Kenya Medical Training College
Department of Clinical Medicine
Year Two Semester One
Pearson's Correlation Coefficient: Worked Examples
26th November 2020

Willis J. Opalla

Learning Objective

- To apply Pearson's Correlation Coefficient formula in determining the relationship between various variables.

Learning Outcomes

- By the end of this session, you should be able to
  1. Explain the appropriate formulae for Pearson's Correlation Coefficient.
  2. Make statistical inferences on relationship between independent and dependent variables through application of Pearson's Correlation Coefficient.

Correlation.

- The Mean, Median, Mode Range and Standard Deviation are univariate as it describes only one variable at a time.
- Description for two variables is done in terms of relationship.
- The most common bivariate descriptive statistics include cross tabulation tables, correlation and regression.
- The cross tab table is same as contingency table.

Types of Correlation Coefficient

- Based on the direction of changes;
  a) Perfect Positive Correlation:

     X is directly proportional to Y. e.g. Designation and Salary. r = 1.

  b) Perfect Negative Correlation:

     X and Y are inversely proportionate. r= -1.

     e.g. Insulin and blood sugar.

  c) Moderately Positive Correlation:

     A type of positive correlation.

  d) Moderately Negative Correlation.

     A type of negative correlation.

  e) No Correlation. No relation. r = 0.

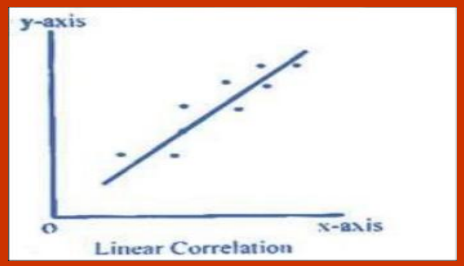
     smoking and type of housing.

Types of Correlation Coefficient

- Based on number of variables;
  a) Simple: Only two variables.
  b) Multiple: More than two variables.
  c) Partial: More than two variables but correlation is study for only two variables by keeping the third variable as constant.
  - e.g. X= yield, y = fertilizer, z = amount of rainfall.
    - Simple = r(xy), r(yz), r(xz)
    - Multiple= r(xyz)
    - Partial = r(xy)z

Types of Correlation Coefficient

- Based on Linearity;
  - Linear:
    - If the changes in one variable bears a constant amount of change or solid pattern of change in another variable then the correlation is said to be linear.



  - Non Linear:
    - If the ratio of change is not constant, i.e. when all the points on the scatter diagram tend to lie near a smooth curve, the correlation is said to be non linear (curvilinear).

Methods of Correlation Coefficient

1. Karl Pearson's method of correlation
2. Spearman's rank correlation.
3. Scatter Plot or graph or scatter diagram method.

Karl Pearson's Correlation Method

- Is a measure of the strength of a linear association between two variables.
- Is denoted by r or rxy(x and y being the two variables involved).
- It attempts to draw a line of best fit through the data of two variables.
- The value of r, indicates how far away all these data points are from this line of best fit.
- Treats all variables equally: i.e. does not consider whether the variable is dependent or independent.

Properties of Pearson's Method

- r is unit-less, hence it may be used to compare association between different bivariate populations.
- Its value always lies between +1 and - 1.
- The following degrees of association can be seen between the variables:
  - A value > 0 indicates a positive association i.e. as the value of one variable increases, so does the value of the other variable.
  - A value < 0 indicates a negative association i.e. as the value of one variable increases, the value of the other variable decreases.

Karl Pearson's Correlation Coefficient

- Interpretation of Pearson's method

Strength of Association	Negative r	Positive r
Weak	-0.1 to -0.3	0.1 to 0.3
Average	-0.3 to -0.5	0.3 to 0.5
Strong	-0.5 to -1	0.5 to 1
Perfect	-1	+1

- The coefficient of correlation is " zero" when the variables X and Y are independent.

Assumptions of Karl Pearson's Correlation Coefficient

- The relationship between the variables is "Linear", i.e. when the two variables are plotted, a straight line is formed by the points plotted.
- The variables are independent of each other.
- The coefficient of correlation r = -0.67, shows correlation is negative because the sign is '-' and the magnitude is 0.67.

Karl Pearson's Correlation Coefficient

- Can be calculated using the formula:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

r_{xy} = Pearson r correlation coefficient between x and y

n = number of observations

x_i = value of x (for i^{th} observation)

y_i = value of y (for i^{th} observation)

Karl Pearson's Correlation Coefficient

- Or,

$$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2}\sqrt{(Y-\bar{Y})^2}}$$

Where, $\bar{X}$ = mean of X variable

$\bar{Y}$ = mean of Y variable

- In case of grouped data "x" and "y" can be taken as the mid value of the class interval.

Karl Pearson's Correlation Coefficient

- Compute Pearson's correlation coefficient from the following data;

Weight in Kg.	60	70	80	90
Cholesterol	120	130	140	150

- Create the table.
- Find the mean of "x" and "y"

Karl Pearson's Correlation Coefficient

- Assumptions of Pearson's method

x	y	$X-\bar{x}$	$Y-\bar{y}$	$(X-\bar{x})(Y-\bar{y})$
60	120	-15	-15	225
70	130	-5	-5	25
80	140	5	5	25
90	150	15	15	225
$\sum x = 300$	$\sum y = 540$			$\sum(X-\bar{x})(Y-\bar{y}) = 500$

Karl Pearson's Correlation Coefficient

- Pearson's method

$$r = \frac{500}{\sqrt{500 \times 500}}$$

$$= \frac{500}{\sqrt{2,50,000}}$$

$$= \frac{500}{500}$$

$$= 1$$

$(x - \bar{x})^2$	$(y - \bar{y})^2$
225	225
25	25
25	25
225	225
$\Sigma(x - \bar{x})^2$ 500	$\Sigma(y - \bar{y})^2$ 500

Hence there is perfect correlation between weight and patients' cholesterol level.

Practice Question

- Compute the correlation coefficient from the following data;

Age	30	40	50	60	70
Blood pressure	120	130	140	150	160

Pearson's Correlation Coefficient

- Nine students held their breath, once after breathing normally and relaxing for one minute and once after hyperventilating for one minute. The table shows the length (in seconds) each held their breath. Is there an association between the two variables?

Subject	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Normal	56	56	65	65	50	25	87	44	35
Hypervent	87	91	85	91	75	28	122	66	58

Pearson's Correlation Coefficient

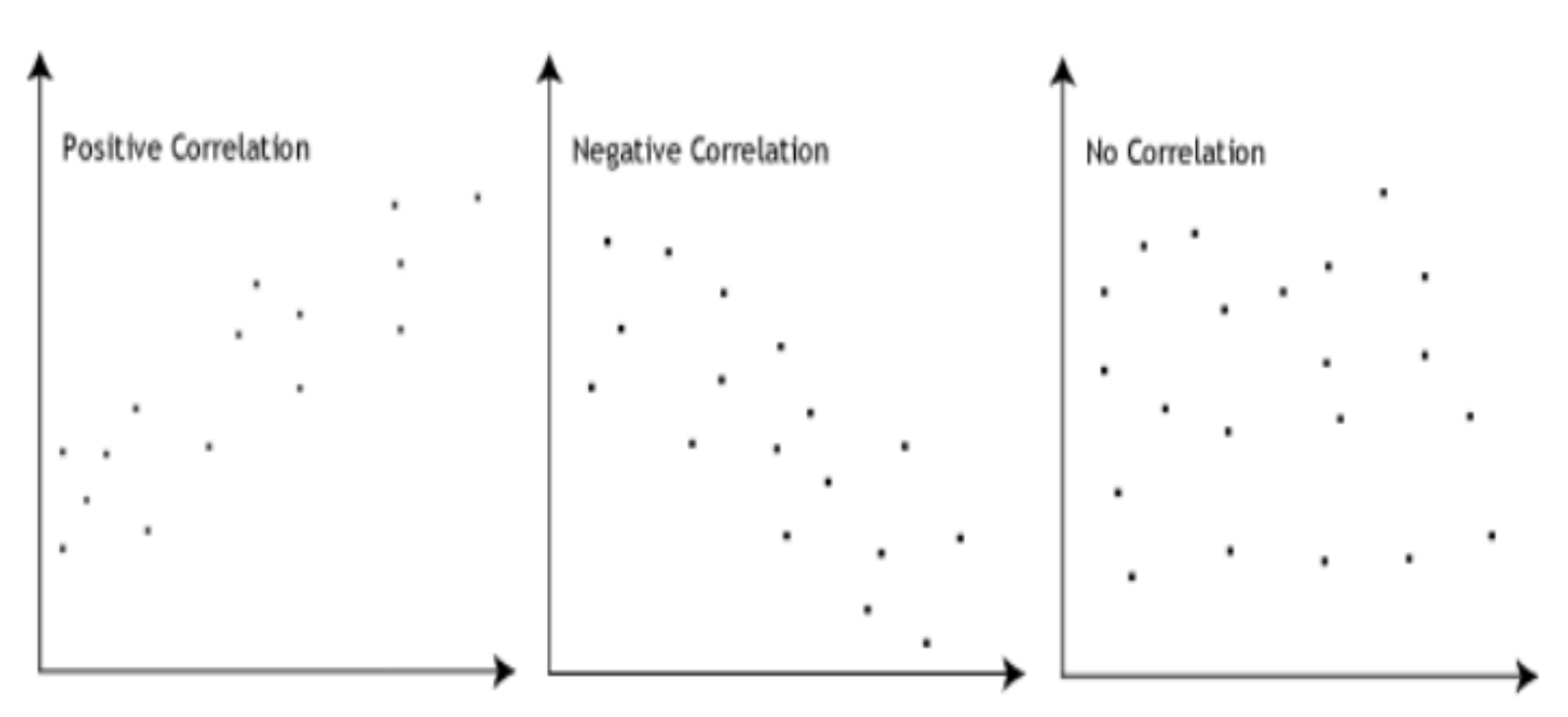
- Hyperventilating times are considered to be the dependent variable, so are plotted on the vertical axis.
- Pearson correlation coefficient attempts to draw a line of best fit through the data of two variables.
- The 'r' indicates how far away all the data points are from the line of best fit (i.e., how well the data points fit the line of best fit).

Pearson's Correlation Coefficient

- *r*, values can range from +1 to -1.
- A value of 0: there is no association between the two variables.
- A value > 0 (i.e. +1): a positive association; as the value of one variable increases, so does the value of the other variable.
- A value < 0 (i.e. -1): indicates a negative association; as the value of one variable increases, the value of the other variable decreases.

Pearson's Correlation Coefficient

- Diagrammatically:



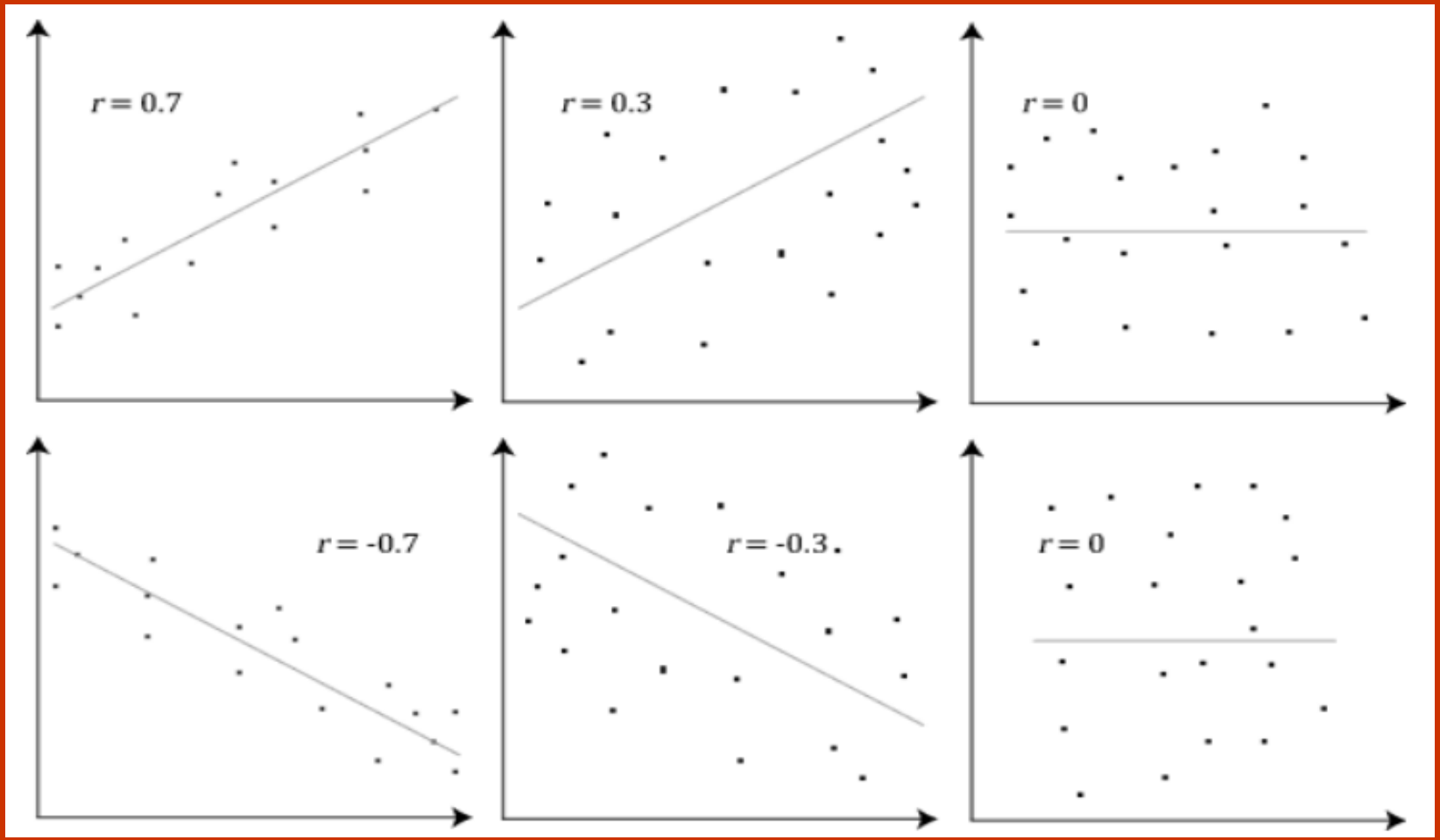
Pearson's Correlation Coefficient

- The stronger the association of the two variables, the closer the Pearson correlation coefficient, *r*, to either +1 or -1 depending on whether the relationship is +ve or -ve.
- +1 or -1 means that all data points are included on the line of best fit, i.e. there are no data points that show any variation away from this line.

Pearson's Correlation Coefficient

- Values for *r* between +1 and -1 (e.g. *r* = 0.8 or -0.4) indicate that there is variation around the line of best fit.
- The closer the value of *r* to 0 the greater the variation around the line of best fit.
- Different relationships and their correlation coefficients are shown in the following diagram:

Pearson's Correlation Coefficient



Pearson's Correlation Coefficient

	Coefficient, r	
Strength of Association	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to -1.0

Pearson’s Correlation Coefficient

- Two variables have to be measured on either an interval or ratio scale.
- Both variables do not need to be measured on the same scale (e.g., one variable can be ratio and one can be interval).
- For ordinal data, Spearman's rank-order correlation are used instead of the Pearson product-moment correlation.

Pearson’s Correlation Coefficient

- The two variables can be measured in different units.
  - e.g, correlating a person's age with blood sugar levels.
  - Here, age is measured in years and blood sugar level measured in mmol/L.
- The units of measurement do not affect the calculation of Pearson's correlation coefficient.
  - This allows the correlation coefficient to be comparable and not influenced by the units of the variables used.

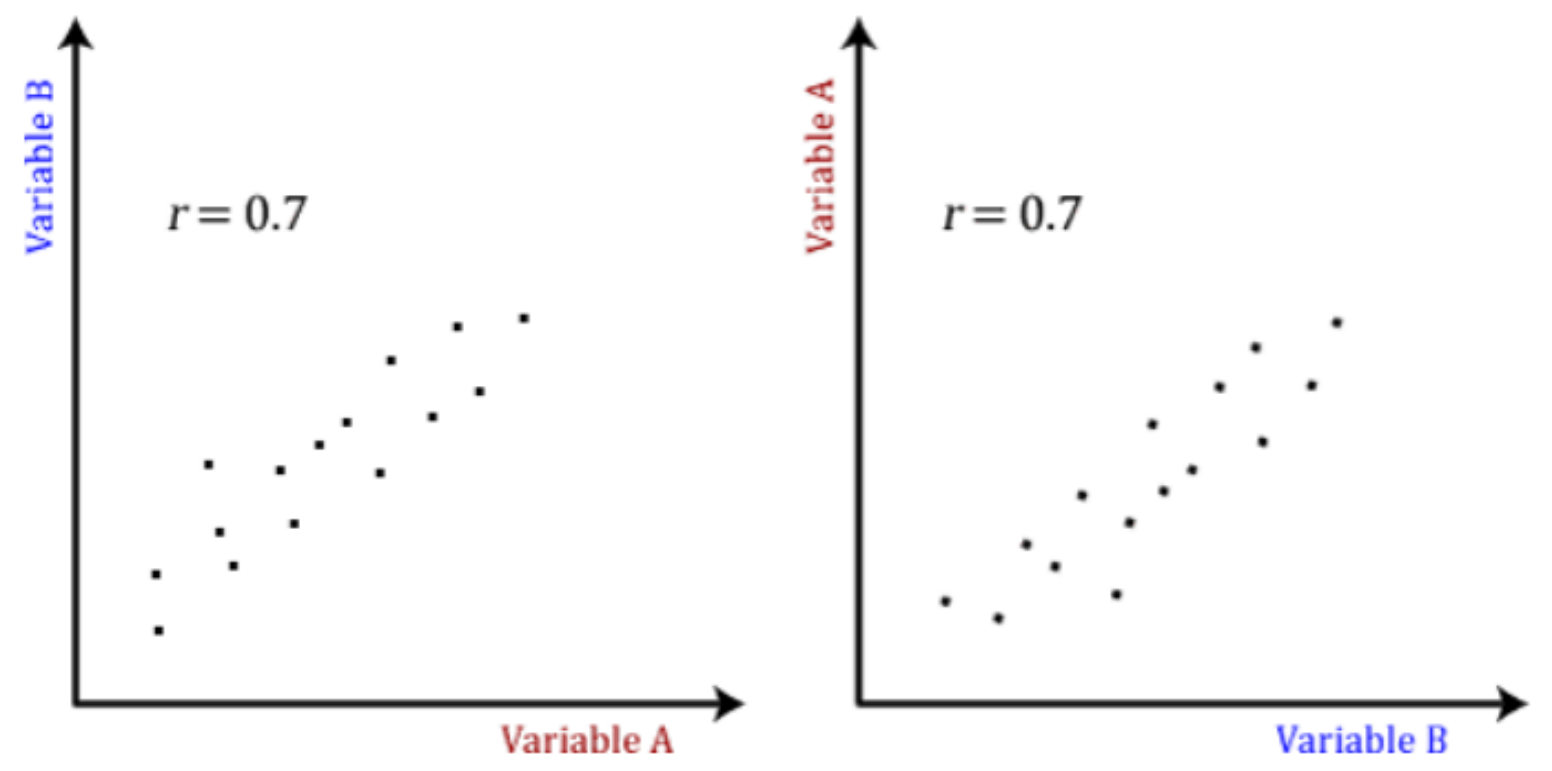
Pearson’s Correlation Coefficient

- Does not take into consideration whether a variable is a dependent or independent variable.
- It treats all variables equally.
- e.g., to find out whether basketball performance is correlated to a person's height, a graph of performance against height is plotted, then Pearson correlation coefficient calculated.

Pearson's Correlation Coefficient

- If, e.g., r = .67. i.e, as height increases so does basketball performance., this makes sense.
- However, if the variables are interchanged when plotting as if to determine whether height was determined by one's basketball performance, r is still = .67.
  - The Pearson correlation coefficient takes no account of any theory behind why the two variables were chosen for comparison.

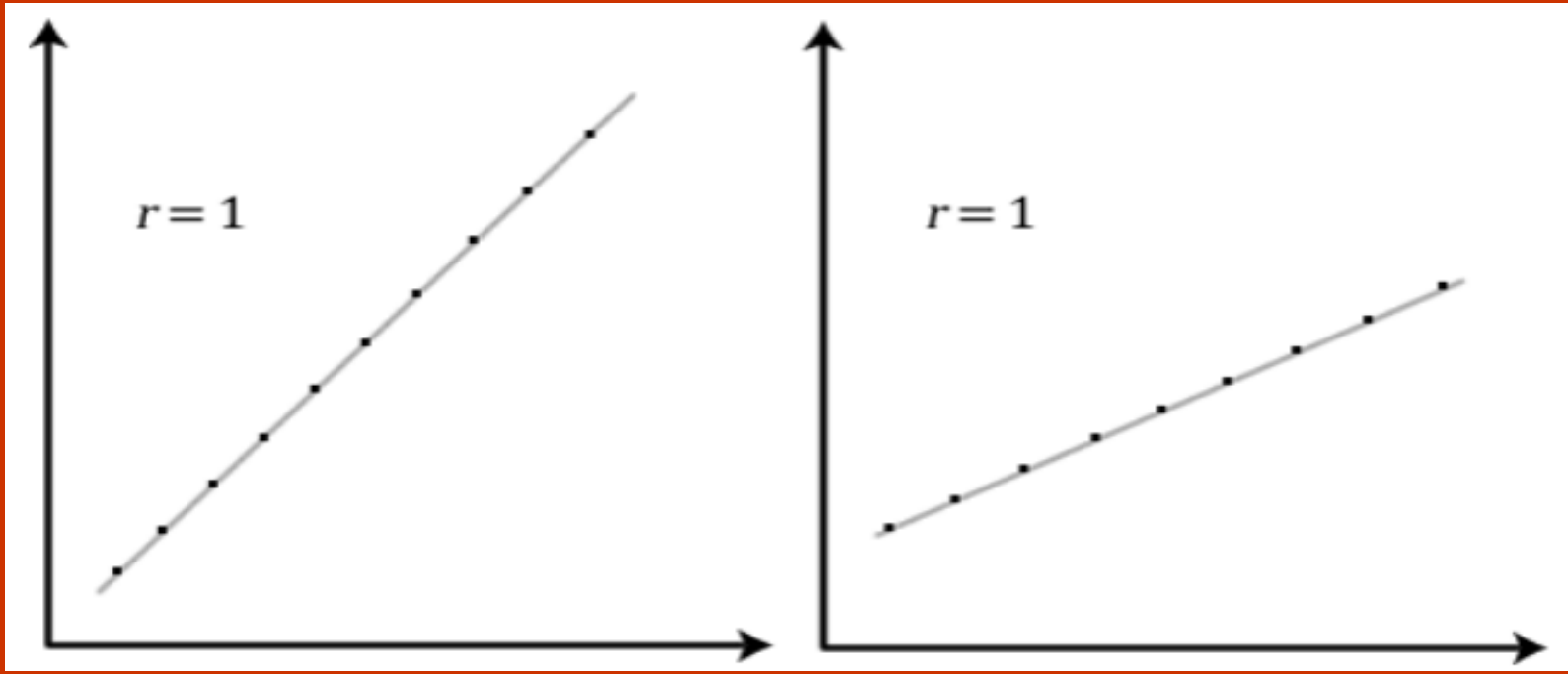
Pearson's Correlation Coefficient



Pearson's Correlation Coefficient

- The Pearson correlation coefficient, *r*, does not represent the slope of the line of best fit.
- Therefore a Pearson correlation coefficient of +1 does not mean that for every unit increase in one variable there is a unit increase in another.
- It means there is no variation between the data points and the line of best fit:

Pearson's Correlation Coefficient



Pearson's Correlation Coefficient

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Pearson's Correlation Coefficient

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

- r_{xy} = Pearson r correlation coefficient between x and y
  n = number of observations
  x_i = value of x (for ith observation)
  y_i = value of y (for ith observation)

Merits and Demerits of Pearson's Method

- Merits:
  - It summarizes the correlation and if plotted on a graph with a linear line, it shows the direction too.
- Demerits:
  - The correlation coefficient always assumes linear relationship regardless of the fact that assumption is correct or not.
  - The value of the coefficient is unduly affected by the extreme values.
  - It cannot be used for ordinal data
  - It is time consuming method.

Summary

- Spearman's correlation is a non-parametric technique *for measuring the relationship between paired observations of two variables when data is ranked.*
- Widely distributed or scattered points on the graph of scatter diagram indicates very little relationship between the independent and dependent variables.

References

- Statistics How To (2020) *Pearson's Correlation Coefficient*, [Online] Available: https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/ (Accessed 23.11.2020)
- Joseph, J. K. (n.d) *Measures of Relationship*, [Online] Available: https://www.slideshare.net/JohnykuttyJoseph/measures-of-relationship, (Retrieved 26.11.2020)