A POSSIBLE IMPROVEMENT IN SPEARMAN'S RANK CORRELATION COEFFICIENT, WHEN KARL PEARSON'S CORRELATION COEFFICIENT IS NEGATIVE

GURMEET SINGH, RUPINDERJEET KAUR

Abstract: A Non Parametric Rank Correlation, proposed by Spearman is a measure of the strength between two variables. It is a monotone association used in distribution of data when the data is misleading. It is not a measure of Linear Relationship between variables. Unlike Pearson's Product Moment Correlation coefficient, it does not require the assumption that the relationship between variables is Linear. Nor does it require any interval scale for measurability of variables. The idea of this paper is to compare the values of Pearson's Product Moment Correlation coefficient and Spearman's Rank Correlation Coefficient, in case, the Pearson's Product Moment Correlation coefficient is negative.

Keywords: Pearson's Product Moment Correlation coefficient, Rank Correlation Coefficient, Comparison of Correlation Coefficient, Linear Relationship of Variables.

Introduction: Historical Background

Correlations between variables can be measured with the use of different indices (coefficients). The two most popular are: Pearson's coefficient(r) and Spearman's rho coefficient (r_s).

The idea of the paper is to compare the values of Pearson's correlation coefficient treating data in a quantitative way versus the values of Spearman's rank correlation coefficient treating the same data in a somewhat 'qualitative' way for real data sets, when the Product Moment Correlation coefficient is negative. Coming back to the history of developing the idea of measuring correlation strength, one should mention the set of papers by Galton, Yule and Pearson (listed in the references) which created the basis for a proper and correct application and interpretation of correlations (in the modern meaning of the word).

The history and properties of Pearson's correlation coefficient were also described by Pearson (1920), Weida (1927), Walker (1928), Stigler (1988), and Piovani (2008).

Pearson's coefficient of correlation was discovered by Bravais in 1846, but Karl Pearson was the first to describe, in 1896, the standard method of its calculation and show it to be the best one possible. Pearson also offered some comments about an extension of the idea made by Galton (who applied it to anthropometric data). He called this method the "product-moments" method (or the Galton function for the coefficient of correlation r). An important assumption in Pearson's 1896 contribution is the normality of the variables analyzed, which could be true only for quantitative variables. Pearson's correlation coefficient is a measure of the strength of the linear relationship between two such variables.

In 1904 Spearman adopted Pearson's correlation coefficient as a measure of the strength of the relationship between two variables that cannot be measured quantitatively. He noted: "The most fundamental requisite is to be able to measure our observed correspondence by a plain numerical symbol. There is no reason whatever to be satisfied either with vague generalities such as "large", "medium", "small," or, on the other hand, with complicated tables and compilations.

Spearman's rank correlation coefficient is a nonparametric (distribution-free) rank statistic proposed as a measure of the strength of the association between two variables. It is a measure of a monotone association that is used when the distribution of data makes Pearson's correlation coefficient undesirable or misleading. Spearman's coefficient is not a measure of the linear relationship between two variables. It assesses how well an arbitrary monotonic function can describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables. Unlike Pearson's product-moment correlation coefficient, it does not require the assumption that the relationship between the variables is linear, nor does it require the variables to be measured on interval scales; it can be used for variables measured at the ordinal level. In principle, r_s is simply a special case of Pearson's product-moment coefficient in which the data are converted to ranks before calculating the coefficient. Spearman's statistical accomplishments of 1904 were not appreciated by his University College colleague Karl Pearson, and there was a long-standing disagreement between them. The history and subsequent practice showed that it was Spearman who was right, and nowadays coefficient r_s is widely used in statistical analyses.

The use of Pearson's product-moment correlation coefficient and Spearman's rank correlation coefficient for geographical data was examined by Haining (1991) and Griffith (2003).

In the present paper we would like to compare the values and significance of Pearson's and Spearman's coefficients for the same sets of data (original data for r and ranked data for r_s) when the Product Moment Correlation Coefficient is negative.

MAIN RESULTS: We give some examples of negative correlation, where there is a significant difference between the values of Product Moment Correlation coefficient and Rank Correlation Coefficient. We will also give a new formula for Corrected Rank Correlation and through examples, it will be visual that Corrected Coefficient of Rank Correlation is much closer to Product Moment Correlation coefficient than the original Rank Correlation, which is the aim of this research paper.

Spearman's Correlation coefficient formula is

$$r = 1 - \frac{6\left(\sum D^2 + \sum\left(\frac{m^3-m}{12}\right)\right)}{n^3-n}$$

Corrected Rank Correlation Coefficient formula when the Product Moment Correlation Coefficient is negative is

$$r = 1 - \frac{6(\sum D^2 + \sum(\frac{m^3-m}{6}))}{n^3-n}$$

We start with the following example:

Karl Pearson's Method: $r = \frac{\frac{\sum XY}{N} - \bar{X}\bar{Y}}{\sqrt{\left(\frac{\sum X^2}{N} - \bar{X}^2\right)\left(\frac{\sum Y^2}{N} - \bar{Y}^2\right)}}$

$$= \frac{\frac{122}{8} - (4.5)(4.5)}{\sqrt{\left(\frac{202}{8} - 4.5^2\right)\left(\frac{202}{8} - 4.5^2\right)}} = -1$$

Spearman's Method: $r = 1 - \frac{6\left(\sum D^2 + \sum\left(\frac{m^3-m}{12}\right)\right)}{n^3-n}$

R_X	R_Y	R_X^2	R_Y^2	$R_X R_Y$	D^2
1	8	1	64	8	49
2	7	4	49	14	25
4	5	16	25	20	1
4	5	16	25	20	1
4	5	16	25	20	1
6	3	36	9	18	9
7	2	49	4	14	25
8	1	64	1	8	49
		$R_X^2 = 202$	$R_Y^2 = 202$	$R_X R_Y = 122$	$D^2 = 160$

$$= 1 - \frac{6\left(160 + \left(\frac{3^3-3}{12} + \frac{3^3-3}{12}\right)\right)}{8^3-8} = -0.95238095$$

New Method: $r = 1 - \frac{6(\sum D^2 + \sum(\frac{m^3-m}{6}))}{n^3-n}$ $= 1 - \frac{6(160 + \left(\frac{3^3-3}{6} + \frac{3^3-3}{6}\right))}{8^3-8} = 1 - 2 = -1$

Let us see another example:

Karl Pearson's Method: $r = \frac{\frac{\sum XY}{N} - \bar{X}\bar{Y}}{\sqrt{\left(\frac{\sum X^2}{N} - \bar{X}^2\right)\left(\frac{\sum Y^2}{N} - \bar{Y}^2\right)}}$

$$= \frac{\frac{230}{10} - (5.5)(5.5)}{\sqrt{\left(\frac{375}{10} - 5.5^2\right)\left(\frac{375}{10} - 5.5^2\right)}} = \frac{-7.25}{\sqrt{(7.25)(7.25)}} = -1$$

R_X	R_Y	R_X^2	R_Y^2	$R_X R_Y$	D^2
1	10	1	100	10	81
2	9	4	81	18	49
5	6	25	36	30	1
5	6	25	36	30	1
5	6	25	36	30	1
5	6	25	36	30	1
5	6	25	36	30	1
8	3	64	9	24	25
9	2	81	4	18	49
10	1	100	1	10	81
		$R_X^2 = 375$	$R_Y^2 = 375$	$R_X R_Y = 230$	$D^2 = 290$

Spearman's Method: $r = 1 - \frac{6\left(\sum D^2 + \sum\left(\frac{m^3-m}{12}\right)\right)}{n^3-n}$

$$= 1 - \frac{6(290 + \left(\frac{5^3-5}{12} + \frac{5^3-5}{12}\right))}{10^3-10} = -0.878787$$

New Method: $r = 1 - \frac{6(\sum D^2 + \sum(\frac{m^3-m}{6}))}{n^3-n}$

$$=1 - \frac{6(290 + \left(\frac{5^3-5}{6} + \frac{5^3-5}{6}\right))}{10^3-10} = 1 - 2 = -1$$

Let us take another example:

R_X	R_Y	R_X^2	R_Y^2	$R_X R_Y$	D^2
1	10	1	100	10	81
2	9	4	81	18	49
6	5	36	25	30	1
6	5	36	25	30	1
6	5	36	25	30	1
6	5	36	25	30	1
6	5	36	25	30	1
6	5	36	25	30	1
6	5	36	25	30	1
10	1	100	1	10	81
		$R_X^2 = 357$	$R_Y^2 = 357$	$R_X R_Y = 248$	$D^2 = 218$

Karl Pearson's Method: $r = \frac{\frac{\sum XY}{N} - \bar{X}\bar{Y}}{\sqrt{\left(\frac{\sum X^2}{N} - \bar{X}^2\right)\left(\frac{\sum Y^2}{N} - \bar{Y}^2\right)}}$

$$= \frac{\frac{248}{10} - (5.5)(5.5)}{\sqrt{\left(\frac{357}{10} - 5.5^2\right)\left(\frac{357}{10} - 5.5^2\right)}} = -1$$

Spearman's Method:

$$r = 1 - \frac{6(\sum D^2 + \sum(\frac{m^3 - m}{12}))}{n^3 - n}$$

$$= 1 - \frac{6(218 + \left(\frac{5^3 - 5}{12} + \frac{5^3 - 5}{12}\right))}{10^3 - 10} = -0.66060606$$

New Method: $r = 1 - \frac{6\left(\sum D^2 + \sum\left(\frac{m^3 - m}{6}\right)\right)}{n^3 - n}$

$$= 1 - \frac{6(218 + \left(\frac{5^3 - 5}{6} + \frac{5^3 - 5}{6}\right))}{10^3 - 10} = -1$$

We have another example:

R_X	R_Y	R_X^2	R_Y^2	$R_X R_Y$	D^2
1	8	1	64	8	49
2	7	4	49	14	25
4	5	16	25	20	1
4	5	16	25	20	1
4	5	16	25	20	1
7	2	49	4	14	25
7	2	49	4	14	25
7	2	49	4	14	25
		$R_X^2 = 200$	$R_Y^2 = 200$	$R_X R_Y = 124$	$D^2 = 152$

Karl Pearson's Method: $r = \frac{\frac{\sum XY}{N} - \bar{X}\bar{Y}}{\sqrt{\left(\frac{\sum X^2}{N} - \bar{X}^2\right)\left(\frac{\sum Y^2}{N} - \bar{Y}^2\right)}}$

$$= \frac{\frac{124}{8} - (4.5)(4.5)}{\sqrt{\left(\frac{200}{8} - 4.5^2\right)\left(\frac{200}{8} - 4.5^2\right)}} = -1$$

Spearman's Method: $r = 1 - \frac{6\left(\sum D^2 + \sum\left(\frac{m^3 - m}{12}\right)\right)}{n^3 - n}$

$$= 1 - \frac{6\left(152 + \left(\frac{3^3 - 3}{12} + \frac{3^3 - 3}{12} + \frac{3^3 - 3}{12} + \frac{3^3 - 3}{12}\right)\right)}{8^3 - 8} = -0.9047619048$$

New Method: $r = 1 - \frac{6(\sum D^2 + \sum(\frac{m^3 - m}{6}))}{n^3 - n}$

$$= 1 - \frac{6(152 + \left(\frac{3^3 - 3}{6} + \frac{3^3 - 3}{6} + \frac{3^3 - 3}{6} + \frac{3^3 - 3}{6}\right))}{8^3 - 8} = -1$$

We have another example:

R_X	R_Y	R_X^2	R_Y^2	$R_X R_Y$	D^2
1	12	1	144	12	121
2	11	4	121	22	81
4	9	16	81	36	25
4	9	16	81	36	25
4	9	16	81	36	25
8	5	64	25	40	9
8	5	64	25	40	9
8	5	64	25	40	9
8	5	64	25	40	9
8	5	64	25	40	9
11	2	121	4	22	81
12	1	144	1	12	121
		$R_X^2 = 638$	$R_Y^2 = 638$	$R_X R_Y = 376$	$D^2 = 524$

Karl Pearson's Method: $r = \frac{\frac{\sum XY}{N} - \bar{X}\bar{Y}}{\sqrt{\left(\frac{\sum X^2}{N} - \bar{X}^2\right)\left(\frac{\sum Y^2}{N} - \bar{Y}^2\right)}}$

$$= \frac{\frac{376}{12} - (6.5)(6.5)}{\sqrt{\left(\frac{638}{12} - 6.5^2\right)\left(\frac{638}{12} - 6.5^2\right)}} = -1$$

Spearman's Method: $r = 1 - \frac{6(\sum D^2 + \sum(\frac{m^3 - m}{12}))}{n^3 - n}$

$$= 1 - \frac{6(524 + \left(\frac{3^3 - 3}{12} + \frac{3^3 - 3}{12} + \frac{5^3 - 5}{12} + \frac{5^3 - 5}{12}\right))}{12^3 - 12} = -0.9160839161$$

New Method: $r = 1 - \frac{6\left(\sum D^2 + \sum\left(\frac{m^3 - m}{6}\right)\right)}{n^3 - n}$

$$= 1 - \frac{6(524 + \left(\frac{3^3 - 3}{6} + \frac{3^3 - 3}{6} + \frac{5^3 - 5}{6} + \frac{5^3 - 5}{6}\right))}{12^3 - 12} = -1$$

Let us take another example:

$Judge_1$	$Judge_2$	R_X	R_Y	R_X^2	R_Y^2	$R_X R_Y$	D^2
1	4	1.5	7.5	2.25	56.25	11.25	36
2	3	3.5	5	12.25	25	17.5	2.25
3	3	5.5	5	30.25	25	27.5	.25
4	2.5	7.5	2	56.25	4	15	30.25
1	4	1.5	7.5	2.25	56.25	11.25	36
2	3	3.5	5	12.25	25	17.5	2.25
3	2.5	5.5	2	30.25	4	11	6.25
4	2.5	7.5	2	56.25	4	15	30.25
				[illegible] = 202	[illegible] = 199.5	[illegible] = 126.25	[illegible] = 143.5

Karl Pearson's Method: $\square = \frac{\frac{\sum\square\square}{\square}-\bar{\square}\bar{\square}}{\sqrt{\left(\frac{\sum\square^2}{\square}-\bar{\square}^2\right)\left(\frac{\sum\square^2}{\square}-\bar{\square}^2\right)}}$

$= \frac{\frac{126.25}{8}-(4.5)(4.5)}{\sqrt{\left(\frac{202}{8}-4.5^2\right)\left(\frac{199.5}{8}-4.5^2\right)}} = -0.9230610308$

Spearman's Method: $\square = 1 - \frac{6(\sum\square^2+\sum(\frac{\square^3-\square}{12}))}{\square^3-\square}$

$= 1 - \frac{6(143.5+\left(\frac{2^3-2}{12}+\frac{2^3-2}{12}+6\,\square\square\square\square\square\right))}{8^3-8} =-0.7857142857$

New Method: $\square = 1 - \frac{6(\sum\square^2+\sum(\frac{\square^3-\square}{6}))}{\square^3-\square}$

$\backslash\ = 1 - \frac{6\left(143.5+\left(\frac{2^3-2}{6}+\frac{2^3-2}{6}+6\ times\right)\right)}{8^3-8}$

$= -0.8630952381$

Let us take another example:

Karl Pearson's Method: $r = \frac{\frac{\sum XY}{N}-\bar{X}\bar{Y}}{\sqrt{\left(\frac{\sum X^2}{N}-\bar{X}^2\right)\left(\frac{\sum Y^2}{N}-\bar{Y}^2\right)}}$

$= \frac{\frac{230}{10}-(5.5)(5.5)}{\sqrt{\left(\frac{375}{10}-5.5^2\right)\left(\frac{383}{10}-5.5^2\right)}} = -0.94901064$

Spearman's Method: $r = 1 - \frac{6(\sum D^2+\sum(\frac{m^3-m}{12}))}{n^3-n}$

$= 1 - \frac{6(298+\left(\frac{3^3-3}{12}+\frac{5^3-5}{12}\right))}{10^3-10} = -0.878787878$

New Method:

$r = 1 - \frac{6\left(\sum D^2+\sum\left(\frac{m^3-m}{6}\right)\right)}{n^3-n}$

R_X	R_Y	R_X^2	R_Y^2	R_XR_Y	D^2
1	10	1	100	10	81
2	9	4	81	18	49
3	8	9	64	24	25
6	6	36	36	36	0
6	6	36	36	36	0
6	6	36	36	36	0
6	4	36	16	24	4
6	3	36	9	18	9
9	2	81	4	18	49
10	1	100	1	10	81
		$R_X^2 = 375$	$R_Y^2 = 383$	$R_XR_Y = 230$	$D^2 = 298$

$= 1 - \frac{6(298+\left(\frac{3^3-3}{6}+\frac{5^3-5}{6}\right))}{10^3-10}$

$= -0.951515$

Conclusion: From the above examples, it is clear that the formula

$$r = 1 - \frac{6(\sum D^2+\sum(\frac{m^3-m}{6}))}{n^3-n}$$

gives the Coefficient of Correlation more closer to Product Moment Correlation coefficient than the Original Spearman's Formula $r = 1 - \frac{6(\sum D^2+\sum(\frac{m^3-m}{12}))}{n^3-n}$ when Product Moment Correlation coefficient is negative.

References:

1. Anderson T.W., 1996. R.A. Fisher and multivariate analysis.Statistical Science 11 (1): 20–34.
2. Bravais A.,1846. Analyse mathématique sur les probabilities des erreurs de situation d'un point. Mémoires présentés par divers savants à l'Académie Royale des Sciences de l'Institut de France 9: 255–332.
3. S. Sudha,R. Ramachandran, Star – Acyclic Coloring of Generalized Petersen; Mathematical Sciences International Research Journal ISSN 2278 – 8697 Vol 3 Issue 1 (2014), Pg 69-71
4. Daniels H.E., 1944. The relation between measures of correlation in the universe of sample permutations. Biometrika 33 (2): 129–135.
5. Fisher R.A., 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 10: 507–521.
6. Fisher R.A., 1921. On the "probable error" of a coefficient of correlation deduced from a small sample. Metron 1:3–32.
7. Galton F., 1869. Hereditary genius. An inquiry into its laws and consequences. MacMillan, London.
8. Galton F., 1875. Statistics by intercomparison. Philosophical Magazine 49: 33–46.
9. Galton F., 1885. Regression towards mediocrity in hereditary stature. Journal of the Anthropological Institute 15: 246–263.
10. Galton F., 1877. Typical laws of heredity. Proceedings of the Royal Institution 8: 282–301.
11. Galton F., 1888. Co-relations and their measurement, chiefly from anthropometric data. Proceedings of the Royal Society of London 45: 135–145.
12. Galton F., 1890. Kinship and correlation. North American Review 150: 419–431.

13. Griffith D.A., 2003. Spatial autocorrelation and spatial filtering. Springer, Berlin.
14. Haining R., 1991. Bivariate correlation with spatial data. Geographical Analysis 23 (3): 210–227.
15. Kendall M.G., 1938. A new measure of rank correlation. Biometrika 30: 81–89.
16. Kendall M.G.,1948. Rank correlation methods. 4th ed. Griffin, London.
17. Pearson K., 1896. Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. Philosophical Transactions of the Royal Society Ser. A 187:253–318.
18. Pearson K., 1900. Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. Philosophical Transactions of the Royal Society Ser. A 195: 1–47.
19. Pearson K., 1920. Notes on the history of correlation. Biometrika 13: 25–45. Piovani J.I., 2008. The historical construction of correlation as a conceptual and operative instrument for empirical research. Quality & Quantity 42: 757–777.
20. Plata S., 2006. A note on Fisher's correlation coefficient. Applied Mathematical Letters 19: 499–502.
21. Rodgers J.L. & Nicewander W.A., 1988. Thirteen ways to look at the correlation coefficient. The American Statistician 42 (1): 59–66.
22. Spearman C.E, 1904a. The proof and measurement of association between two things. American Journal of Psychology 15: 72–101.
23. J. Krishnaveni, J. Subhashini, C. Sekar, Some Separation Properties of Čech Soft Closure Spaces; Mathematical Sciences international Research Journal ISSN 2278 – 8697 Vol 3 Issue 2 (2014), Pg 698-700
24. Spearman C.E, 1904b. General intelligence, objectively determined and measured. American Journal of Psychology 15: 201–293.
25. Spearman C.E., 1910. Correlation calculated from faulty data. British Journal of Psychology 3: 271–295.
26. Stigler S.M., 1988. Francis Galton's account of the invention of correlation. Statistical Science 4 (2): 73–86.
27. Student, 1908. Probable error of a correlation coefficient. Biometrika 6: 302–310.
28. Valz P.D. & Thompson M.E., 1994. Exact inference for Kendall's S and Spearman's rho. Journal of Computational and Graphical Statistics 3: 459–472. Walker H. M., 1928. The relation of Plana and Bravais to theory of correlation. Isis 10 (2): 466–484.
29. Weida F.M., 1927. On various conceptions of correlation. The Annals of Mathematics, Second Series 29 (1/4): 276–312.
30. Yule G.U., 1897a. On the significance of Bravais' formulae for regression, in the case of skew correlation. Proceedings of the Royal Society of London Ser. A 60: 477–489
31. Yule G.U., 1897b. On the theory of correlation. Journal of the Royal Statistical Society Ser. A 60: 812–854.
32. *Dr. P.Puthiyanayagam, K.Saravanan,* New A-D- and E-Optimal N-Ary Block Designs With Unequal; Mathematical Sciences International Research Journal ISSN 2278 – 8697 Vol 2 Issue 1 (2013), Pg 69-73
33. Yule G.U., 1903. Notes on the theory of association of attributes in statistics. Biometrika 2: 121–134.
34. Yule G.U., 1907. On the theory of correlation for any number of variables, treated by a new system of notation. Proceedings of the Royal Society of London 79: 182–193.

Rupinderjeet Kaur (Corresponding Author)
Khalsa College, Patiala
E-Mail: Rupinderjeet_Batth@Yahoo.Com