Journal of Advances in Mathematics and Computer Science

30(2): 1-8, 2019; Article no.JAMCS.45496
ISSN: 2456-9968
(Past name: British Journal of Mathematics & Computer Science, Past ISSN: 2231-0851)

Confidence Interval Estimate of the Correlation Coefficient for Age and Systolic Blood Pressure of 20, 30 and 50 Individuals

Onyedikachi O. John[1*]

[1]Department of Physical Sciences, Rhema University, Nigeria.

Author's contribution

The sole author designed, analysed, interpreted and prepared the manuscript.

Article Information

DOI: 10.9734/JAMCS/2019/45496
Editor(s):
(1) Dr. Morteza Seddighin, Professor, Indiana University East Richmond, USA.
(2) Dr. Sheng Zhang, Professor, Department of Mathematics, Bohai University, Jinzhou, China.
Reviewers:
(1) Eric S. Hall, USA.
(2) Cliff Richard Kikawa, Namibia University of Science and Technology, Namibia.
(3) Siana Halim, Petra Christian University Surabaya, Indonesia.
(4) Kamil Alakuş, Ondokuz Mayıs University, Turkey.
Complete Peer review History: http://www.sciencedomain.org/review-history/28053

Original Research Article

Received: 03 October 2018
Accepted: 18 December 2018
Published: 01 January 2019

Abstract

A formula for the confidence interval of correlation coefficient based on the Fishers' z-transformation, (where z = arctanh(r) and is used to convert the skewed distribution of sample correlation, r to a normal distribution) was applied to the age and systolic blood pressure of 20 individuals. Confidence intervals using percentile and bias-corrected and accelerated (BCa) bootstrap with 2000 replications were also obtained for the same sample sizes respectively. The interval lengths from the derived formula are 0.5997, 0.5073 and 0.3297 for 20, 30 and 50 observations respectively. The interval lengths from the percentile bootstrap are 0.6223, 0.5303 and 0.3077, while interval lengths from BCa bootstrap are 0.6214, 0.4958 and 0.3031 respectively. The interval length decreases as the sample size increases, giving a more accurate confidence interval. The derived formula gives a slightly shorter interval length for $n = 20$.

Keywords: Bootstrap; confidence interval; correlation coefficient; Fishers' z-transformation; jackknife.

1 Introduction

The correlation coefficient describes the degree and direction of relationship between two quantitative variables, [1,2]. A positive sign indicates a direct relationship, while a negative sign indicates an inverse

**Corresponding author: E-mail: johnkady@yahoo.com;*

relationship. The Pearson product-moment correlation coefficient formula r is usually applied to estimate the population correlation coefficient, [3]. Since the population correlation coefficient is not known, it is important to estimate the interval within which it is expected to lie. This confidence interval has the property that if random sampling were repeated infinitely many times, $100(1-\alpha)\%$ of the generated set of points representing the confidence interval will contain the true parameter, [4].

The bootstrap method proposed by Efron and Tibshirani, [5], can be used to compute the confidence interval for the unknown correlation coefficient. Hall, [6], gives proof based on the Edgeworth expansion that the BCa method produces a confidence interval that is second-order accurate, irrespective of the distribution. Methods that provide confidence intervals without the need for Monte Carlo approximations to the bootstrap distribution for the estimator are found in [7] and [8]. Phuenaree and Sanorsap, [9], compared the performance of standard bootstrap and percentile bootstrap for different distributions. Tsagris et al. [10], examined the correlation coefficient in the bivariate Poisson and Negative Binomial distributions. The question of convergence of bootstrap estimates has also been of concern. The law of large numbers implies that $S_B^{*2} \to S^2$ almost surely as $B \to \infty$, where B is the number of bootstrap samples and S the standard deviation, according to Tsagris et al. [10]. Bickel and Freedman [11] did extensive work on the convergence of bootstrap distribution.

In this paper, results from a formula for the confidence interval of the correlation coefficient based on the Fishers' z transformation, [12], were compared to the results from bootstraps confidence interval methods for different sample sizes. The data used were obtained from Clinitek Medical Diagnostic Laboratories, 2018 records.

2 Fisher's z Transformation

The transformation due to Fisher, [12], expresses that the statistic

$$z_r = 0.5 \ln\left(\frac{1+r}{1-r}\right) \tag{2.1}$$

is approximately distributed with mean

$$\mu_{z_r} = 0.5 \ln\left(\frac{1+\rho}{1-\rho}\right) \text{ and standard deviation } \sigma_{z_r} = \frac{1}{\sqrt{n-3}} \tag{2.2}$$

If X and Y are two random variables having a joint bivariate normal distribution with correlation coefficient ρ, then the $100(1-\alpha)\%$ confidence interval for μ_{z_r} is given by

$$z_r - z_{\alpha/2}\left(1/\sqrt{n-3}\right) < \mu_{z_r} < z_r + z_{\alpha/2}\left(1/\sqrt{n-3}\right) \tag{2.3}$$

Based on this Fisher's z transformation, substituting equation 2.2 in equation 2.3 will give:

$$z_r - z_{\alpha/2}\left(1/\sqrt{n-3}\right) < 0.5 \ln\left[(1+\rho)/(1-\rho)\right] < z_r + z_{\alpha/2}\left(1/\sqrt{n-3}\right) \tag{2.4}$$

$$e^{2z_r - z_{\alpha/2}\left(2/\sqrt{n-3}\right)} < (1+\rho)/(1-\rho) < e^{2z_r + z_{\alpha/2}\left(2/\sqrt{n-3}\right)}$$

Hence the $100(1-\alpha)\%$ confidence interval for the correlation coefficient ρ is obtained as

$$\frac{e^{2z_r - z_{\alpha/2}\left(2/\sqrt{n-3}\right)} - 1}{e^{2z_r - z_{\alpha/2}\left(2/\sqrt{n-3}\right)} + 1} < \rho < \frac{e^{2z_r + z_{\alpha/2}\left(2/\sqrt{n-3}\right)} - 1}{e^{2z_r + z_{\alpha/2}\left(2/\sqrt{n-3}\right)} + 1} \tag{2.5}$$

3 Percentile Bootstrap

From the original data, $(x_1, y_1),...,(x_n, y_n)$, bootstrap data sets $(x_1^*, y_1^*),...,(x_n^*, y_n^*)$ are generated with B replications. For each of the B replications $(X^{*1}, Y^{*1}),...,(X^{*B}, Y^{*B})$, we compute the Pearson product-moment correlation coefficients $r_1^*,...,r_B^*$. Based on these B bootstrapped correlation values, the $100(1-\alpha)\%$ bootstrap percentile interval for the correlation coefficient ρ is given by

$$r^*_{B(\alpha/2)} < \rho < r^*_{B(1-(\alpha/2))} \tag{3.1}$$

where $r^*_{B(\alpha/2)}$ is the $B(\alpha/2)th$ value in the ordered list of the bootstrap distribution of r^*, and $r^*_{B(1-(\alpha/2))}$ is the $B(1-(\alpha/2))th$ value in the ordered list of the bootstrap distribution of r^*, [5].

4 Bias-Corrected and Accelerated BCa Bootstrap

The bias-corrected and accelerated bootstrap is an improvement of the percentile bootstrap where the interval limits are given by percentiles that depend on the accelerated, $\hat{a}$, and bias-correction $\hat{z}_0$ values. Thus the $100(1-\alpha)\%$ BCa confidence interval for the correlation coefficient according to Efron and Tibshirani, [5] is given by

$$r^*_{B\alpha_1} < \rho < r^*_{B\alpha_2} \tag{4.1}$$

where

$$\left.\begin{aligned} \alpha_1 &= \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1-\hat{a}\left(\hat{z}_0 + z_{\alpha/2}\right)}\right) \\ \alpha_2 &= \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-(\alpha/2)}}{1-\hat{a}\left(\hat{z}_0 + z_{1-(\alpha/2)}\right)}\right) \end{aligned}\right\} \tag{4.2}$$

and $\hat{z}_0$ is computed directly from the proportion of the correlation coefficients of the B bootstrap replications that is less than the correlation coefficient of the original sample data:

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#(r_B^* < r)}{B}\right) , \tag{4.3}$$

The expression for the accelerated value $\hat{a}$ is given by

$$\hat{a} = \frac{\sum_{i=1}^{n}\left(r_{(i)} - r_{(\cdot)}\right)^3}{6\left\{\sum_{i=1}^{n}\left(r_{(i)} - r_{(\cdot)}\right)^2\right\}^{3/2}} , \tag{4.4}$$

where $r_{(i)}$ is the ith jackknife replication of the correlation coefficient r and $r_{(\cdot)} = \left(\sum_{i=1}^{n} r_{(i)}\right)\Big/ n$.

5 Results and Discussion

The Pearson product-moment correlation coefficient of age and systolic blood pressure of a sample of 20 individuals gave, $r = 0.6097$. The 95% confidence limits for the correlation coefficient, using equation 2.5 are $(0.2289, 0.8286)$. The 95% percentile bootstrap confidence limits based on 2000 replications are $(0.2280, 0.8503)$. The values of $\hat{a}$ and $\hat{z}_0$ are respectively 0.011 and 0.00827, and upon substitution into equation 4.2 gives $(\alpha_1, \alpha_2) = (0.0269, 0.9780)$. Thus the 95% BCa bootstrap confidence limits based on 2000 replications are $(0.2326, 0.8540)$. Age and Systolic Blood Pressure of 20, 30, and 50 individuals have been depicted in Appendix I.

The sample was increased to 30 individuals and the Pearson correlation gave, $r = 0.571$. The 95% confidence limits for the correlation coefficient, ρ, using equation 2.5 was computed to be $(0.2655, 0.7723)$. The 95% percentile bootstrap confidence limits based on 2000 replications are $(0.2676, 0.7879)$. The values of $\hat{a}$ and $\hat{z}_0$ are computed to be $\hat{a} = 0.0173$ and $\hat{z}_0 = 0.01$, which when substituted in equation 4.2 yields $(\alpha_1, \alpha_2) = (0.0302, 0.9798)$. Thus the 95% BCa bootstrap confidence limits, based on 2000 replications are $(0.2785, 0.7743)$. The BCa has an interval length of 0.4958 which is slightly shorter than the interval lengths of the percentile bootstrap and that obtained from equation 2.5. However, the interval length based on Fishers' z transformation is slightly shorter than that of the percentile bootstrap. The sample was also increased to 50 and the Pearson correlation coefficient was computed to be 0.654. The 95% confidence limits using equation 2.5, which is based on Fishers' z transformation, are $(0.4589, 0.7886)$ with an interval length of 0.3297. The 95% confidence limits for the correlation coefficient based on percentile bootstraps with 2000 replications are $(0.4812, 0.7889)$. For the BCa, the values of $\hat{a}$ and $\hat{z}_0$ are obtained as $(\hat{a}, \hat{z}_0) = (0.02, 0.042)$, resulting in $(\alpha_1, \alpha_2) = (0.0354, 0.9833)$ upon application of equation 4.2. Hence, the 95% BCa confidence limits based on 2000 replications are $(0.4949, 0.7980)$ with an interval length of 0.3031. This indicates that

when the sample size is small, the interval estimate based on Fishers' z transformation gives a better result than the percentile bootstrap. However, as the sample size increases, the percentile bootstrap performs better. The BCa bootstrap gives a more accurate result.

Table 1. 95% confidence interval and interval length

n	Method	Lower Limit	Upper Limit	Interval Length
20	Based on Fishers'	0.2289	0.8286	0.5997
	Percentile bootstrap	0.2280	0.8503	0.6223
	BCa bootstrap	0.2326	0.8540	0.6214
30	Base on Fishers'	0.2655	0.7723	0.5068
	Percentile bootstrap	0.2676	0.7879	0.5203
	BCa bootstrap	0.2785	0.7743	0.4958
50	Fishers'	0.4589	0.7886	0.3297
	Percentile bootstrap	0.4812	0.7889	0.3077
	BCa bootstrap	0.4949	0.7980	0.3031

6 Conclusion

The result from confidence interval formula derived from Fishers' z transformation compares favourably with the bootstrap confidence interval methods, namely percentile bootstrap and BCa bootstraps for a sample size of 20. However, as the sample size increases to 30, and finally to 50, the BCa gives a better and more accurate confidence interval for the correlation coefficient.

Competing Interests

Author has declared that no competing interests exist.

References

[1] Triola MF. Elementary statistics with multimedia study guide. Perason Education Inc. Boston; 2008.

[2] Agresti A, Franklin CA. Statistics: The art and science of learning from data. Pearson Education Inc. New Jersey; 2006.

[3] Mann PS. Introductory statistics: Second edition. John Wiley & Sons Inc. New Jersey; 2010.

[4] Chernick MR, LaBudde RA. An introduction to bootstrap methods with application to R. John Wiley & Sons Inc., New Jersey; 2011.

[5] Efron B, Tibshirani RJ. An introduction to the bootstrap. Chapman and Hall, New York; 1993.

[6] Hall P. The bootstrap and edgeworth expansion. Springer-Verlag, New York; 1992.

[7] DiCiccio TJ, Efron B. More accurate confidence intervals in exponential families. Biometrika. 1992; 79:231-245.

[8] DiCiccio TJ, Romano JP. The automatic percentile method: Accurate confidence limits in parametric models. Can. J. Stat. 1989;17:155-169.

[9] Phuenaree B, Sanorsap S. An interval estimate of pearson's correlation coefficient by bootstrap methods. Asian Journal of Applied Sciences. 2017;05(03).

[10] Tsagris M, Elmatzoglou I, Frangos CC. Assessment of performance of correlation estimates in discrete bivariate distribution using bootstrap methodology; 2015.
Available:https://arxiv.org/pdf/1151.01677

[11] Bickel PJ, Freedman DA. Some asymptotic theory for the bootstrap. Annals of Statistics. 1981;9(6): 1196-1217.

[12] Fisher RA. Statistical methods for research works. Oliver and Boyd, London; 1934.

APPENDIX I

Age and Systolic Blood Pressure of 20, 30, and 50 individuals

Age	Systolic blood pressure
61	150
48	100
32	100
38	110
40	130
57	140
46	110
42	140
51	140
54	130
48	120
53	130
69	160
36	120
60	140
55	150
43	160
39	130
36	120
46	140

Age	Systolic blood pressure
61	150
48	100
32	100
38	110
40	130
57	140
46	110
42	140
51	140
54	130
48	120
53	130
69	160
36	120
60	140
55	150
43	160
39	130
36	120
46	140
42	120
40	150
52	120
64	140
53	140
58	160
70	150
60	140
60	120
37	100

Age	Systolic blood pressure	Age	Systolic blood pressure
67	160	70	150
40	120	75	160
47	130	65	150
60	140	52	140
55	150	55	160
37	100	61	150
43	160	48	100
50	170	32	100
60	120	38	110
30	110	40	130
39	130	57	140
36	120	37	90
46	140	46	110
42	120	42	140
40	130	51	140
33	100	60	110
52	120	54	130

Age	Systolic blood pressure	Age	Systolic blood pressure
64	140	32	100
41	150	30	100
40	140	30	110
50	110	48	120
60	140	53	130
53	140	69	160
58	160	54	160
39	130	36	120

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://www.sciencedomain.org/review-history/28053