ISSN 2347-1921

Confidence Interval For The Estimation of The Pearson's Correlation Coefficient

Rawiyah Muneer Alraddadi
2055 Napoleon Road, Unit 8 H, Bowling Green, OH 43402 USA

Abstract:

Pearson’s correlation coefficient is used to measure the influence of one quantitative variable on another quantitative variable. Basing on the sign of correlation, the type of dependence (positive/negative) can be decided in bi-variate data. The estimation of parameter can be done in two methods, point estimation and interval estimation. In this paper, various methods to find the confidence interval for the correlation are discussed. As the population correlation coefficient is estimated by the Pearson’s correlation coefficient, the Monte-Carlo simulation will give the approximate the estimation of the Pearson’s correlation coefficient ($\hat{r}$). The Fisher Z, Bootstrap method and variance reduction methods are discussed in this paper.

Keyword: correlation; interval; Monte Carlo; Fisher; Bootstrap

Council for Innovative Research
Peer Review Research Publishing System

Journal: JOURNAL OF ADVANCES IN MATHEMATICS

Vol .11, No.1

www.cirjam.com , editorjam@gmail.com

1.INTRODUCTION

The Pearson’s correlation coefficient ($\hat{r}$) is a point estimator which estimates r , population correlation coefficient. The statistic ($\hat{r}$) is used to describe the linear relationship between two variables, which are normally distributed. If the theoretical sampling distribution is not available, Monte Carlo (MC) procedure to approximate the estimation of the Pearson’s correlation coefficient ($\hat{r}$)

The goal is used Monte Carlo that would improve the accuracy of the correlation coefficient. Then, variance reduction is used to improve the efficiency of Monte Carlo methods.

In this study, the Monte Carlo, the Fisher’s z Method and the Bootstrap have been used for producing good approximate confidence intervals for the correlation coefficient. Also, in this paper the coverage of confidence interval for the correlation coefficient of the all methods have been used and the average estimate of correlation coefficient, standard error, confidence interval (CIs) and the width for the correlation coefficients are established. Finally, variance reduction method is used to reduce variability. (e.g. Importance Sampling).

2. Simulation Data

In order to simulate the paired data, two factors are considered, the true correlation and the sample size (n). For the purpose of this study, there will be five selected values of correlation, ρ (ρ = 0, 0.2, 0.4, 0.6, 0.8) and three values of sample size (n=15, 25, 50). In this study, there are total 15 simulation conditions (5*3).

Four methods, the Fisher’s z method, the Monte Carlo Method, the Variance reduction (e.g. Importance Sampling) and the Bootstrap Method, are used. Also, variance reduction method will be used to discover the small standard error.

3. Methodology

3.1. Naïve Monte Carlo

This study will estimate the confidence interval of the correlation coefficient using the independent and identically distributed pairs (X_{1i}, Y_{1i}), i=1... n from bivariate distribution. Then, the confidence interval of the correlation coefficient can be determined by using the Monte Carlo Simulation.

Here, the Monte Carlo Method will approximate the coverage of the confidence interval for the correlation coefficient and to estimate the true Pearson's correlation($\hat{r}$) as in this study, there is no theoretical sampling distribution. For a certain value of ρ, sample size (n=20, 25, 50) and 10000 (S=10000) random samples will be produced from $N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right)$

In simple linear regression, one typically estimates the correlation coefficient ρ between two normally distributed variables by its sample analog r_n (Samaniego 252). It can be shown that:

$$\sqrt{n}(\hat{r}_n - \rho) \overset{D}{\rightarrow} N(0, (1-\rho^2)^2) \quad \text{As} \quad n \rightarrow \infty \qquad (1)$$

In this method, the pair (X_{1i}, Y_{1i}) is simulated from the bivariate normal distribution $N_2[0,0,1,1,\rho]$. Then

$Z = \frac{\hat{r}_n - \rho}{\frac{(1-\rho^2)}{\sqrt{n}}}$ is calculated by using the central limit theorem. If Z > 1.96 then I (Z > 1.96) = 1and zero otherwise, this process will be repeated S times to estimate the mean and standard error of MC.

Monte Carlo Confidence intervals for r_n are generally based on the equation (1) and are given below:

$$\left[\hat{r}_n - \Phi(\tfrac{\alpha}{2})\, \frac{(1-\rho^2)}{\sqrt{n}}, \hat{r}_n + \Phi(1 - \tfrac{\alpha}{2}) \frac{(1-\rho^2\ )}{\sqrt{n}}\right] \qquad (2)$$

The width of the confidence interval for the correlation coefficient is calculated in this study.

3.2. Fisher’s z Method

R. A. Fisher recommended transforming to the variable $z(\rho)$ via the transformation, where $z(\rho)$ is given by:

$$z(\rho) = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} \qquad (3)$$

The transformation was motivated by the fact that its asymptotic variance does not depend on ρ and that it converges more quickly than *r* does (Samaniego 252). This statistic is now called **“**Fisher's Z". This transformation stabilizes the variance so that:

$$\sqrt{\boldsymbol{n}}\big(z(\hat{\rho}_n)\ - z(\rho)\big) \overset{D}{\rightarrow} N(0, 1) \qquad (4)$$

In this method also, (X_{1i}, Y_{1i}) are simulated from the bi-variate normal distribution $N_2[0,0,1,1,\rho]$. Then, $Z = \frac{z(\hat{\rho}_n)\ - z(\rho)}{\sqrt{n}}$ is computed. If Z > 1.96 then I (Z > 1.96) = 1 and zero otherwise. Also, the process is repeated again S times and the mean and standard error of Fisher's Z are calculated.

A confidence interval for Fisher's Z**,** $z(\rho)$, is computed by using the below equation:

$$\left[z(\hat{\rho}_n) - \frac{\Phi(\frac{\alpha}{2})}{\sqrt{\mathrm{n}}}, z(\hat{\rho}\ _n) + \frac{\Phi(1-\frac{\alpha}{2})}{\sqrt{\mathrm{n}}}\right] = [z^l, z^u] \qquad (5)$$

Then, the confidence interval $z(\rho)$ is inverted to estimate the confidence interval for ρ:

$$\left[\frac{e^{2z^l}-1}{e^{2z^l}+1}, \frac{e^{2z^u}-1}{e^{2z^u}+1}\right] \qquad (6)$$

3.3. Bootstrap Method

The aim here is to get the asymptotic distribution of the sample correlation coefficient, *r*. If we let

$$\begin{pmatrix} m_x \\ m_y \\ m_{xx} \\ m_{yy} \\ m_{xy} \end{pmatrix} = \frac{1}{n} \begin{pmatrix} \sum x_i \\ \sum y_i \\ \sum {x_i}^2 \\ \sum {y_i}^2 \\ \sum x_i y_i \end{pmatrix}$$

And $s_x = \sqrt{m_{xx} - (m_x)^2}$,$s_y = \sqrt{m_{yx} - (m_y)^2}$, and $s_{xy} = m_{xy} - m_x - m_y$, then $r = \frac{s_{xy}}{s_x s_y}$.

The delta method is found by using the central limit theorem:

$$\sqrt{n}\left\{\begin{pmatrix} m_x \\ m_y \\ m_{xx} \\ m_{yy} \\ m_{xy} \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ {\sigma_x}^2 \\ {\sigma_y}^2 \\ \sigma_{xy} \end{pmatrix}\right\} \overset{D}{\rightarrow} N_5 \left\{\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} - \Sigma\right\}$$

Using this method, the coverage probability is determined.

3.4. Variance Reduction: Importance Sampling

Here, the data (y) is generated from the bivariate normal distribution $N_2[0,0,1,1,\rho]$. Then $w(y) = \frac{f(y)}{g(y)} = \frac{\text{dmnorm (y,mu,sigma)}}{\text{dmnorm (y,mu2,sigma)}}$ is computed. In the next step, $Z = \frac{\hat{r}_n - \rho}{\frac{(1-\rho^2)}{\sqrt{n}}}$ is calculated.

If Z > 1.96 then I (Z > 1.96) *w(y). Finally, all the steps are repeated S times and the mean and standard error of MC are calculated.

4. Results

Table 1 shows the coverage of the confidence interval for the correlation coefficient using the MC sampling distribution, the Bootstrap (Bootstrap MSE) Method and the Fisher's z Method. In this study, the Fisher's z Method shows a bigger width than the BT method. Therefore, the Fisher's z Method shows a larger coverage of the intervals than the Bootstrap Simulation. Comparing the coverage between the Bootstrap and the Fisher's z Methods, it can be observed that Bootstrap became smaller as the sample size increases.

Table 2 displays the confidence interval, the standard error and the width of CI for Monte Carlo. It is clear that the width of the confidence interval decreases as the sample size increases for each ρ. Also, the standard error decreases when the sample size increases for each ρ. This result shows that a bigger sample size is related with a true estimate of ρ.

Table 3 presents the variance reduction by using importance sampling - unstandardized weights. The table shows the standard error of this method, the confidence interval and the width of CI. It is shown that the width of the confidence interval increases as the sample size increases for each ρ. The standard error here decreases more than the Naïve Monte Carlo for each ρ.

Table 1: Coverage of intervals for the estimation of the correlation coefficient for MC: Naive Monte Carlo Method, Fisher: Fisher Method, sim.MSE: simulation Method, sim.q: simulation Method based on quantiles

ρ	n	MC	Fisher's Z	Bootstrap MSE	Sim.q
0	15	0.8752	0.9148	0.9625	0.9468
	25	0.9065	0.9324	0.9468	0.9501
	50	0.9319	0.9431	0.9420	0.9498
0.2	15	0.8801	0.9231	0.9606	0.9441
	25	0.9140	0.9365	0.9483	0.9528
	50	0.9276	0.9400	0.9344	0.9466
0.4	15	0.8832	0.9226	0.9697	0.9470
	25	0.9004	0.9244	0.9422	0.9441
	50	0.9275	0.9394	0.9380	0.9446
0.6	15	0.8811	0.9274	0.9752	0.9464
	25	0.9068	0.9368	0.9534	0.9503
	50	0.9279	0.9412	0.9378	0.9475
0.8	15	0.8783	0.9226	0.9829	0.9445
	25	0.9068	0.9327	0.9585	0.9492
	50	0.9282	0.9432	0.9436	0.9487

Table 2: Naive Monte Carlo

ρ	n	$\hat{\hat{r}}_n$	SE	Lower limit	Upper limit	Width
0	20	- 0.0006841315	0.22741436	- 0.4492943	0.4366636	0.8859586
	25	0.0031949871	0.20391388	- 0.39216347	0.3943166	0.7864800
	50	- 0.0001045960	0.14333765	- 0.27946727	0.2784509	0.5579182
0.2	20	0.1935608907	0.22112551	- 0.26264024	0.5959719	0.8586122
	25	0.1966145010	0.19706008	- 0.21376567	0.5503549	0.7641205
	50	0.1980162855	0.13788643	- 0.08354724	0.4587595	0.5423068
0.4	20	0.3929804751	0.19673612	- 0.03605734	0.7218536	0.7579109
	25	0.3929961484	0.17401970	0.01611421	0.6887630	0.6726488
	50	0.3971043350	0.12125242	0.14459053	0.6109903	0.4663997
0.6	20	0.5899780688	0.15450382	0.22672329	0.8285812	0.6018579
	25	0.5938542714	0.13541788	0.28072914	0.8122404	0.5315113
	50	0.5962610979	0.09224753	0.39774157	0.7548107	0.3570691
0.8	20	0.7917126904	0.09178333	0.56291471	0.9226877	0.3597730
	25	0.794568243	0.07940452	0.60618550	0.9117592	0.3055737
	50	0.7965827399	0.05382884	0.67442324	0.8849812	0.2105580

Table 3: Importance Sampling

ρ	n	SE	Lower limit	Upper limit	Width
0	20	0.010332500	0.5371283	1.425926	0.8887972
	25	0.010118754	0.5602496	1.697652	1.1374025
	50	0.010941349	0.5055317	1.658526	1.1529945
0.2	20	0.008926193	0.7046792	1.471421	0.7667415
	25	0.009097279	0.6066346	1.393767	0.7871328
	50	0.011084896	0.5336968	1.569317	1.0356197
0.4	20	0.011756620	0.5555878	1.517867	0.9622768
	25	0.010475668	0.6766047	1.728799	1.0521939
	50	0.009403558	0.6753384	1.462166	0.7868273
0.6	20	0.010240777	0.7916117	1.780885	0.989273
	25	0.009317214	0.8294164	1.620201	0.7907845
	50	0.010703651	0.7320566	1.736110	1.0040537
0.8	20	0.010703651	0.6291310	1.880384	1.2512527
	25	0.013465334	0.7063542	1.581554	0.8751995
	50	0.010205929	0.7278479	1.516217	0.7883692

5. Conclusion

In conclusion, the Fisher's z and the Bootstrap Simulations do better than the Monte Carlo. In this study the Fisher's z Method is recommended, because it is not changed by increasing the sample size. Both the Monte Carlo and the Fisher's z Methods are easier to program than the Bootstrap Method. The data in the Monte Carlo and the Fisher's z Method are simulated from a bivariate normal distribution. Since the data is generated from a bivariate normal distribution the process is easier to arrive on a conclusion.

The Monte Carlo standard error decrease by increasing the sample size and
ρ. In the variance reduction method, the standard error is smaller compared to the standard error in Naïve Monte Carlo method.

References

[1] Madel, Micha. "Simulation-Based Confidence Intervals for Functions With Complicated Derivatives." (2013). Print.

[2] Samaniego, Francisco J. *Stochastic Modeling and Mathematical Statistics: A Text for Statisticians and Quantitative Scientists*. Print.