Parameter: Journal of Statistics
Available at https://bestjournal.untad.ac.id/index.php/parameter
Vol. 5 No. 1 2025, 43 - 49
DOI: *https://doi.org/10.22487/27765660.2025.v5.i1.17444*
eISSN: 2776-5660

ROBUST PERMUTATION TEST FOR SPEARMAN CORRELATION AND ITS APPLICATION TO TESTING THE RELATIONSHIP BETWEEN OPEN UNEMPLOYMENT RATE AND NUMBER OF CRIMES

Indriyani[1] dan Suliadi Suliadi[2*]

[1,2]Dept. of Statistics, Bandung Islamic University

[2]ORCHID ID: https://orcid.org/0000-0002-1201-1044

***e-mail:** suliadi@gmail.com

ABSTRACT

The Pearson correlation coefficient can be inaccurate for data that doesn't follow a normal distribution or has outliers, which makes it difficult to measure how strong the linear relationship is. Instead, the Spearman correlation coefficient can be used to measure relationships that go in the same direction without needing a normal distribution and can handle outliers. While the t-test method for Spearman correlation is often taught, it isn't always suitable and can lead to a Type I error when data isn't normal or the sample size is small. A robust permutation test method on the Spearman correlation coefficient using studentized statistics is designed to overcome these limitations. This research discusses the application of that method to analyze data about the open unemployment rate and the number of reported crimes in Indonesia. Using the permutation test, the value of r_s=0.2437 was obtained with a p-value <0.05, indicating a significant correlation between the two variables. The findings of this study are expected to provide a foundation for effective policy recommendations that consider unemployment factors in reducing crime rates.

Keywords: *Crime, Non-normality, Permutation Test, Robust, Rank Spearman Correlation, Unemployment*

Cite: Indriyani, I., & Suliadi, S. (2025). Robust permutation test for Spearman correlation and its application to testing the relationship between open unemployment rate and number of crimes. *Parameter: Journal of Statistics*, *5*(1), 43–49. https://doi.org/10.22487/27765660.2025.v5.i1.17444

*Received: May 2025; Reviewed: May 2025; Published: Jul 2025

INTRODUCTION

Correlation analysis is an important tool for measuring the strength and direction of relationships between variables in research. Pearson correlation is usually used to measure linear relationships between variables, but it requires the assumption of normal distribution, which is often not met in real practice. In addition, Pearson correlation is also very sensitive to outliers that can significantly affect the results of the magnitude of the correlation coefficient (Conover, 1999). In situations where data do not meet the assumption of normality or contain outliers, the Spearman correlation is a better choice. It does not require the assumption of bivariate normality and is often considered more robust for non-normal data (Myers & Well, 2003).

A common approach in testing the significance of the Spearman correlation coefficient is t-test. However, Yu & Hatson (2022) obtained that this approach is theoretically incorrect when data deviate from normality or when sample sizes are small (e.g., $n \leq 50$), specifically it cannot control the type I error. The coefficient of rank Spearman correlation is robust to normality deviation and outliers, but not for testing hypotheses by using t-test. To overcome this shortcoming, Yu & Hatson (2022) proposed using a permutation test as a more robust method. The design of this test aims to yield more accurate results and enhance Type I error control compared to conventional methods.

This approach is particularly relevant for social research where data is often non-normal and has outliers, such as in studies on unemployment and crime (Elamir & Mousa, 2021). The Open Unemployment Rate (OER) is an important indicator in labor market analysis that reflects the proportion of the labor force that is actively seeking work but has not yet found a job (World Bank, 2021). High unemployment can negatively affect individual well-being, mental health, and lead to social and political instability, as well as contribute to an increase in crime. Crime, as a complex phenomenon, involves unlawful acts that adversely affect society, and is often influenced by socio-economic factors such as poverty and unemployment (Sampson & Laub, 1993). Economic deprivation and limited opportunities can increase the risk of involvement in criminal activity.

Based on the description above, a problem can be identified, namely how the relationship between the open unemployment rate and reported crime in Indonesia based on the results of the analysis using the Spearman coefficient correlation permutation test. The purpose of this article is to test the strength of the relationship through Spearman's coefficient correlation between the open unemployment rate and the number of crimes reported in Indonesia, using a permutation test approach. By using a more robust permutation test, the results can more accurately evaluate the relationship between these social variables and provide a stronger basis for formulating effective policies.

MATERIALS AND METHODS

Variabels and Data Sources

The variables used are the open unemployment rate and the number of crimes reported in Indonesia in 2021. Data was obtained from BPS-Statistics Indonesia publications for each province, covering a total of 34 provinces. The open unemployment rate is defined as the percentage of the labor force that is unemployed and looking for work, while the number of reported crimes refers to the total number of criminal incidents officially recorded by law enforcement authorities. Both variables are measured at the provincial level, with the open unemployment rate expressed as a percentage and the number of reported crimes expressed as a count (number of cases).

Analysis Methods

Normality Test

Suppose there are n vector of observations, $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ where $\boldsymbol{x}_i = (x_{i1}, x_{i2})^t, i = 1, 2 \ldots, n$. Before applying the correlation analysis, a normality test was conducted using the Mardia test to determine if the data followed a multivariate normal distribution. This test involves two main components: skewness and kurtosis. Skewness ($b_{1,2}$) assesses the degree asymmetry of the distribution, while kurtosis ($b_{2,2}$) measures the skewness of the distribution. The Mardia test is a generalization of the univariate skewness and kurtosis tests (Tabachnick & Fidell, 2013). For bivariate data (p= 2), skewness and kurtosis are calculated using the formula:

$$b_{1,2} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ (\boldsymbol{X}_i - \bar{\boldsymbol{X}})^t \boldsymbol{S^{-1}} (\boldsymbol{X}_j - \bar{\boldsymbol{X}}) \right\}^3 \quad (1)$$

$$b_{2,2} = \frac{1}{n^2}\sum_{i=1}^{n}\{(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^t \boldsymbol{S}^{-1}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})\}^2 \quad (2)$$

where n is the number of observations, $\boldsymbol{X}_i$is the i-th observation vector, $\overline{\boldsymbol{X}}$ is the sample mean, and S is the covariance matrix.

The p-value for skewness is calculated using the chi-square distribution with degrees of freedom $\frac{m(m+1)(m+2)}{6}$, for m=2:

$$x^2_{skewness} = \frac{n}{6} b_{1,2} \quad (3)$$

$$p - value_{skewness} = x^2_{(\alpha,df)} > x^2_{skewness} \quad (4)$$

The p-value for kurtosis is calculated using the normal distribution (z):

$$z_{kurtosis} = \frac{b_{2,2} - m(m+2)}{\sqrt{\frac{8m(m+2)}{n}}} \quad (5)$$

$$p - value_{kurtosis} = 2(1 - \Phi(|z_{kurtosis}|)) \quad (6)$$

The hypotheses for the Mardia Skewness/Kurtosis are:

H_0 : Data comes from multivariate normal distribution

H_1 : Data does not come from multivariate normal distribution

Test criteria: H_0 is rejected if the p-value of Mardia Skewness/Kurtosis < α (significance level).

Outlier Check

Mahalanobis Distance measures the distance of a data point from the distribution by considering the covariance between variables, making the data standardized and uncorrelated (Rencher, 2002). This method is often applied in multivariate outlier detection and can provide reliable identification of unusual observations (Filzmoser et al., 2005). For bivariate data, the Mahalanobis Distance is calculated by:

$$D_i(\boldsymbol{X_i}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sqrt{(\boldsymbol{X_i} - \mu)^t \Sigma^{-1}(X_i - \mu)} \quad (7)$$

where μ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix. The i-th data point is considered an outlier if

$$D_i(\boldsymbol{X_i}, \overline{\boldsymbol{x}}, \boldsymbol{S}) > \chi^2_{(2,\alpha)} \quad (8)$$

The hypotheses for the Mahalanobis Distance are:

H_0 : The observation is not an outlier

H_1 : The observation is an outlier

Test criteria: H_0 is rejected if $D_i > \chi^2_{(2,\alpha)}$. with 2 being the degrees of freedom and α being the level of significance.

Spearman Correlation Coefficient

Spearman correlation is a non-parametric statistical method used to measure the strength and direction of a monotonic relationship between two variables (Pallant, 2016). Unlike the Pearson correlation which requires the assumption of normality, the Spearman correlation calculates the relationship based on the rank of the data, so it is more resistant to outliers and does not require the assumption of a normal distribution. The Spearman correlation coefficient (r_s) is calculated by converting the variable values into ranks (Field, 2013), which is written in equation (9):

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(a_i - b_i)^2}{n(n^2-1)} \quad (9)$$

where a_i and b_i are the rank of the $x_{i1}$and x_{i2} values, respectively, among n observations of its variable.

Spearman Correlation Coefficient Permutation Test

Yu & Hatson (2022) proposed a new approach in testing the significance test for rank Spearman correlation. This test is robust to the normality violation and appropriate for small sample size. It can also control the type I error. The proposed test is based on a permutation test. The test is a non-parametric method used to assess the statistical significance of test without assuming a particular distribution. It works by creating an empirical distribution of the test statistic through repeated randomization of the data (Good, 2005). The steps to perform the permutation test for Spearman correlation are as follows:

1. Convert *X* and *Y* data into ranks (a_i and b_i). If there is the same data, use the average rank.
2. Use equation (9) to calculate the Spearman correlation coefficient.
3. Calculate the variance estimate of the Spearman correlation coefficient

$$\hat{\tau}^2 = \frac{\hat{\mu}_{22}}{\hat{\mu}_{20}\hat{\mu}_{02}} \tag{10}$$

$$\hat{\mu}_{22} = \frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{a})^2(b_i - \bar{b})^2 \tag{11}$$

$$\hat{\mu}_{20} = \frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{a})^2 \tag{12}$$

$$\hat{\mu}_{02} = \frac{1}{n}\sum_{i=1}^{n}(b_i - \bar{b})^2 \tag{13}$$

4. Calculate the studentized statistic (R_s) for the original data using the formula:

$$R_s = \frac{r_s}{\hat{\tau}} \tag{14}$$

Where r_s is the Spearman rank correlation coefficient.

5. Randomize the rank of variable *Y*, while the rank of variable *X* remains fixed *B* times (*B* = *n*!), or do a random permutation if *B* is greater than 1 million.
6. For each permutation, calculate the studentized statistic (R_s^k) using the formula:

$$R_s^k = \frac{r_s^{(k)}}{\hat{\tau}^{(k)}} \tag{15}$$

and variance estimation ($\hat{\tau}^{2(k)}$) with the formula:

$$\hat{\tau}^{2(k)} = \frac{\hat{\mu}_{22}^k}{\hat{\mu}_{20}\hat{\mu}_{02}} \tag{16}$$

7. Calculate the p-value or ASL (Achieved Significance Level*)* by using:

$$ASL = p = 2.min\left(\frac{1}{B}\sum_{k=1}^{B} I\left(R_s^k < R_s\right), \frac{1}{B}\sum_{k=1}^{B} I\left(R_s^k > R_s\right)\right) \tag{17}$$

The hypotheses for the Spearman Correlation Coefficient Permutation Test are:

H_0 : There is no monotonic correlation between variables X and Y ($p = 0$).

H_1 : There is a monotonic correlation between variables X and Y ($p \neq 0$).

Test criteria: H_0 is rejected if the p-value (ASL) < α (significance level).

RESULTS AND DISCUSSION

Descriptive statistics of the research variables are depicted in Table 1.

Table 1. Descriptive Statistics of Research Variables

Variables	N	Minimum	Maximum	Average	Standard Deviation
Number of Crimes	34	971	36543	7044	7702.057
Open Unemployment Rate	34	3.01	9.91	5.492	1.819

The average number of crimes was 7044, with a range of 971 to 36543, with a standard deviation of 7702.057. This shows a large variation in the number of crimes between provinces. The average TPT was 5.492%, with a range of 3.01% to 9.91% and a standard deviation of 1.819, showing quite clear differences between provinces.

Normality Test

Testing the normality of the data on the number of crimes and the open unemployment rate using the Mardia test. The results are shown in Table 2.

Table 2. Mardia test results for data on the number of crimes and open unemployment rate

Test	Statistic	P-value	Result
Mardia Skewness	42.8795	1.096×10^{-8}	Not Normal
Mardia Kurtosis	4.4022	1.071×10^{-5}	Not Normal

Test Criteria

p-value Mardia Skewness = 1.096×10^{-8}. If it used α =0.05 then p-value Mardia Skewness < α. It's mean that H_0 is rejected.

p-value Mardia Kurtosis = 1.071×10^{-5}. If it used α =0.05 then p-value Mardia Kurtosis < α. It's mean that H_0 is rejected.

From the results of the Mardia test that has been carried out, with a real level of 5%, it can be concluded that the data does not come from a multivariate normal distribution. Because the data does not meet the assumption of normality, the analysis can continue with the rank Spearman method without the need to check for outliers first.

Spearman Correlation Coefficient

The data for both variables were converted into ranks, and then the Spearman correlation coefficient was calculated, as follows:

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(a_i - b_i)^2}{n(n^2-1)} = 1 - \frac{6(4949{,}5)}{34(34^2-1)} = 0.2437$$

The Spearman correlation coefficient of the original data shows a value of r_s = 0.2437, indicating a weak positive relationship between the number of reported crimes and the open unemployment rate in the 34 provinces of Indonesia.

Spearman Correlation Coefficient Permutation Test

The process starts by calculating the variance estimates and studentized statistics for the original data. The calculation results show:

$$\hat{\tau}^2 = \frac{\hat{\mu}_{22}}{\hat{\mu}_{20}\hat{\mu}_{02}} = \frac{257.809}{(3272.5)(3.272)} = 0.024$$

$$R_s = \frac{r_s}{\hat{\tau}} = \frac{0.2437}{0.1552} = 1.57$$

The next step involves randomizing the open unemployment rate variable while the rank of the number of crimes remains unchanged. Given the very large number of permutation combinations (34! about 2.952×10^{38}) for 34 data pairs, this study uses random permutation with one million permutations for efficiency.

For example, in the first permutation, a new ranking of the open unemployment rate is calculated, followed by a recalculation of the Spearman correlation coefficient with the values obtained being:

$$r_s^{(1)} = 1 - \frac{6\sum_{i=1}^{n}\left(a_i - b_{(1)i}\right)^2}{n(n^2-1)} = 1 - \frac{6(7374{,}5)}{34(34^2-1)} = -0.1628$$

It is followed by the calculation of variance estimation and studentized statistics for the first permutation, the values obtained are:

$$\hat{\tau}^{2(1)} = \frac{\hat{\mu}_{22}^{(1)}}{\hat{\mu}_{20}\hat{\mu}_{02}} = \frac{289.029}{(3272.5)(96.23529)} = 0.9178$$

$$R_s^{(1)} = \frac{r_s^{(1)}}{\hat{\tau}^{(1)}} = \frac{-0.1628}{0.958} = -0.1324$$

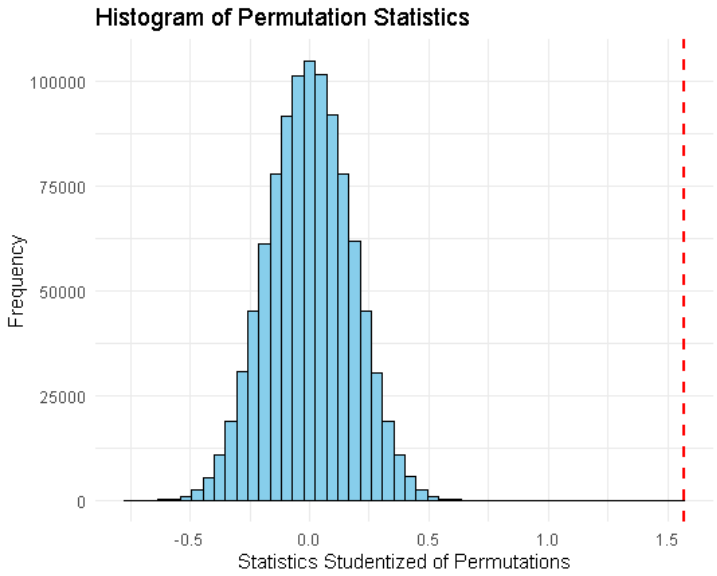

Figure 1. Histogram of Permutation Results

After 1.000.000 runs, it was obtained $R_s^{(1)}, R_s^{(2)}, \ldots, R_s^{(1.000.000)}$ and depicted in Figure 1. The histogram shows that the R_s values of the original data (indicated by the red solid line) is at the extreme right compared to the distribution of the R_s^k values of the permutations, in which all the R_s^k values of permutations being smaller than the $R_s$values of the original data. This indicates that the correlation between the number of reported crimes and the open unemployment rate is significant.

ASL value or P-value is calculated with the formula:

$$ASL = p = 2.\min\left(\frac{1}{1.000.000}\sum_{k=1}^{1.000.000} I\left(R_s^k < R_s\right), \frac{1}{1.000.000}\sum_{k=1}^{1.000.000} I\left(R_s^k > R_s\right)\right)$$
$$= 2.\min(1,0) = 0$$

With the result of ASL = 0, which is much smaller than the significance level of α=0.05, we can conclude that the null hypothesis (H_0) is rejected, indicating a significant correlation between the number of crimes and the unemployment rate. Although the correlation is significant, the value of r_s =0.2437 indicates that the strength of the relationship between the two variables is quite weak.

CONCLUSION

The analysis indicates that the Spearman correlation coefficient of the original data is r_s=0.2437, which means there is a weak positive relationship between the number of reported crimes and the open unemployment rate in 34 provinces in Indonesia. In the permutation test, the studentized statistic was R_s = 1.57. After performing one million permutations, the p-value was found to be 0, which is much smaller than the significance level α = 0.05. Therefore, we reject the null hypothesis, demonstrating a significant correlation between the number of reported crimes and the open unemployment rate.

REFERENCES

Badan Pusat Statistik (BPS) (2024, July 24). *Tenaga Kerja*. Retrieved from https://mahulukab.bps.go.id/subject/6/tenaga-kerja.html.

Conover, W. J. (1999). *Practical Nonparametric Statistics* (3rd ed.). John Wiley & Sons.

Elamir, E. A. H., & Mousa, A. E. (2021). *Robust statistical methods for skewed and heavy-tailed data in social sciences*. International Journal of Statistics and Data Science, 2(1), 12–23.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (5th ed.). SAGE Publications Ltd.

Filzmoser, P., Varmuza, K., & Reimann, C. (2009). *Introduction to multivariate analysis: From data to models*. Springer.

Good, P. I. (2005). *Permutation, Parametric and Bootstrap Tests of Hypotheses* (3rd ed.). Springer.

Myers, J. L., & Well, A. D. (2003). *Research Design and Statistical Analysis* (2nd ed.). Lawrence Erlbaum Associates.

Pallant, J. (2016). *SPSS Survival Manual: A Step by Step Guide to Data Analysis Using IBM SPSS* (6th ed.). Open University Press.

Rencher, A. C. (2002). *Methods of Multivariate Analysis* (2nd ed.). Wiley-Interscience.

Sampson, R. J., & Laub, J. H. (1993). *Crime and Deviance in the Life Course*. University of Chicago Press.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics* (6th ed.). Pearson.

World Bank. (2021). *Unemployment, total (% of total labor force) (modeled ILO estimate)*. World Bank. Retrieved from https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS

Yu, H., & Hutson, A. D. (2022). *A robust Spearman correlation coefficient permutation test*. Communications in Statistics - Theory and Methods, 53(6), 2141–2153.