurement on the second variable (Performance in Statistics) is on an interval scale, it becomes necessary to convert the interval data on Performance in Statistics to an ordinal scale.
- It is assumed that the two variables are linearly related.
- Therefore, the appropriate statistical test to use to answer the question is

the Spearman rank-order correlation coefficient (Spearman r_s).

Step 2: Statement of Hypotheses

Let r_s represent the correlation in the population. For the Statistics teacher to recommend that good performance in Mathematics be made a pre-requisite for enrolment into the Statistics course, then good standing in Mathematics should go with good standing in Statistics. Therefore, a significant positive correlation should exist between Performance in Mathematics and Performance in Statistics. The research hypothesis is therefore directional. The null hypothesis (H_o) and the alternative hypothesis (H_1) may therefore be stated as follows:

H_o : There is no significant correlation between performance in Mathematics and performance in Statistics, or even that performance in Mathematics is significantly negatively correlated with performance in Statistics [$\rho_s \leq 0$].

H_1: A significant positive correlation exists between performance in Mathematics and performance in Statistics [

ρ_s

Step 3: Decision Rules

Given: 0.05 level of significance, a one-tailed Spearman test, n = 16, the critical values of r_s in the Spearman r_s tables is equal to ± 0.430. Since we expect the r_s value to be positive, the critical value of r_s in this case is equal to + .430.

Therefore If $r_{s\ obs}$ < 0.430, retain H_o , and If $r_{s\ obs}$ ≥ 0.430, reject H_o .

OR use the t conversion formula

Step 3: Decision Rules cont:

[The Decision Rules can also be alternatively stated as follows:

Given : 0.05 level of significance, a one-tailed Spearman r_s test, n = 16 , the r_s value obtained through calculation ($r_{s\,obs}$) may be converted to a t value using the conversion formula:

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-(r_s)^2}}$$

.

The computed t value (t_{obs}) may be referred to the t tables with n - 2 = 16 - 2 = 14 df . From the t tables , the critical values of t for a one-tailed test is equal to ± 1.761. Since we expect a positive correlation, then the t_{crit} in this case is equal to + 1.761. This means that t_{obs} must fall on the positive tail end of the t distribution.

Therefore **if t_{obs} < 1.761, retain H_o ; and**

if t_{obs} ш 1.761, reject H_o .

Step 4: Computation

- Let X stand for *Performance in Mathematics* and Y stand for *Performance in Statistics*. With the usual notations, the computational formula for the Spearman r_s is given by:

$$r_s = 1 - \frac{6\Sigma D_i^2}{n(n^2 - 1)}$$

- where D is the difference between paired ranks (i.e. Or $R_{xi} - R_{yi}$ *or* $R_{yi} - R_{xi}$) performance in the Statistics course, which is on an interval scale, will have to be converted to an ordinal scale. The data is therefore re-arranged as in the following table and D_i and D_i^2 values are calculated as follows:

Step 4: Computation

- Computing all the required summations and substituting into the formula, we have:

$$r_s = 1 - \left[\frac{6 \times 177.50}{16(256 - 1)} \right]$$

$$= 1 - \left[\frac{1065}{16 \times 255} \right]$$

$$= 1 - \left[\frac{1065}{4080} \right]$$

$$= 0.7389706$$

- i.e. $r_{s\ obs} \approx 0.739$ (corrected to 3 decimal places)

Step 5: Decision

- Referring $r_{s\ obs} = 0.739$ to the *Decision Rules* in (Step 3), we observe that $(r_{s\ obs} = 0.739) > (r_{s\ crit} = 0.430)$.

$\therefore$ $\boldsymbol{H_0}$ **is rejected at the .05 level of significance.**

[Alternatively, the decision can be stated as follows: check on next page)

Step 5: Decision

- Using the conversion formula: $t = \dfrac{r_S\sqrt{n-2}}{\sqrt{1-(r_S)^2}}$

and substituting the values of n and rs obs into this formula, we get, $t = 0.7389706\sqrt{14}\Big/\sqrt{1-(0.7389706)^2}$

$$= 4.1039337$$

i.e. $t_{obs} \approx 4.104$ (corrected to 3 decimal places).

- Referring $t_{obs} = 4.104$ to the *Decision Rules* (Step 3), we note that ($t_{obs} = 4.104$) > ($t_{crit} = 1.761$) .

$\therefore$ ***H_0* is rejected at the 0.05 level**

Step 6: Interpretation

- At the .05 level of significance, there is a strong **significant positive correlation** between performance in Mathematics and performance in Statistics. Therefore, if I were the teacher of the course in Statistics, I would strongly recommend that good standing in Mathematics be made a pre-requisite for enrolment into the Statistics course.