An Interval Estimation of Pearson's Correlation Coefficient by Bootstrap Methods

Bumrungsak Phuenaree* and Sirikun Sanorsap

Department of Mathematics, Faculty of Science,
Burapha University, Chonburi, Thailand.

**Corresponding author's email: bumrungsak [AT] buu.ac.th*

ABSTRACT— ***In this paper, we compare three confidence intervals for Pearson's correlation coefficient which are Fisher's transformation, standard bootstrap and percentile bootstrap methods. The performance of these confidence intervals is considered by the coverage probability and the average width. Monte Carlo simulation results for generating non-normal distribution show that the percentile bootstrap confidence interval is the best method, when the distribution is a uniform distribution and the sample sizes are larger than or equal to 50. For the logistic and Laplace distributions, the percentile bootstrap method is the most efficiency method when the sample sizes are larger than or equal to 200 and the correlation coefficients are at least 0.5. However, the Fisher method gives the best confidence interval when the correlation coefficients are 0.2.***

Keywords— Confidence Interval, Pearson's correlation, Bootstrap method, Fisher's transformation.

1. INTRODUCTION

The study of the relationship between two quantitative variables commonly uses the correlation coefficient to determine strength and direction of the correlation between them, such as the temperature and the relative humidity. The degree of the relationship can be somewhere between -1 and +1. If it is close to ±1, the two variables have strong correlation. Considering the direction when the coefficient is positive, the variables are directly correlated. But if the coefficient is negative, the two variables are inversely correlated.

Pearson's correlation coefficient is the most widely used. It is statistical measure of linear relationship between two variables. In the statistical inference, the confidence interval for Pearson's correlation coefficient can be computed from the Fisher's transformation (Fisher, 1934). However, there are the assumptions about the estimating Pearson's correlation coefficient, such as the normality assumption. But in practice, it is very difficult to know the distribution of the data. For this reason, the Fisher's transformation may not be a suitable method in this situation.

Efron and Tibshirani (1994) proposed the bootstrap method to estimate the standard error of some interest statistic for unknown distributed data, and used this method to calculate the confidence interval for unknown parameter.

The object of this study is to compute the confidence interval for Pearson's correlation coefficient by using Fisher transformation and bootstrap methods when the data is non-normal. In this research we shall consider only the symmetric distributions.

2. PEARSON'S CORRELATION COEFFICIENT

Pearson's correlation coefficient is a statistical measure of linear relationship between two variables, X and Y. This value is represented by ρ, referred to the population correlation coefficient, and by r, referred to the sample correlation coefficient. Suppose $\{(x_1, y_1),(x_2, y_2),...,(x_n, y_n)\}$ is a set of pairwise random samples of size n. The formula for r can be obtained by:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}. \quad (1)$$

The Pearson's correlation coefficient has been being used in many fields, such as the medical research (Mukaka, 2012), geometric growth model for small-world networks (Shang, 2014) etc.

3. FISHER'S TRANSFORMATION

The Fisher's transformation is a method to transform the sampling distribution of Pearson's correlation coefficient so that it becomes normally distributed. Define the Fisher's transformation (Fisher, 1934) of r (z_r) as follows:

$$z_r = 0.5 \ln \frac{1+r}{1-r}. \quad (2)$$

If two random variables, X and Y, in section 2 have a joint bivariate normal distribution with the correlation ρ, then z_r is approximately normally distributed. The standard error of z_r is equal to $\frac{1}{\sqrt{n-3}}$. So, the confidence interval of Pearson's correlation can be constructed as:

$$\left(z_r - z_{\alpha/2}\frac{1}{\sqrt{n-3}}, z_r + z_{\alpha/2}\frac{1}{\sqrt{n-3}}\right), \quad (3)$$

where z_r is the Fisher's transformation of r, $z_{\alpha/2}$ is the $\frac{100\alpha}{2}$ quantile of a standard normal distribution and n is the sample size. Then the lower and upper limits of the confidence interval of r can be shown as follows:

$$r = \frac{\exp(2z_r) - 1}{\exp(2z_r) + 1} \quad (4)$$

4. BOOTSTRAP METHODS

The bootstrap method relies on a simple idea without knowing about the underlying distributions of our observations. The random samples are taken with replacement from the original data, and then the estimate of interested parameter is computed. Resampling the samples many time in order to obtain the distribution of bootstrap estimates (Efron, 1994), so this estimate distribution can be used to calculate the confidence interval of Pearson's correlation coefficient.

4.1 Standard Bootstrap

The bootstrap samples $(x_1^*, y_1^*), (x_2^*, y_2^*), ..., (x_n^*, y_n^*)$ are drawn with replacement from the random samples of size n orders $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. For each of these bootstrap samples we calculate the Pearson's correlation coefficients, $r_1^*, r_2^*, ..., r_B^*$, where B is the number of bootstrap replications. The standard bootstrap interval can be written as:

$$\left(r_B - z_{\alpha/2} SE(r_B), r_B + z_{\alpha/2} SE(r_B)\right), \quad (5)$$

where

$$r_B = \frac{1}{B}\sum_{i=1}^{B} r_i^*,$$

$$SE(r_B) = \sqrt{\frac{\sum_{i=1}^{B}\left(r_i^* - r_B\right)^2}{B-1}},$$

and $z_{\alpha/2}$ is the $\frac{100\alpha}{2}$ quantile of a standard normal distribution.

4.2 Percentile Bootstrap

Based on *B* bootstrapped values $r_1^*, r_2^*, ..., r_B^*$ from the previous section, the bootstrap percentile interval of ρ can be defined by

$$\left(r_{(B\alpha)}^*, r_{(B(1-\alpha))}^*\right), \tag{6}$$

where $r_{(B\alpha)}^*$ is the $B\alpha$ th order statistics of the bootstrap distribution r^*, $r_{(B(1-\alpha))}^*$ is the $B(1-\alpha)$ th order statistics of the bootstrap distribution of r^* and *B* is the number of bootstrap replications.

5. COVERAGE PROBABILITY

The confidence coefficient is the percentage of the confidence intervals that contain the true value of interested parameter. This value is set by researcher to calculate the confidence interval. The coverage probability of a confidence interval is the proportion of the time that the true parameter value is in the interval. Both values are equivalent by the definition; in other word, the coverage probability should be close to the nominal confidence coefficient.

The criteria of coverage probability is based on hypothesis testing of the proportion at 0.05 nominal level considered in this case. If the coverage probability is in the following interval:

$$\left(p_0 - z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{N}}, p_0 + z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{N}}\right), \tag{7}$$

where p_0 is the nominal confidence coefficient and N is the number of Monte Carlo simulations, that means it is close to the nominal confidence coefficient.

6. SIMULATION RESULTS

This section provides simulation case studies for the coverage probabilities and the average widths of the confidence intervals. The 5,000 random samples of sizes *n* = 10, 30, 50, 100, 200 and 400 are generated from three symmetric distributions; Laplace, Uniform and logistic distributions, with the same variance which are Laplace(0,1.28189), Uniform(0,6.2832) and Logistic(0,1). The confidence coefficient is 0.95. The 1,000 bootstrap samples are drawn from the original sample. Three Pearson's correlation coefficients of 0.2, 0.5 and 0.95 are considered. The results for coverage probability (CP) and average width (AW) are shown in Table 1-3.

Table 1 illustrates the coverage probabilities and the average widths for the uniform distribution. Two bootstrap confidence intervals have coverage probabilities closed to the nominal confidence coefficient when the sample sizes are 100, 200 and 400. Besides a coverage probability, the percentile bootstraps interval has the shortest average width but this following cases; for large correlation coefficient (ρ = 0.95), the average widths of two bootstrap methods are the same. However, for *n* = 10, 30, 50, and small correlation coefficient; ρ = 0.2, the Fisher method gives the best results.

In Table 2 and Table 3 show the results of all confidence intervals in logistic and Laplace distribution respectively. It can be seen that two bootstrap methods have the coverage probabilities closed to the nominal confidence coefficient when the sample sizes are large; *n* = 200, 400, and ρ = 0.5, 0.95. Apart from a coverage probability, the percentile bootstrap interval has the shortest average widths. However, for small correlation coefficient; ρ = 0.2, the Fisher method gives the best results.

Table 1: The coverage probabilities and the average widths of 95% confidence intervals for Uniform(0,6.2832)

n	ρ	Fisher		Percentile Bootstrap		Standard Bootstrap	
		CP	AW	CP	AW	CP	AW
10	0.2	0.9540*	1.1390	0.9032	-	0.8668	-
	0.5	0.9566	-	0.9108	-	0.8888	-
	0.95	0.9858	-	0.9314	-	0.9274	-
30	0.2	0.9532*	0.6762	0.9380	-	0.9196	-
	0.5	0.9642	-	0.9336	-	0.9238	-
	0.95	0.9940	-	0.9436	-	0.9398	-
50	0.2	0.9554*	0.5275	0.9436	-	0.9364	-
	0.5	0.9668	-	0.9458*	0.3910	0.9412	-
	0.95	0.9946	-	0.9476*	0.0396	0.9418	-
100	0.2	0.9558*	0.3745	0.9496*	0.3697	0.9496*	0.3710
	0.5	0.9650	-	0.9494*	0.2731	0.9450*	0.2736
	0.95	0.9962	-	0.9560*	0.0269	0.9552*	0.0269
200	0.2	0.9510*	0.2653	0.9458*	0.2618	0.9450*	0.2625
	0.5	0.9704	-	0.9540*	0.1928	0.9514*	0.1931
	0.95	0.9956	-	0.9538*	0.0187	0.9530*	0.0187
400	0.2	0.9558*	0.1879	0.9504*	0.1854	0.9500*	0.1857
	0.5	0.9678	-	0.9486*	0.1360	0.9490*	0.1362
	0.95	0.9972	-	0.9554*	0.0131	0.9546*	0.0131

*The coverage probability is in the interval (0.9440, 0.9560), so the average width would be considered in this case.

Table 2: The coverage probabilities and the average widths of 95% confidence intervals for Logistic(0,1)

n	ρ	Fisher		Percentile Bootstrap		Standard Bootstrap	
		CP	AW	CP	AW	CP	AW
10	0.2	0.9458*	1.1133	0.9036	-	0.8626	-
	0.5	0.9414	-	0.9036	-	0.8626	-
	0.95	0.9192	-	0.8986	-	0.8930	-
30	0.2	0.9462*	0.6738	0.9232	-	0.9060	-
	0.5	0.9326	-	0.9226	-	0.9056	-
	0.95	0.9022	-	0.9194	-	0.9168	-
50	0.2	0.9458*	0.5259	0.9324	-	0.9244	-
	0.5	0.9356	-	0.9324	-	0.9202	-
	0.95	0.8980	-	0.9324	-	0.9286	-
100	0.2	0.9506*	0.3745	0.9434	-	0.9394	-
	0.5	0.9342	-	0.9420	-	0.9348	-
	0.95	0.8958	-	0.9370	-	0.9378	-
200	0.2	0.9494*	0.2654	0.9472*	0.2655	0.9458*	0.2660
	0.5	0.9346	-	0.9480*	0.2190	0.9460*	0.2197
	0.95	0.8944	-	0.9498*	0.0335	0.9466*	0.0336
400	0.2	0.9484*	0.1878	0.9448*	0.1887	0.9442*	0.1892
	0.5	0.9394	-	0.9494*	0.1563	0.9482*	0.1565
	0.95	0.8894	-	0.9506*	0.0236	0.9496*	0.0237

*The coverage probability is in the interval (0.9440, 0.9560), so the average width would be considered in this case.

Table 3: The coverage probabilities and the average widths of 95% confidence intervals for Laplace(0,1.28189)

n	ρ	Fisher		Percentile Bootstrap		Standard Bootstrap	
		CP	AW	CP	AW	CP	AW
10	0.2	0.9448*	1.1278	0.9166	-	0.8746	-
	0.5	0.9160	-	0.9008	-	0.8658	-
	0.95	0.8654	-	0.8896	-	0.8832	-
30	0.2	0.9450*	0.6736	0.9202	-	0.9048	-
	0.5	0.9136	-	0.9214	-	0.9012	-
	0.95	0.8368	-	0.9192	-	0.9102	-
50	0.2	0.9454*	0.5259	0.9272	-	0.9164	-
	0.5	0.9122	-	0.9258	-	0.9140	-
	0.95	0.8150	-	0.9232	-	0.9208	-
100	0.2	0.9444*	0.3741	0.9312	-	0.9268	-
	0.5	0.9080	-	0.9326	-	0.9264	-
	0.95	0.8098	-	0.9330	-	0.9336	-
200	0.2	0.9498*	0.2654	0.9486*	0.2664	0.9488*	0.2674
	0.5	0.9072	-	0.9441*	0.2350	0.9456*	0.2359
	0.95	0.8062	-	0.9442*	0.0410	0.9448*	0.0412
400	0.2	0.9484*	0.1880	0.9480*	0.1908	0.9444*	0.1912
	0.5	0.9130	-	0.9446*	0.1689	0.9440*	0.1693
	0.95	0.8028	-	0.9458*	0.0291	0.9490*	0.0292

*The coverage probability is in the interval (0.9440, 0.9560), so the average width would be considered in this case.

7. CONCLUSION

The problem of non-normality data for estimating the correlation coefficient has considered in this research. When the distribution is a uniform distribution and the sample sizes are larger than or equal to 50, the percentile bootstrap confidence interval is the best method.

For the logistic and Laplace distributions, the percentile bootstrap method is the most efficiency method when the sample sizes are larger than or equal to 200 and the correlation coefficients are at least 0.5. However, the Fisher method gives the best confidence interval when the correlation coefficients are 0.2.

8. REFERENCES

[1] Efron B., and Tibshirani R.J, An introduction to the bootstrap, Chapman & Hall, USA, 1994
[2] Fisher R.A., Statistical methods for research workers, Oliver and Boyd, London, 1934
[3] Mukaka M.M., "A guide to appropriate use of Correlation coefficient in medical research", Malawi Medical Journal, vol. 24, no. 3, pp.69-71, 2012
[4] Shang Y., "Geometric Assortative Growth Model for Small-World Networks", The Scientific World Journal, Article ID759391, vol. 2014, 8 pages, 2014.
[5] Weaver B., and Koopman R., "An SPSS Macro to Compute Confidence Intervals for Pearson's Correlation", The Quantitative Methods for Psychology, vol. 10, no. 1, pp.29-39, 2014