

Survey on Data Analysis Using Correlation Methods

Katru Akhila¹, Chilamakuru Ravalipriya², Bonda Himaja³, Deepa N⁴

¹School of information technology and engineering, Vit university, Vellore, Tamilnadu, India

²School of information technology and engineering, Vit university, Vellore, Tamilnadu, India

³School of information technology and engineering, Vit university, Vellore, Tamilnadu, India

⁴Asst.prof, school of information technology and engineering, Vit university, Vellore, Tamilnadu, India

Abstract - In this paper, we are going to use the correlation methods for analysis of data. The correlation methods are Pearson and Spearman methods. These are mainly used in the software process to identify the complexity between the various modules in the software. For calculating the correlation between the modules we need to collect the data. For obtaining the data we use the code analyzer tool. It is an open source tool in which we able to analyze the metrics of an application at a particular time. The main advantage of this tool is its support the various languages.

Key Words: Pearson, spearman, correlation methods, coefficient, linear correlation, complexity

1. INTRODUCTION

Using the code analyzer tool we can able to calculate the total number of lines, number of code lines, whitespace lines and comment lines. In this paper we take the total lines and code lines as our data and calculate the correlation between them. Correlation is a measure of the extent to which two variables are related. If an increase in one variable depends on the increase of the other variable then it is positive correlation. Correlation helps to identify the complexity between the modules. The major requirement for the good software is to maintain the complexity between the various modules being low. If the complexity of the system is more than the fault occurrence in the system also increases and it is difficult for system building.

2. PEARSON METHOD

Correlation between the data is a measure of how well they are related. The most common method of correlation is Pearson Correlation. Pearson correlation coefficient(r) is a measure of the strength and the linear relationship between the two sets of data. It is used when there are two quantitative variables. The Pearson coefficient value lies between the -1 and +1. It has the positive linear correlation, negative linear correlation and no linear correlation.

For calculating the Pearson correlation we need to follow the below steps:

- i. Make a table with the obtained data for two columns (x, y) and add three columns for (xy), (x²) and (y²).
- ii. Complete the table using manual calculation.
- iii. After completing the table we need to add all the columns from top to bottom
- iv. Use this formula to find Pearson correlation coefficient

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

Where,

n =number of pairs of scores

$\sum xy$ =sum of the products of paired scores

$\sum x$ =sum of x scores

$\sum y$ =sum of y scores

$\sum x^2$ =sum of squared x scores

$\sum y^2$ =sum of squared y scores

Table -1: Pearson Correlation Results

X	Y	XY	X ²	Y ²
60	53	3180	3600	2809
98	78	7644	9604	6084
128	107	13696	16384	11449
68	58	3944	4624	3364
64	48	3072	4096	2304
61	48	2928	3721	2304
65	48	3120	4225	2304
61	54	3294	3721	2916
146	125	18250	21316	15625
62	51	3162	3844	2601

n =10,

$\sum x = 813, \sum y = 670,$

$\sum xy = 62290,$

$\sum x^2 = 75135, \sum y^2 = 5170$

By using the above formula, we can obtain the Pearson correlation coefficient (r) as

$$r = 0.99228$$

3. SPEARMAN METHOD

Spearman correlation coefficient (ρ) is based on the ranked values for each variable for a given data set. It is a technique which can be used to measure the strength and relationship between two variables. It is also measure of the strength of an association between two variables. In spearman the results always lies between the +1 and -1. It calculates the monotonic relationship between two variables and the variables tend to change together but not at a constant rate.

For calculating the Spearman correlation coefficient we need to follow the below steps:

- i. Create a table from the given data.

- ii. Rank the given two data sets based on ascending order.
- iii. If we have tied scores we have to give the same rank for all those scores.
- iv. Find the difference between the ranks of two data sets.
- v. Square the difference of the ranks to remove the negative values and sum them.

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

Where,

$\sum d_i^2 =$ sum of squares of difference of ranks

n =number of scores

Table -2: Spearman Correlation Results

X	Rank of X	Y	Rank of Y	d	d ²
60	10	53	6	4	16
98	3	78	3	0	0
128	2	107	2	0	0
68	4	58	4	0	0
64	6	48	8	-2	4
61	8	48	8	0	0
65	5	48	8	-3	9
61	8	54	5	3	9
146	1	125	1	0	0
62	7	51	7	0	0

$$\sum d^2 = 38$$

n =10

By using the above formula, we can obtain the spearman rank coefficient (ρ) as

$\rho = 0.77$

4. COMPARISON BETWEEN PEARSON AND SPEARMAN METHOD

Pearson correlation measures the linear relationship between two continuous variables and the Spearman correlation measures the data sets based on their ranks for each variable. Spearman does not require assumption of the relationship between variables but it is always good to examine the relationship between variables. The main difference between them is that Pearson is most appropriate for measurements from an interval scale and the Spearman correlation is appropriate for measurements taken from ordinal scales.

5. CONCLUSION

By using the above methods such as Pearson correlation and spearman correlation, we can conclude that the Pearson coefficient has higher correlation than the Spearman coefficient. This is because the Pearson coefficient value is greater than Spearman coefficient value.

REFERENCES

- [1] Myers, Leann, and Maria J. Sirois. "Spearman Correlation Coefficients, Differences between." Encyclopedia of Statistical Sciences
- [2] Cornbleet, P. Joanne, and Margaret C. Shea. "Comparison of product moment and rank correlation coefficients in the assessment of laboratory method-comparison data." *Clinical chemistry* 24.6 (1978): 857-861.
- [3] Bookbinder, Michael J., and Kim J. Panosian. "Using the coefficient of correlation in method-comparison studies." *Clinical chemistry* 33.7 (1987): 1170-1176.
- [4] Tuğran, Elif, et al. "A Simulation Based Comparison of Correlation Coefficients with Regard to Type I Error Rate and Power." *Journal of Data Analysis and Information Processing* 3.03 (2015): 87.
- [5] Morrison, Julie L., et al. "GeneRank: using search engine technology for the analysis of microarray experiments." *BMC bioinformatics* 6.1 (2005): 233.
- [6] Matzke, Dora, et al. "Bayesian Inference for Correlations in the Presence of Measurement Error and Estimation Uncertainty."
- [7] Benesty, Jacob, et al. "Pearson correlation coefficient." *Noise reduction in speech processing*. Springer Berlin Heidelberg, 2009. 1-4.
- [8] Wiedermann, Wolfgang, and Michael Hagmann. "Asymmetric properties of the Pearson correlation coefficient: Correlation as the negative association between linear regression residuals." *Communications in Statistics-Theory and Methods* 45.21 (2016): 6263-6283.
- [9] Xiong, Hui, et al. "Exploiting a support-based upper bound of Pearson's correlation coefficient for efficiently identifying strongly correlated pairs." *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004.
- [10] Son, Young Sook, and Jangsun Baek. "A modified correlation coefficient based similarity measure for clustering time-course gene expression data." *Pattern Recognition Letters* 29.3 (2008): 232-242.
- [11] Emerson, Robert Wall. "Causation and Pearson's Correlation Coefficient." *Journal of Visual Impairment & Blindness (Online)* 109.3 (2015): 242.
- [12] McDonald, John H. *Handbook of biological statistics*. Vol. 2. Baltimore, MD: Sparky House Publishing, 2009.
- [13] Kruschke, John. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, 2014.