CORRELATION

Ansar Shahzadi

School of Electrical Engineering & Computer Science

National University of Science and Technology(NUST)

CORRELATION

- What is correlation
- What does its indicates?
- What does coefficient of Correlation Indicates ?
- Properties of Correlation Coefficients.
- Spearman Rank Correlation Coefficient (r_s)

CORRELATION

- Correlation is a statistical technique used to determine the degree to which two variables are related
- Measured with a correlation coefficient.
- Most popularly seen correlation coefficient: Pearson Product-Moment Correlation

TYPES OF CORRELATION

- Positive correlation
  - High values of X tend to be associated with high values of Y.
  - As X increases, Y increases
  - the length of an iron bar will increase as the temperature increases.
- Negative correlation
  - High values of X tend to be associated with low values of Y.
  - As X increases, Y decreases
  - the volume of gas will decrease as the pressure increase
- No correlation
  - No consistent tendency for values on Y to increase or decrease as X increases
  - If there is no relationship between the two variables such that the value of one variable change and the other variable remain constant

CORRELATION COEFFICIENT

A measure of the strength of the linear relationship between two variables that is defined in terms of the (sample) covariance of the variables divided by their (sample) standard deviations Represented by "r". For ungrouped data, we use following formula

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

OR

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

CORRELATION COEFFICIENT

- A measure of degree of relationship.
- r lies between 1 and -1.
- The sign of r denotes the nature of association .
- While the value of r denotes the strength of association.

Correlation

High positive correlation

Zero correlation

High negative correlation

stronger

+1.00

perfect positive

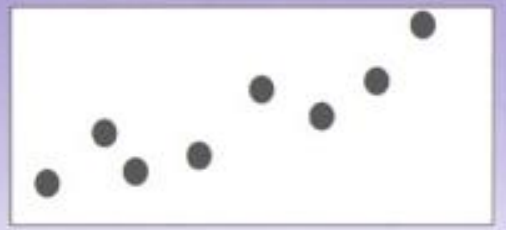
as one event increases, the second exactly increases

weaker

+.50

positive

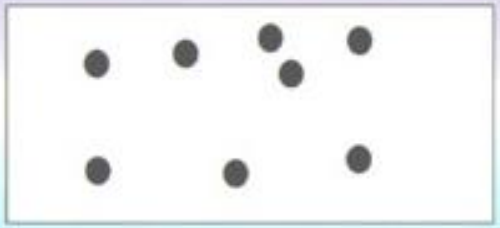
as one event increases, the second sometimes increases



0

zero correlation

no relationship between the events

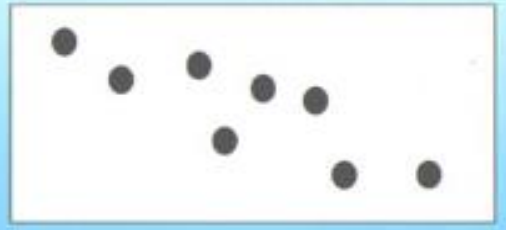


weaker

negative

-.50

as one event increases, the second sometimes decreases

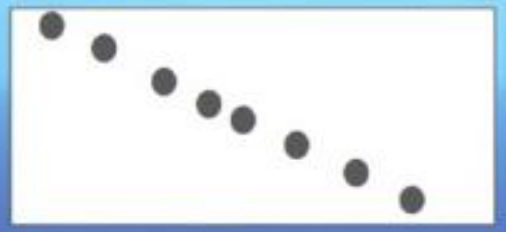


stronger

perfect negative

as one event increases, the second exactly decreases

-1.00



PROPERTIES OF CORRELATION COEFFICIENT

The most important properties of r are as follows:

- The value of r does not depend on which of the two variables under study is labeled X and which is labeled Y. Here $r_{XY} = r_{YX}$
- The value of r is independent of the units in which X and Y are measured.
- The value of r lies between 1 and -1.
- The correlation coefficient is independent of the origin and scale, i.e . $r_{XY} = r_{UV}$
- In case of a bivariate population where both X and Y are random variables, r is the geometric mean between the two regression coefficients, i.e $\mathrm{r} = \pm\sqrt{b_{xy}.b_{yx}}$

EXAMPLE 1

Calculate the correlation coefficient of the following data.
Also draw the scatter diagram.

X	78	89	97	69	59	68	61	79
Y	125	137	156	112	107	123	108	136

SOLUTION

X	Y	X^2	Y^2	XY
78	125	6084	15625	9750
89	137	7921	18769	12193
97	156	9409	24336	15132
69	112	4761	12544	7728
59	107	3481	11449	6313
68	123	4624	15129	8364
61	108	3721	11664	6588
79	136	6241	18496	10744
$\sum X$ =600	$\sum Y$ =1004	$\sum X^2$=46242	$\sum Y^2$=128012	$\sum XY$=76812

SOLUTION

Here n=8, $\sum X$ =600, $\sum Y$ =1004, $\sum \boldsymbol{X^2}$=46242, $\sum \boldsymbol{Y^2}$=128012 and $\sum XY$=76812

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

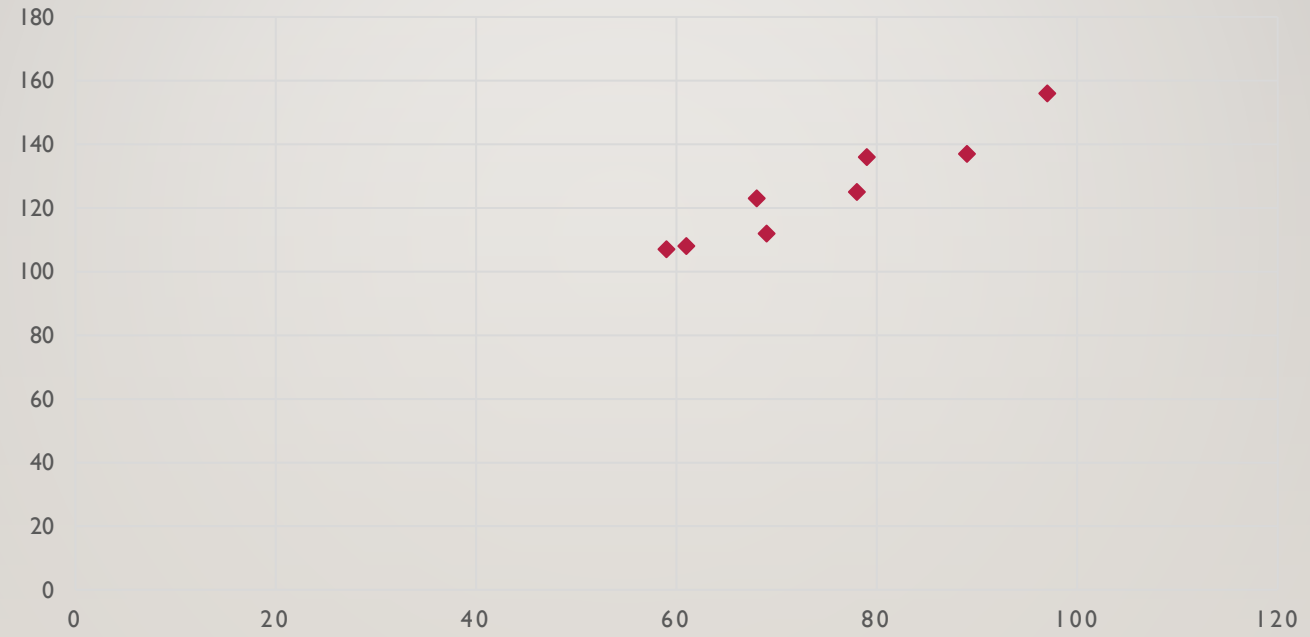
$$r=\frac{8(76812)-(600*1004)}{\sqrt{[8(46242)-(600)^2][8(128012)-(1004)^2}}$$

$$r=\frac{8(76812)-(600*1004)}{\sqrt{[8(46242)-(600)^2][8(128012)-(1004)^2}}$$

$$r=\frac{12096}{12640}=0.956958$$

Hence the value of r indicates strong correlation between X and Y.

SCATTER DIAGRAM



SPEARMAN RANK CORRELATION COEFFICIENT (r_s)

- Sometimes, the actual measurements or counts of individuals or objects are either not available or accurate assessment is not possible. They are then arrange according to some characteristics of interest. Such an ordered arrangement is called a ranking.
- This procedure makes use of the two sets of ranks that may be assigned to the sample values of X and Y.
- Spearman Rank correlation coefficient could be computed in the following cases:
  - Both variables are quantitative.
  - Both variables are qualitative ordinal.
  - One variable is quantitative and the other is qualitative ordinal.

PROCEDURE:

- Rank the values of X from 1 to n where n is the numbers of pairs of values of X and Y in the sample.
- Rank the values of Y from 1 to n.
- Compute the value of di for each pair of observation by subtracting the rank of Yi from the rank of Xi
- Square each di and compute $\sum {d_i}^2$ which is the sum of the squared values.
- Apply the following formula $r_s = 1 - \frac{6 \sum {d_i}^2}{n(n^2-1)}$
- The value of r_s denotes the magnitude and nature of association giving the same interpretation as simple r.

EXAMPLE#2

Find the coefficient of rank correlation from the following ranking of 10 students in statistics and mathematics.

Statistics(x)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Mathematics(y)	2	4	3	1	7	5	8	10	6	9

SOLUTION

x	y	d=x-y	d^2
1	2	-1	1
2	4	-2	4
3	3	0	0
4	1	3	9
5	7	-2	4
6	5	1	1
7	8	-1	1
8	10	-2	4
9	6	3	9
10	9	1	1
			$\sum d^2=34$

SOLUTION

Hence using Spearman Rank Correlation Coefficient f formula, we get

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

$$r_s = 1 - \frac{6(34)}{10(10^2-1)}$$

$$r_s = 1 - 0.2 = 0.8$$

The value of r indicates a high correlation between statistics and mathematics.

QUESTION#1

A computer while calculating the correlation coefficient between two variables *X* and *Y* from 25 pairs of observation obtained the following sums, $\sum X = 125,\ \sum Y = 100,$ $\sum XY = 508,\ \sum X^2 = 650,\ \sum Y^2 = 460.$ It was, however, later discovered at the time of checking that he had copied down two pairs as $(6,14)$ and $(8,6)$ while the corrected value were $(8,13)$ and $(5,7)$.Obtain the correct value of the coefficient of correlation.

QUESTION#2

Researchers interested in determining if there is a relationship between death anxiety and religiosity conducted the following study. Subjects completed a death anxiety scale (high score = high anxiety) and also completed a checklist designed to measure an individuals degree of religiosity (belief in a particular religion, regular attendance at religious services, number of times per week they regularly pray, etc.) (high score = greater religiosity. A data sample is provided below:

Death Anxiety	38	42	29	31	28	15	24	17	19	11	8	19	3	14	6
Religiosity	4	3	11	2	9	6	14	9	10	15	19	17	10	14	18