e-text

Paper-CC10 (*U-IV*)

Research Methodology and Quantitative Techniques

Spearman's Rank Correlation Coefficient

Dr. Supriya

Assistant Professor (Guest)

Ph. D: Geography; M.A. in Geography

Post Doc. Fellow (ICSSR), UGC- NET-JRF

Department of Geography

Patna University, Patna

Mob: 9006640841

Email: supriyavatsa52256@gmail.com

Content Writer & Affiliation	***Dr Supriya, Asst. Professor (Guest), Patna University***
Subject Name	***Geography***
Paper Code	***CC-12***
Paper Name	***Human and Social Geography***
Title of Topic	***Spearman's Rank Correlation Coefficient***
Objectives	***To understand the method of determination of Rank Correlation by Spearman***
Keywords	***Correlation, Coefficient, Ranking, Negative, Positive***

Spearman's Rank Correlation Coefficient

Dr. Supriya

Introduction: The Spearman's rank correlation coefficient (r_s) is a method of testing the strength and direction (positive or negative) of the correlation (relationship or connection) between two variables.

In statistics, Spearman's rank correlation coefficient or Spearman's ρ, named after Charles Spearman and often denoted by the Greek letter ρ or as (rho) or as r_s, is a nonparametric measure of rank correlation. rho is statistical dependence between the rankings of two variables. It assesses how well the relationship between two variables can be described using a monotonic function.

The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not). If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other.

Intuitively, the Spearman correlation between two variables will be high when observations have a similar (or identical for a correlation of 1) rank, (i.e. relative position label of the observations within the variable: 1st, 2nd, 3rd, etc.) between the two variables, and low when observations have a dissimilar (or fully opposed for a correlation of −1) rank between the two variables.

Spearman's coefficient is appropriate for both continuous and discrete ordinal variables. Both Spearman's ρ (rho) and Kendall's τ (tau) can be formulated as special cases of a more general correlation coefficient.

Definition and calculation

The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the rank variables.

For a sample of size n, the n raw scores $X_{i,}$ Y_i are converted to ranks $rg_{Xi,}$ rg_{Yi} and r_s is computed as

$$r_s = \rho_{rg_X, rg_Y} = \frac{\operatorname{cov}(rg_X, rg_Y)}{\sigma_{rg_X} \sigma_{rg_Y}},$$

where

ρ denotes the usual Pearson correlation coefficient, but applied to the rank variables,

$\operatorname{Cov}(rg_x, rg_y)$ is the covariance of the rank variables,

σ_{rgx} and σ_{rgy} are the standard deviations of the rank variables.

Only if all n ranks are distinct integers, it can be computed using the popular formula

$$r_s = 1 - \frac{6 \sum d^2}{N(N^2 - 1)}$$

Where

D= the difference between the two ranks of each observation,

N= is the number of observations.

Identical values are usually assigned to each fractional ranks equal to the average of their positions in the ascending order of the values, which is equivalent to averaging over all possible permutations.

LIMITATIONS:

- The data must be linear (draw a scatter graph with the line of best fit)
- The data must be independent from each other (ex: HDI and Fertility does NOT work because HDI is calculated using Fertility)
- There should be between 10 and 30 pairs of data

Note that a strong correlation does not necessarily mean cause and effect. Ex: % of households owning a camera and % of person dying of lung cancer in the United States are both increasing between 1950 and 2000... But both elements are clearly NOT correlated!

STEPS OF CALCULATING CORRELATION BETWEEN TWO DATASET OF VARIABLES:

Step 1: If you wanted to find relationship among total irrigated land and the productivity of wheat; could start by suggesting that there will be no relationship between area and crop production. This will be called **Null Hypothesis**.

To begin with Rho test from collected data from selected samples/districts you have to collect two sets of data for each of the samples/district of both variables total irrigated area and wheat production. To avoid working with decimal points, turned all the figures from to and convert them into whole numbers. For example 98.1 becomes 98.

State	Percentage Cultivated area under irrigation	R1	Productivity	R2
Punjab	98.1		4.2	
Haryana	87.6		3.3	
Uttar Pradesh	75.9		2.3	
Andhra Pradesh	63.9		2.7	
Bihar	63.4		1.7	
Tamil Nadu	63.1		2.2	
West Bengal	48.2		2.4	
Gujarat	44.7		1.5	
Madhya Pradesh	44.5		1.1	
Uttarakhand	42.9		1.7	
Orissa	33.6		1.3	
Karnataka	28.5		1.5	
Chattisgarh	27.6		1	
Rajasthan	26.4		1.2	
Maharashtra	16.8		1	
Jharkhand	5.4		1.7	
Assam	4.9		1.5	

Step 2: Than need to rank each set of data, to do this separately for each of the pairs of data. Look for the lowest values in a column set to start, then rank from lowest to highest for the entire column.

State	Percentage Cultivated area under irrigation	R1	Productivity	R2
Punjab	98.1		4.2	
Haryana	87.6		3.3	
Uttar Pradesh	75.9		2.3	
Andhra Pradesh	63.9		2.7	
Bihar	63.4		1.7	
Tamil Nadu	63.1		2.2	
West Bengal	48.2		2.4	
Gujarat	44.7		1.5	
Madhya Pradesh	44.5		1.1	
Uttarakhand	42.9		1.7	
Orissa	33.6		1.3	
Karnataka	28.5		1.5	
Chattisgarh	27.6		1	
Rajasthan	26.4		1.2	
Maharashtra	16.8		1	
Jharkhand	5.4		1.7	
Assam	4.9		1.5	

Step 3: There are two empty columns to the right hand side of your table. now going to use those columns. Label the first of your two columns d (R1- R2)' .In this column, for each pair of data, subtract the rank r2 from the rank r1 to give d. d stands for difference.Note that some figures will be minus (-) and some will be plus (+) figures.

State	Percentage Cultivated area under irrigation	R1	Productivity	R2	d(R1-R2)
Andhra Pradesh	64	4	3	3	1
Assam	5	17	2	12	5
Bihar	63	5	2	7	-2
Chattisgarh	28	13	1	16	-3
Gujarat	45	8	2	10	-2
Haryana	88	2	3	2	0
Jharkhand	5	16	2	9	7
Karnataka	29	12	2	11	1
Madhya Pradesh	45	9	1	15	-6
Maharashtra	17	15	1	17	-2
Orissa	34	11	1	13	-2
Punjab	98	1	4	1	0

Rajasthan	26	14	1	14	0
Tamil Nadu	63	6	2	6	0
Uttar Pradesh	76	3	2	5	-2
Uttarakhand	43	10	2	8	2
West Bengal	48	7	2	4	3
		153		**153**	**0**

Step 4: Now complete the last column in the table. Simply square the value of d (To square something is to multiply a number by itself. We do this procedure to get rid of all the negative values (a minus value multiplied by a minus value is a plus value).

State	Percentage Cultivated area under irrigation	R1	Productivity	R2	d(R1-R2)	d2
Andhra Pradesh	64	4	3	3	1	1
Assam	5	17	2	12	5	25
Bihar	63	5	2	7	-2	4
Chattisgarh	28	13	1	16	-3	9
Gujarat	45	8	2	10	-2	
Haryana	88	2	3	2	0	
Jharkhand	5	16	2	9	7	
Karnataka	29	12	2	11	1	
Madhya Pradesh	45	9	1	15	-6	
Maharashtra	17	15	1	17	-2	
Orissa	34	11	1	13	-2	
Punjab	98	1	4	1	0	
Rajasthan	26	14	1	14	0	
Tamil Nadu	63	6	2	6	0	
Uttar Pradesh	76	3	2	5	-2	
Uttarakhand	43	10	2	8	2	
West Bengal	48	7	2	4	3	
Total						

Step 5: You should now have a go at completing the whole table here to make sure you understand the process before you go on to step six.

State	Percentage Cultivated area under irrigation	R1	Productivity	R2	d(R1-R2)	d2
Andhra Pradesh	64	4	3	3	1	1
Assam	5	17	2	12	5	25
Bihar	63	5	2	7	-2	4
Chattisgarh	28	13	1	16	-3	9

Gujarat	45	8	2	10	-2	4
Haryana	88	2	3	2	0	0
Jharkhand	5	16	2	9	7	49
Karnataka	29	12	2	11	1	1
Madhya Pradesh	45	9	1	15	-6	36
Maharashtra	17	15	1	17	-2	4
Orissa	34	11	1	13	-2	4
Punjab	98	1	4	1	0	0
Rajasthan	26	14	1	14	0	0
Tamil Nadu	63	6	2	6	0	0
Uttar Pradesh	76	3	2	5	-2	4
Uttarakhand	43	10	2	8	2	4
West Bengal	48	7	2	4	3	9
Total						**154**

Step 6 :This is what your table should now look like; Now apply the Spearmans Rank Equation using the figures in your table.

$$R = 1 - \frac{6\sum d^2}{n(n^2 - n)}$$

Here is how this would worked through

$R = 1 - 6(154) / 17\text{-}(17^2\text{-}17)$

R= 1-6(154)/17(289-17)

R= 1-6(154)/17(272)

R= 1-924/4624

R= 1- 0.1998

R= 0.800

Step 7: Is the correlation a positive or negative one?

The value will always come out between -1.0 and +1.0. What does this mean? Negative correlations are those between -1 and zero. Values of less than zero, i.e. those with a minus value show a negative correlation. This means that, as one value rises, the other value falls. An example of negative correlations in geography might be:

Life expectancy increases as % of medical facilty increases

A **perfect negative correlation** would be -1. A figure near to zero (e.g. -0.3) would be considered to be weak. for figures of between -0.7 and -1 to indicate a strong negative correlation

Positive correlations are those between +1 and zero. Values of more than zero, i.e. those with a plus value, show a positive correlation. This means that as one value rises the other value rises or as one value falls the other value falls. An example of positive correlations in geography might be: The larger the town the greater the number of retail outlets it has

In above example, as our R value is + 0.8; we can see that There is a positive correlation and that correlation is considered to be strong.

Step eight: Can we trust our result?

When you are using statistics you need to be very careful to check that you have a meaningful result and not one which is just down to chance.

There are two checks after calculation :

1. Have you used more than 10 sets of data? If not your sample may not be representative.

2. Check to see if final figure registers as at least good enough to be **confident** of, on the **significance table**. Look for the column shown as 98% confidence (or sometimes listed as 0.5 level or 5%) If your result comes out as higher than the number shown for the number of data sets used then you can be confident the result is not down to random chance. 95% is a pretty high rate of confidence, so that is good enough. Our result of 0.8 is above the figures shown for 17 sets of data at the 95% confidence level . (0.475) so we can trust that this result is significant. We can be confident it is not down to chance.

Significance level table

Number of pairs of data (n)	10% chance 90% confident	5% chance 95% confident	2% chance 98% confident	1% chance 99% confident
5	0.9	1	1	1
6	0.829	0.886	0.943	1
7	0.714	0.786	0.893	0.929
8	0.643	0.738	0.833	0.881
9	0.6	0.683	0.783	0.833
10	0.564	0.648	0.746	0.794
12	0.506	0.591	0.712	0.777
14	0.456	0.544	0.645	0.715
16	0.425	0.506	0.601	0.665
18	0.399	0.475	0.564	0.625
20	0.377	0.45	0.534	0.591
22	0.359	0.428	0.508	0.562
24	0.343	0.409	0.485	0.537
26	0.329	0.392	0.465	0.515
28	0.317	0.377	0.448	0.496
30	0.306	0.364	0.432	0.478

------*------*------*-------