Kenya Medical Training College
Department of Clinical Medicine
Year Two Semester One
Measures of Relationship: Worked Examples
26th November 2020

Willis J. Opalla

Learning Objective

- To apply the learnt formulae to calculate and make inferences on relationship between various variables.

Learning Outcomes

- By the end of this session, you should be able to
  1. Explain the appropriate formulae for each measure of relationship.
  2. Make statistical inferences on relationship between independent and dependent variables through application of formulae for coefficients of correlation.

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.

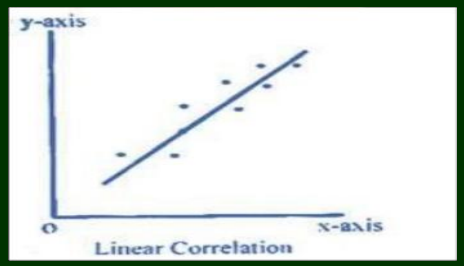
     smoking and type of housing.

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

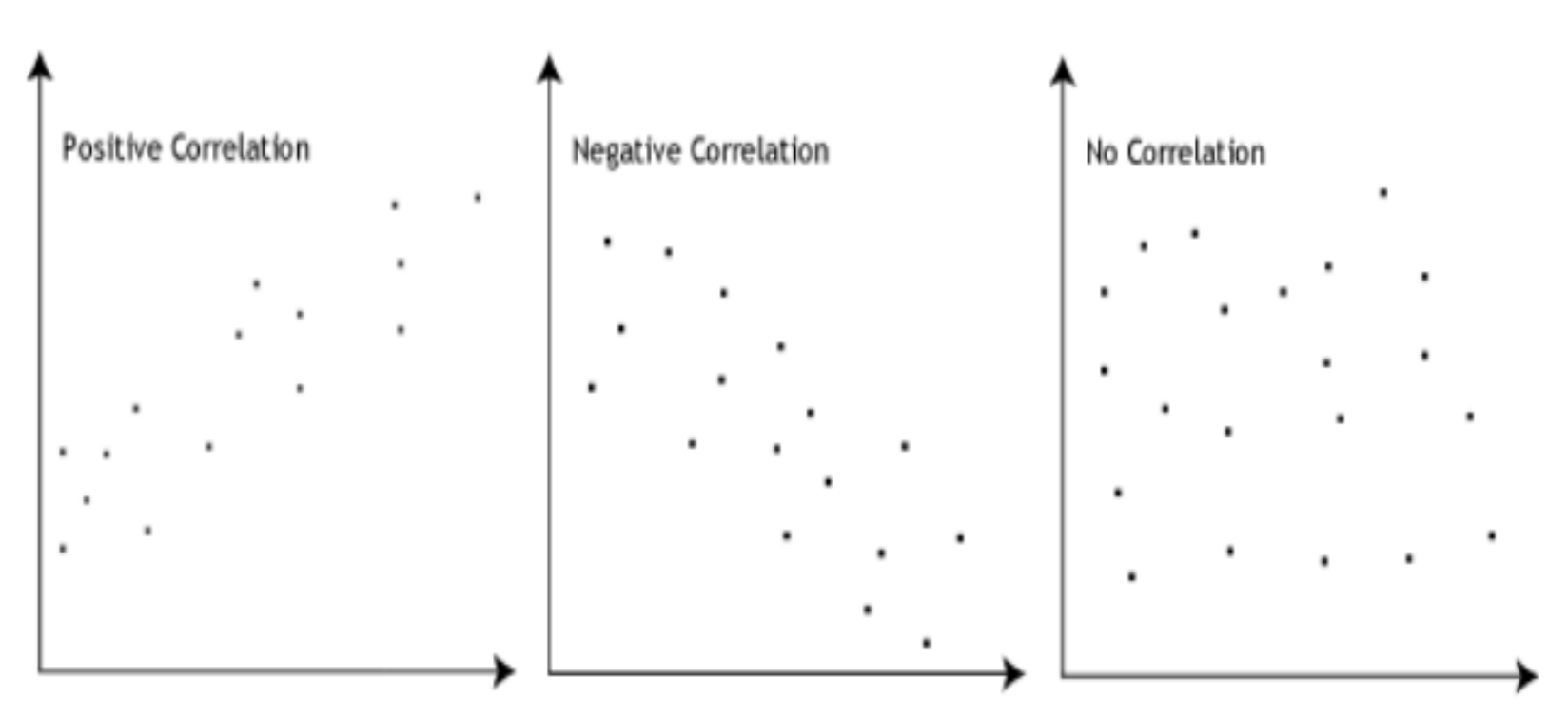
- Hyperventilating times are considered to be the dependent variable, so are plotted on the vertical axis.
- Pearson correlation coefficient attempts to draw a line of best fit through the data of two variables.
- The 'r' indicates how far away all the data points are from the line of best fit (i.e.,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



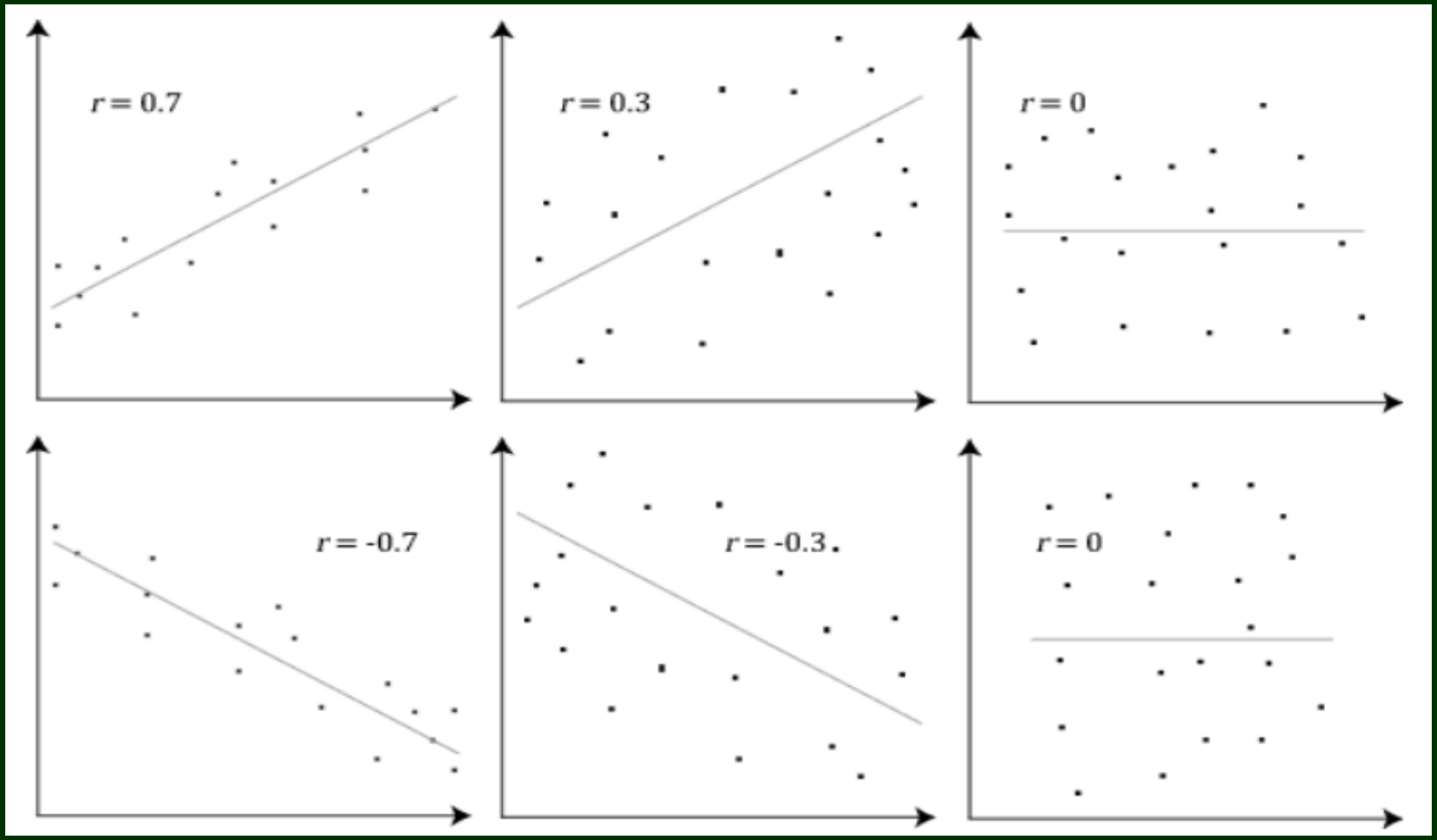
Pearson's Correlation Coefficient

- 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.

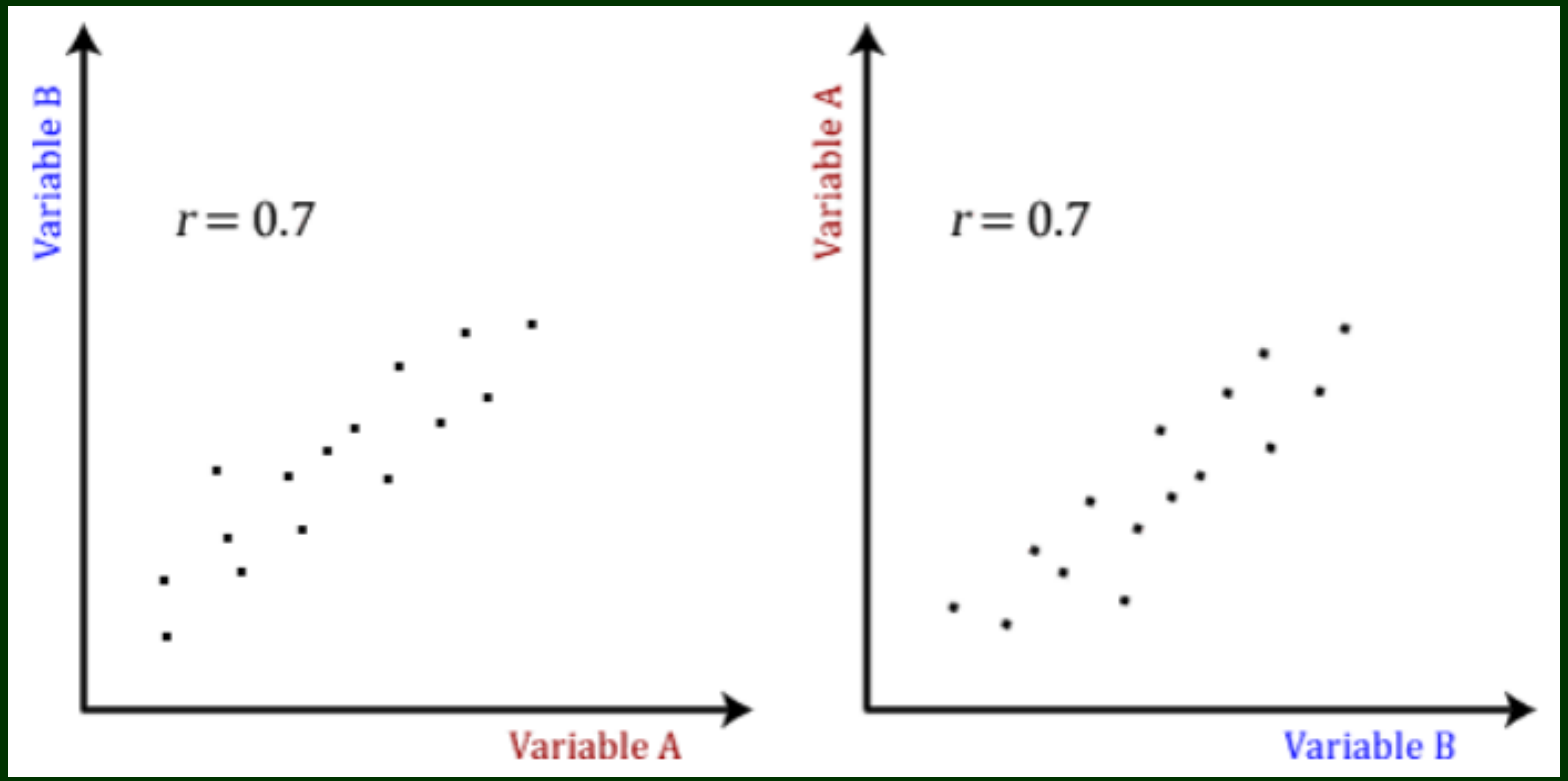
Pearson’s Correlation Coefficient

- 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.

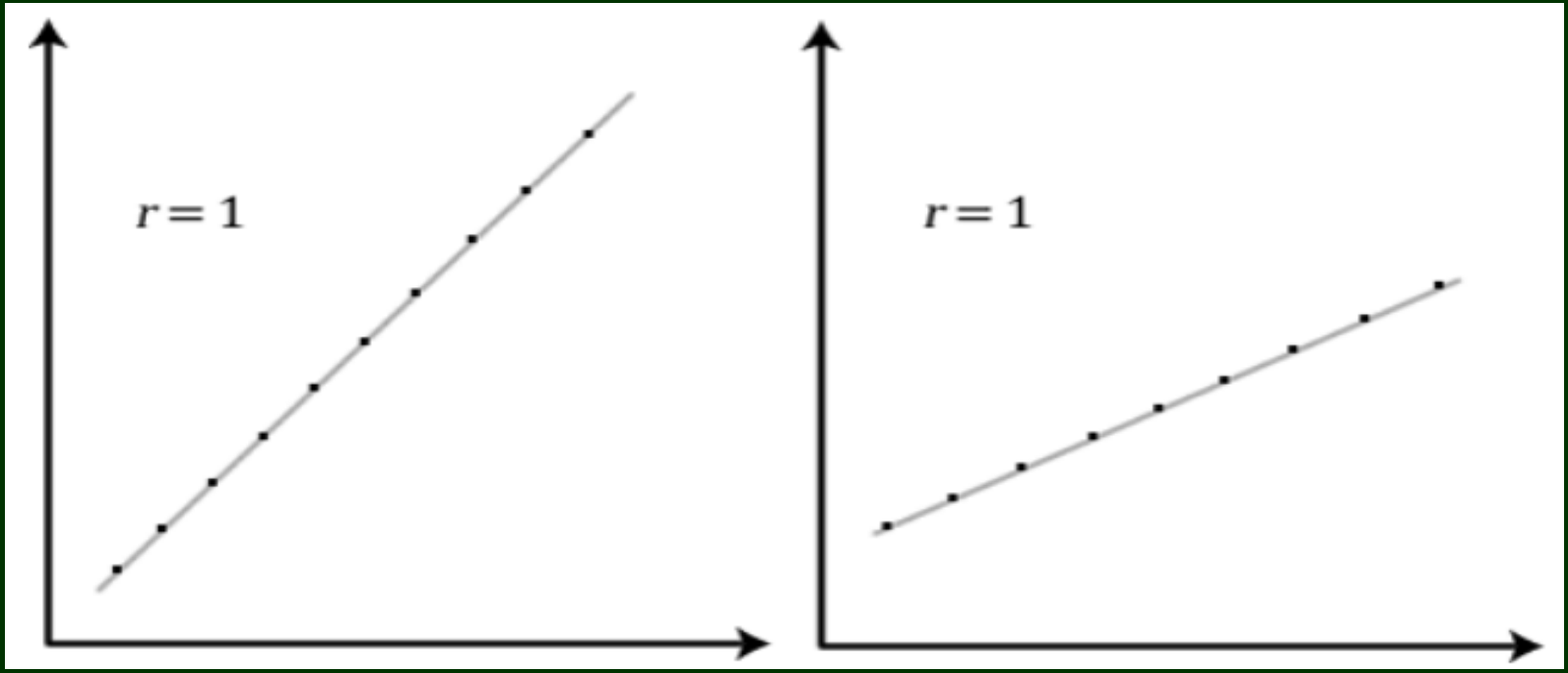
Pearson's Correlation Coefficient



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.

Spearman's Correlation Coefficient

- For finding correlation between two variables by taking their ranks.
  - Useful for qualitative data.
  - Can be used when the actual magnitude of characteristics under consideration is not known, but relative position or rank of the magnitude is known.
  - Is the nonparametric version of the Pearson correlation coefficient.
  - The data must be ordinal, interval or ratio with ranks.

Spearman's Correlation Coefficient

- For summarizing the strength and direction (negative or positive) of a relationship between two variables.
- Will always be between +1 and -1, where:
- +1 = a perfect positive correlation between ranks.
- -1 = a perfect negative correlation between ranks.
- 0 = no correlation between ranks.

Spearman's Correlation Coefficient

- It is denoted by " rho" (ρ).
- There are two cases for calculating rank correlation.
- Case 1.
  - No tie of allotted rank
- Case 2.
  - There is a tie for two or more values or ranks in either "x" or "y" or in both "x" and "y".

Spearman's Correlation Coefficient

- Case 1: No tie of allotted rank:
- In this, none of the values/ranks of x and y are repeated.
- "*p*" can be calculated using the formula:

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

d = difference in the ranks of data set of 'x' and 'y'.

i.e. d = Rx – Ry (i.e. d = rank x – rank y)

Spearman's Correlation Coefficient

- The formula for the Spearman rank correlation coefficient when there are no tied ranks is:

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

- **Example Question:**

  Pathology and physiology scores for nine students are:

  - Pathology: 35, 23, 47, 17, 10, 43, 9, 6, 28
  - Physiology: 30, 33, 45, 23, 8, 49, 12, 4, 31

- Compute the student's ranks in the 2 subjects and compute the Spearman's rank correlation.

Spearman's Correlation Coefficient

- Step 1: Find the ranks for each subject.
  - To rank manually by hand, the scores are ordered from the largest to smallest then assigned the rank 1 to the highest score, 2 to the next highest etc:

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

Spearman's Correlation Coefficient

- Step 2: Add a third column, d.
  - The d is the difference between ranks.
  - e.g., the first student's pathology rank is 3 and physiology rank is 5, so the difference is 3 points.
- In the 6th column, square the values of d.

Pathology	Rank	Physiology	Rank	d	d^2
35	3	30	5	2	4
23	5	33	3	2	4
47	1	45	2	1	1
17	6	23	6	0	0
10	7	8	8	1	1
43	2	49	1	1	1
9	8	12	7	1	1
6	9	4	9	0	0
28	4	31	4	0	0

Spearman's Correlation Coefficient

- Step 3: Add up all values of d^2.
  $4 + 4 + 1 + 0 + 1 + 1 + 1 + 0 + 0 = 12$.

  This will be required for the factor $6\Sigma d^2$ of the formula.
- Step 4: Insert the values into the formula. These ranks are not tied (i.e. not similar) so the first formula is applied:

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

Spearman's Correlation Coefficient

- Substituting:

$$\rho = 1 - (6x12)/\left[9(81-1)\right]$$
$$= 1 - 72/720$$
$$= 1 - 0.1$$
$$= 0.9$$

The Spearman's Correlation for this set of data is 0.9, hence implying a strong positive correlation.

Worked Example

- Calculate the rank correlation of the marks for five students in Microbiology and Immunology.
  - Only the ranks should be arranged in ascending or descending order.
  - One data pair belongs to one student.
  - Prepare a table to calculate Σd^2

Microbiology 85 81 77 68 53

Immunology 78 70 72 62 67

Microbiology	Rank	Immunology	Rank	d	d^2
85	1	78	1	0	0
81	2	70	3	1	1
77	3	72	2	1	1
68	4	62	5	1	1
53	5	67	4	1	1

Worked Example

- Substituting in the equation:

  $\rho = 1 - 6x4 / 5 (25-1)$

  $= 1 - 24/120$

  $= 0.8$

  The marks of the two subjects are strongly positively correlated.

Exercise

- Calculate the Spearman's correlation coefficient for the temperatures (^oC) of two patients, Adan and Kadzo on different days in one week.

  Adan 20 28 25 23 22 30 31

  Kadzo 15 26 17 19 21 24 27

- First step:

  Feed the data into a 6 x 8 table.

Adan	Rank	Kadzo	Rank	d	d^2

Case 2: Tie of Allotted Rank

- i.e. more than one rank is present in either x or y or both x and y.
- "ρ" is calculated using the Spearman's formula and then adding CF, the Correlation Factor.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} + \text{CF}.$$

- CF has to be calculated for each repeated ranks and then added.
- The CF is calculated using $CF = m\,(m^2 - 1)/12$, where m is the number of times the data repeats.
- d = difference in the ranks of data set of 'x' and 'y' (d = Rx - Ry).

Case 2: Tied Ranks

- Calculate the rank correlation of the following marks obtained by eight students in Medicine and Obstetrics.
  - Medicine 60 81 72 68 53 75 85 68
  - Obstetrics 78 70 72 62 67 70 70 61
  - Here Medicine (x) the value 68 is repeated twice and in Obstetrics (y) the value 70 is repeated thrice.
  - In the first series CF = 2x(4-1)/12 = 0.5
  - In the second series CF = 3x(9-1)/12 = 2

Case 2: Tied Ranks

- Tabulating:

Medicine	Rank, Rx	Obstetrics	Rank, Ry	d	d^2
60	6	78	1	5	25
81	2	70	3	-1	1
72	4	72	2	2	4
68	5	62	5	0	0
53	7	67	4	3	9
75	3	70	3	0	0
85	1	70	3	-2	4
68	5	61	6	-1	1

- $\Sigma d^2 = 44$
- $n = 8$

Case 2: Tied Ranks

- Substituting:

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

$p = 1 – 6x\ 44 + 0.5 + 2 / 8\ (64\text{-}1)$

$= 1 – 266.5/504$

$= 1\text{-}\ 0.53$

$= 0.47$

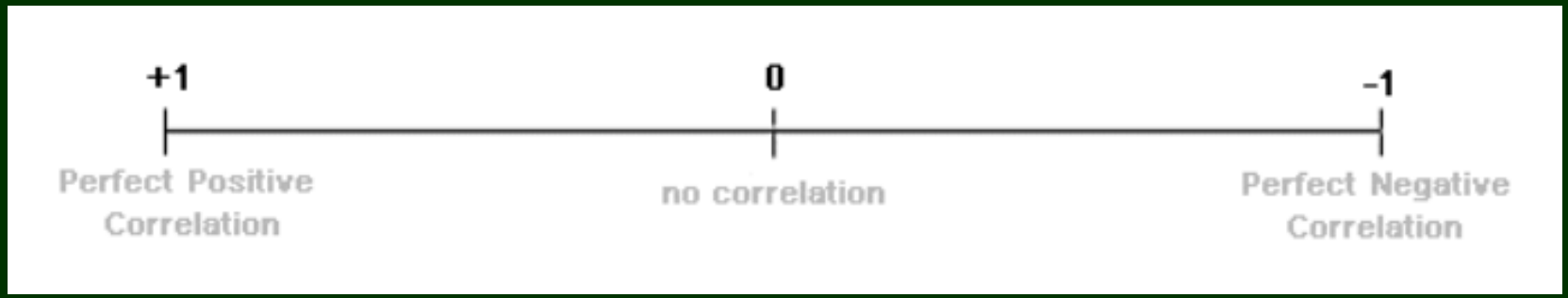
- The marks of the two subjects have a positive correlation.

Merits and Demerits of Spearman's Correlation Coefficient

- Merits
  - Can be used as a measure of degree of association between qualitative data.
  - Is very simple and easily understandable.
  - Can be used when the actual data is given or when only the ranks of the data are given.
- Demerits
  - Ranks coefficient for a frequency distribution, i.e., grouped data, can't be calculated.
  - Calculation gets tedious when the number of observations is large.

Meaning of p Value

- The closer p is to +1 or -1, the stronger the likely correlation.
- A perfect positive correlation is +1 and a perfect negative correlation is -1.
- A p value of -0.73 suggests a fairly strong negative relationship, i.e.



Scatter Diagram Method

- Scatter Diagrams are convenient mathematical tools to study the correlation between two random variables.
- A sheet of paper upon which the data points corresponding to the variables of interest are scattered.
- The association between the two variables can be determined by the pattern that the data points form on the paper.
- The suitable correlation analysis technique can further be applied.

Scatter Diagram Method

- Application
  - A quick way of confirming a hypothesis that two variables are correlated.
  - Provides a graphical representation of the strength of the relationship between two variables.
  - Also helps in understanding cause and effect relationship to evaluate whether manipulation of independent variable (cause) is producing the change in dependent variable (effect.)

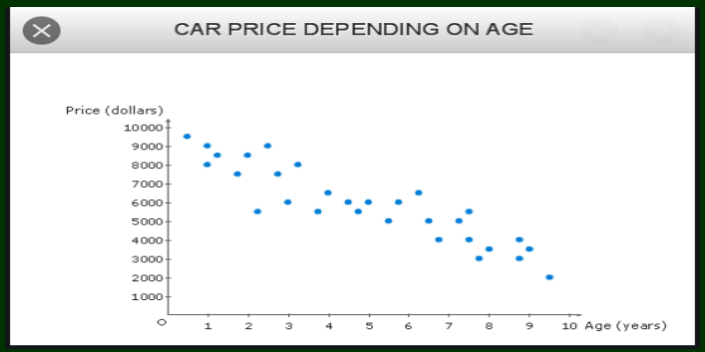
Steps to Construct a Scatter Diagram

- Step 1: Draw a line “L”, on a paper where the horizontal part of “L” is x axis and vertical part of “L” is y axis.

- Step 2: Make the scale units at even multiples such as 10,20,30,40 etc so as to have an even scale system.

Steps to Construct a Scatter Diagram

- Step 3: Place the independent (cause) variable on horizontal axis and dependent (effect) variable on vertical axis.
  - Plot the data points at the intersection of x and y axis.
  - The plots on the graphs generally look scattered and hence named as scatter plot.
  - Interpret the data and find the relationship.



Interpretation of Scatter Diagram

- It suggests the degree and the direction of the correlation.
- The greater the scatter of plotted points, the lesser is the relationship.
- If the points are closer to a diagonal line from left corner to the upper right corner, the correlation is perfectly positive. (r = +1)
- If all the plots are on the diagonal line from upper left corner to the lower right corner, the correlation is perfectly negative. (r = -1)

Interpretation of Scatter Diagram

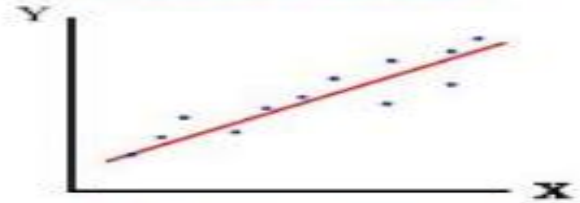
Correlation and Regression

Linear correlation:

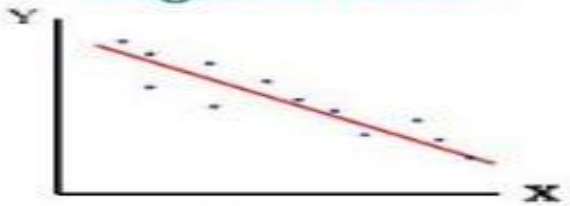
- Does one variable increase or decrease linearly with another?
- Is there a linear relationship between two or more variables?

Types of linear relationships:

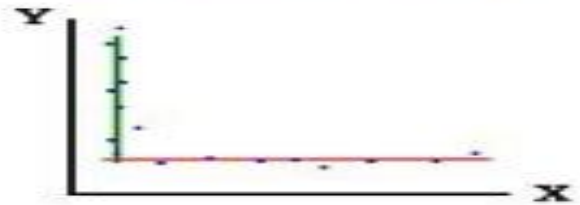
Positive linear



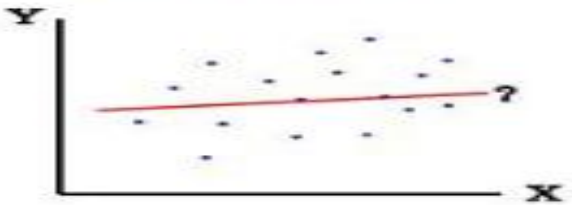
Negative linear



No relationship



None or weak



Interpretation of Scatter Diagram

- If the points are widely distributed or scattered on the graph it indicates very little relationship. (weak positive or weak negative).
- If the plotted points lie on the diagram in disorganized manner it shows absence of correlation.

Merits and Demerits of Scatter Diagram

- Merits
  - It is simple and non mathematical method to study correlation.
  - Easily understood and rough idea can be quickly formed.
  - Not influenced by the extreme values of x and y.
- Demerits
  - Cannot establish the exact degree of correlation.
  - It cannot be always referred as a measure of degree of correlation since it is not mathematical and hence less reliable.

Regression

- Regression analysis is a reliable method of identifying which variables have impact on a topic of interest.
  - Dependent Variable:
    - This is the main factor that the study seeks to understand or predict.
  - Independent Variables:
    - The factors that are hypothesized to have an impact on the dependent variable of the study.

Regression

- Regression is done by deriving a suitable equation on the basis of available bivariate data.
- This equation is called Regression equation and its geometrical representation is called Regression curve.
- The regression equation requires the Regression coefficient.

Regression Analysis

- Regression analysis determines the nature of relationship between the variables.
  - i.e. studies the functional relationship between the variables and thereby providing a mechanism for prediction.
- Regression analysis describes the mathematical relationship between dependent variable (y) and independent variable (x).
- Aims at estimating the unknown values of ‘y’ and for the known values of ‘x’ by use of the equation y = a+bx

Properties of Regression Coefficient

- It is denoted by b.
- Between two variables (x and y), two values of regression coefficient can be obtained.
  - One will is obtained when x is considered as independent and y as the dependent variable and the other when it is reversed.
  - The regression coefficient of y on x is represented as b_{yx} and that of x on y as b_{xy}.
- The square root of the products of two regression coefficients ($b= b_{yx}$ and $b_1= b_{xy}$) is correlation coefficient.

Regression Equations

1. Regression Equation of y on x.
2. Regression equation of x on y.

- Regression Equation of y on x.
  - It is y = a + bx, where y =dependent variable,
    x = independent variable and
    a and b are constants.
- It is also to be noted that

b = byx (regression coefficient of y on x)

$$b = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n\bar{x}^2}$$

$$a = \bar{y} - b\bar{x}$$

Regression Equation of x on y

- It is $x = a_1 + b_1x$

  where x = dependent variable,

  y = independent variable and

  a_1 and b_1 are constants.

- It is also to be noted that

  b_1 = bxy (regression coefficient of x on y).

  $$b_1 = \frac{\Sigma xy - n\bar{x}\,\bar{y}}{\Sigma y^2 - n\bar{y}^2}$$

  $$a_1 = \bar{x} - b_1\bar{y}$$

Types of Regression

- Simple linear regression:
  - It is the relationship between a scalar response or dependent variable and one or more explanatory/independent variables.
- Multiple linear regression:
  - More than one explanatory variable.
  - Multivariate linear regression:
  - Multiple correlated dependent variables are predicted, rather than a single scalar variable.

Types of Regression

- Positive regression:
  - A positive sign indicates that as the predictor variable increases, the response variable also increases.
- Negative regression:
  - A negative sign indicates that as the predictor variable increases, the response variable decreases.

Types of Regression

- Linear and nonlinear Regression:
  - A model is linear when each term is either a constant or the product of a parameter and a predictor variable.
  - It is non linear if the equation does not meet the linear criteria.

Regression Analysis

- Worked Example
  - Fit a regression equation of BP on age based on the following data and estimate the probable BP for a 55 year-old.
    - n = 5 •
    - $\overline{X} = \Sigma x/n = 250/5 = 50$
    - $\overline{Y} = \Sigma y/n = 700/5 = 140$
- The regression equation to be fitted is y = a+bx where y is BP and x is the age.
  - Age 30 40 50 60 70
  - BP 120 130 140 150 160

Regression Analysis

- Regression Equation of y on x.
  - Find b and a using the given formula.

$$b = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n\bar{x}^2} \quad \text{and}$$

$$a = \bar{y} - b\bar{x}$$

x	y	xy	x^2
30	120	3600	900
40	130	5200	1600
50	140	7000	2500
60	150	9000	3600
70	160	11200	4900
$\Sigma x=250$	$\Sigma y=700$	$\Sigma xy=36000$	$\Sigma x^2=13500$

Regression Equation of y on x.

- $b = \dfrac{36000 - 5 \times 50 \times 140}{13500 - 5 \times (50)^2}$

$b = 36000 - 35000/13500 - 12500$

$b = 1000/1000 = 1$

$a = \bar{y} - b\bar{x}$

$a = 140 - 1 \times 50 = 90$

So the fitted regression equation is $y = a+bx$.

$BP = 90 + 1 \times 35 = 90 + 35 = 145mmHg.$

Regression Analysis: Example 2

- Fit the two line of regression equation for the following data.

X	10	20	30	40	50
Y	30	50	70	90	110

n = 5

$\overline{X} = \Sigma x/n = 150/5 = 30$

$\overline{Y} = \Sigma y/n = 350/5 = 70$

- The regression equation to be fitted is y = a+bx and

$x = a_1 + b_1 y$.

X 10 20 30 40 50

Y 30 50 70 90 110

Regression Analysis: Example 2

- Regression Equation of y on x.
  - Find b and a using the given formula

$$b = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - nx^2} \quad \text{and}$$

$$a = \bar{y} - b\bar{x}$$

x	y	xy	x^2	y^2
10	30	300	100	900
20	50	1000	400	2500
30	70	2100	900	4900
40	90	3600	1600	8100
50	110	5500	2500	12100
$\sum x = 150$	$\sum y = 350$	$\sum xy = 1250$	$\sum x^2 = 5500$	$\sum y^2 = 28500$

Regression Equation of y on x.

- $b = \dfrac{12500 - 5x30x70}{5500 - 5x(30)^2}$

  $b = 12500 - 10500 \div (5500 - 4500)$

  $= 2000 \div 1000$

  $= 2$

- • $a = y - bx$

  $a = 70 - 2 \times 30$

  $= 70 - 60$

  $= 10$

- So the fitted regression equation is $y = 10 + 2x$.

Regression Equation of x on y

- Find b_1, a_1 and a using the formula

$$b_1 = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma y^2 - ny^2}$$

$$a_1 = \bar{x} - b\bar{y}$$

- Regression Equation of y on x.

$b_1 = 12500 - 5x30x70 \div (28500 - 5x(70)^2)$

$b_1 = 12500 - 10500 \div (28500 - 24500)$

$b_1 = 2000 \div 4000 = 0.5$

$a_1 = x - b_1 y$

$= 30 - 0.5 \times 70 \quad = 30 - 35 = -5$

So the fitted regression equation is $x = -5 + 0.5y$.

Properties of Regression Coefficient

- The square root of the products of two regression coefficients is correlation coefficient.
- In the given examples

$b = byx = 2$

$b_1 = b_1xy = 0.5$

- $r = \sqrt{(2 \times 0.5)}$

$= \sqrt{1}$

$= 1$

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.
- Regression analysis describes the mathematical relationship between dependent variable (y) and independent variable (x) by estimating the unknown values of 'y' and for the known values of 'x' using the equation y = a+bx.

References

- Kothari, C. R., (2004) *Research Methodology, Methods and Techniques*, 2nd ed., New Age International Publishers, New Delhi.

- Measures of Relationship By; Mr. Johny Kutty Joseph Asstt. Professor
- References
- Clef, T. (2013). Exploratory Data Analysis in Business and Economics: An Introduction Using SPSS, Stata, and Excel. Springer Science