Correlation

S C Agarkar

VNBRIMS

Thane

The Concept of Correlation

- If two or more quantities vary in sympathy so that movements in the one tend to be accompanied by corresponding movements in other then they are said to be correlated (Connor). Examples are:
  - Supply and prices of vegetables
  - Study hours and performance in examination
  - Temperature and resistance of metals
  - Pressure and volume of an enclosed air
  - Amount of Carbon Dioxide and global temperature
  - Add some more

Historical Facts

- The theory of correlation analysis was first propounded by the French AstronomerBravis.
- Linear correlation theory was firstpopoundedby Sir Francis Galton.
- Karl Pearson propounded the mathematical method of calculating coefficient of correlation in 1896 and used it for problems related to Biology and Genetics.
- In 20thcentury this correlation coefficient entered into social sciences.

Pearson Correlation

- Pearson correlation coefficient is given by
- $r = \sum (X_i - x)(Y_i - y) / \sqrt{\sum (X_i - x)^2 \sum (Y_i - y)^2}$ Where
- r = Pearson correlation coefficient
- X = Values of X variables
- Y = Values of Y variables
- x= Arithmetic mean of X variables
- y= Arithmetic mean of Y variables

Calculate Pearson Correlation Coefficient

- Calculate Pearson coefficient of correlation from following data.
- X: 9 8 7 6 5 4 3 2 1; Y: 15 16 14 13 11 12 10 8 9

	X	(X-x)	$(X-x)^2$	Y	(Y-y)	$(Y-y)^2$	(X-x)(Y-y)
1	9	4	16	15	3	9	12
2	8	3	09	16	4	16	12
3	7	2	04	14	2	4	4
4	6	1	01	13	1	1	1
5	5	0	00	11	-1	1	0
6	4	-1	01	12	0	0	0
7	3	-2	04	10	-2	4	4
8	2	-3	09	8	-4	16	12
9	1	-4	16	9	-3	9	12
Sum	45		60	108		60	57

$r = \sum (X_i-x)(Y_i-y)/\sqrt{\sum (X_i-x)^2 \sum (Y_i-y)^2}$

$= 57/\sqrt{60*60} = 0.95$

Problem to solve

- Calculate Pearson coefficient of correlation from following data.
- Adv Cost (in thousands): 39 65 62 90 82 75 25 08 36 78
- Sales (inlakhs): 47 53 58 86 62 68 60 91 51 84

	X	(X-x)	$(X-x)^2$	Y	(Y-y)	$(Y-y)^2$	(X-x)(Y-y)
1	39	-17		47			
2	65	09		53			
3	62	06		58			
4	90	34		86			
5	82	26		62			
6	75	19		68			
7	25	-31		60			
8	08	-48		91			
9	36	-20		51			
10	78	22		84			

$r = \sum (X_i-x)\,(Y_i-y) / \sqrt{\sum (X_i-x)^2 \sum (Y_i-y)^2}$

Ans: +0.117

Spearman Rank Order Correlation

- Step 1: Assigning ranks to all the values of x as well as y variables
- Step 2: In case of repeated values mid rank is to be assigned
- Step 3: Calculating rank differences
- Step 4: Calculating square of deviations
- Step 5: Use the formula

$r_R = 1 - 6\sum D^2/N^3 - N$

r_R = Rank Correlation coefficient

$\sum D^2$ = Total of square of rank differences

N = Number of pairs of observation.

- Step 6: Add $(m^3 - m)/12$ to $\sum D^2$. Here m means the number of times an item has repeated. This correction factor is to be added for each repeated value.

Illustrative Example

- Two teachers were asked to rank seven students on the basis of their singing competitions. The marks assigned to these students out of maximum marks 50 are given below. Calculate Spearman's rank correlation coefficient.
- Student: 1 2 3 4 5 6 7
- Teacher 1: 39 43 45 42 36 32 28
- Teacher 2: 46 41 44 40 38 36 32

Calculating Spearman Correlation Coefficient

Student	Teacher 1	Rank R_1	Teacher 2	Rank R_2	$R_1 - R_2$	$D^2 = (R_1 - R_2)^2$
1	39	4	46	1	3	9
2	43	2	41	3	-1	1
3	45	1	44	2	-1	1
4	42	3	40	4	-1	1
5	36	5	38	5	0	0
6	32	6	36	6	0	0
7	28	7	32	7	0	0

$r_R = 1 - 6\sum D^2/N^3 - N$

$= 1 - (6*12)/343-7$

$= 1 - 0.214$

$= 0.786$

Problem to Solve

- Find the rank correlation coefficient for the following
- Marks in Statistics

48 60 72 62 56 40 39 52 30

- Marks in Accountancy

62 78 65 70 38 54 60 32 31

Hint: Prepare the table with values of Marks in Statistics and in Accountancy. Then give rank, get the difference in the ranks, square it, add all squared values and substitute in the formula.

Ans: + 0.66

Repeated values

- Calculate the coefficient of rank correlation from following data.

X: 48 33 40 9 16 16 65 24 16 57

Y: 13 13 24 6 15 4 20 9 6 19

X	Y	Rank R_1	Rank R_2	$R_1 - R_2$	$(R_1 - R_2)^2$
48	13	3	5.5	-2.5	6.25
33	13	5	5.5	-0.5	0.25
40	24	4	1	3	9.00
9	6	10	8.5	1.5	2.25
16	15	8	4	4	16.00
16	4	8	10	-2	4.00
65	20	1	2	-1	1.00
24	9	6	7	-1	1.00
16	6	8	8.5	-0.5	0.25
57	**19**	**2**	**3**	**-1**	**1.00**

Calculation

$\sum D^2 = 41$

Repeated X values 3 (one time)

Repeated Y values 2 (two times)

$r_R = 1 - 6\ [\sum D^2 + 1/12\ (m^3 - m) + 1/12\ (m^3 - m) + 1/12\ (m^3 - m)]/N^3 - N$

$= 1 - 6\ [41 + 1/12\ (3^3 - 3) + 1/12\ (2^3 - 2) + 1/12\ (2^3 - 2)/\ (10^3 - 10)$

$= 1 - 6\ {}^*44/\ 1000 - 10$

$= 1 - 0.266$

$= 0.734$

Problem to solve

- Calculate Spearman's coefficient rank correlation from the following data
- X: 57 16 24 65 16 16 09 40 33 48
- Y: 19 06 09 20 04 15 06 24 13 13
- Hint: Prepare the table giving the rank to each value of X and Y. Assign average value to the values that get repeated. Calculate difference in ranks, square them and add. Use this value in the formula to get the answer (0.73)

Partial Correlation

- In partial correlation we study the effect of one variable on a dependent variable by excluding the effect of other variables. For example if price is affected by demand, income and exports, we can study the relationship between price and demand excluding the effect of income and exports.
- Simple correlation between two variables is called the Zero order coefficient (no factor is held constant). When partial correlation is studied between two variables by keeping the third variable constant it is a first order coefficient. When two variables are kept constant it is a second order correlation.

An Example

- The following zero order correlation coefficients are given r_{12}= 0.98, r_{13}= 0.44, r_{23}= 0.54. Calculate the partial coefficient correlation between first and third variable keeping the effect of second constant.
- $R_{13.2} = r_{13} - r_{12}r_{23} / \sqrt{1 - r^2_{12}} \sqrt{1 - r^2_{23}}$
- $= 0.44 - (0.98 * 0.54) / \sqrt{1 - (.98)^2} \sqrt{1 - (.54)^2}$
- $= -0.0892/0.199 * 0.842$
- $= - 0.5322$

Multiple Correlation

- In case of multiple correlation the effect of all the independent factors on a dependent factor is studied. It is calculated using the values of zero order correlations.
- If the values of r_{12}, r_{13} , r_{23} are 0.98, 0.44 and 0.54 respectively, then calculate multiple correlation coefficient treating first variable as dependent and second and third as independent.
- $R_{1.23} = \sqrt{r^2_{12} + r^2_{13} - 2\, r_{12} r_{13} r^2_{23} / 1 - r^2_{23}}$
- $= \sqrt{(.98)^2 + (.44)^2 - 2\ .98 * .44 * ,54\ /\ 1 - (.54)^2}$
- $= 0.986$

Thank you, What correlation do you see below?