Colloquium Biometricum 38
2008, 51–56

ON TWO CONFIDENCE INTERVALS FOR THE CORRELATION COEFFICIENT

Joanna Tarasińska

Department of Applied Mathematics and Computer Science
University of Life Sciences in Lublin
Akademicka 13, 20-950 Lublin, Poland
e-mail: joanna.tarasinska@up.lublin.pl

Summary

An interval estimation for the correlation coefficient in bivariate normal distribution is considered. A case of a very small sample size *n*=5 is taken into account. The coverage probability and an average length for two confidence intervals are evaluated by means of a simulation study. One confidence interval is based on Fisher's z transformation and the another on probability density function.

Keywords and phrases: correlation coefficient, interval estimation, coverage probability, length of confidence interval

Classification AMS 2000: 62F25, 62H20

1. Introduction

Let us consider the sample $(X_i, Y_i),\ i = 1\ldots n$ from bivariate normal distribution with covariance matrix $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$. The estimate of correlation coefficient ρ is the sample correlation coefficient

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} .$$

An inference about ρ has been investigated by many authors. The first one was Fisher (1915, 1921). He derived the probability density function (pdf) for r and pointed the famous transformation, called Fisher's z transformation, $z = \frac{1}{2}\log\frac{1+r}{1-r}$ which is extremely useful to obtain confidence intervals for ρ as z has got approximately normal distribution with expected value $\xi = \frac{1}{2}\log\frac{1+\rho}{1-\rho}$ and variance $\frac{1}{n-3}$.

Fisher (1915) gave the formula for the probability density function of r in two forms, one in terms of infinite sums and the other in terms of (n-2)-th derivatives. Hotelling (1953) derived the probability density function of r in the following form:

$$f(r;\rho) = \frac{n-2}{\sqrt{2\pi}}\frac{\Gamma(n-1)}{\Gamma(n-0.5)}\left(1-\rho^2\right)^{0.5(n-1)}\left(1-r^2\right)^{0.5(n-4)}(1-\rho r)^{1.5-n} G\left(\frac{1}{2},\frac{1}{2};n-\frac{1}{2},\frac{1+\rho r}{2}\right), \quad (1.1)$$

where $G(a,b;c,x) = \sum_{j=0}^{\infty}\frac{\Gamma(a+j)}{\Gamma(a)}\frac{\Gamma(b+j)}{\Gamma(b)}\frac{\Gamma(c)}{\Gamma(c+j)}\frac{x^j}{j!}$.

Though the problem of confidence intervals for ρ is very old and has been intensively investigated, it still attracts attention of statisticians. Sun and Wong (2007) did some simulations to compare the coverage probability (i.e. the percentage of a true parameter value falling within the intervals) for nine 95% confidence intervals obtained from different transformations . Among others they considered Fisher's z transformation and four others, more accurate , given by Hotelling (1953). They also considered the c.i. based on the pdf of form (1.1). Their simulation study for sample size $n = 10$ showed the similar coverage probability for all considered intervals. Actually their paper motivated this one. The aim of the paper is to give some results for as small sample size as $n = 5$. The intervals based on Fisher's z transformation and on probability density function (1.1) are considered. Obtained results concern not only coverage probability but also the length of intervals.

The $(1-\alpha)100\%$ confidence interval based on Fisher's *z* transformation is derived from formula:

$$\left(\tanh\left(z-\frac{u_\alpha}{\sqrt{n-3}}\right)\ ;\ \tanh\left(z+\frac{u_\alpha}{\sqrt{n-3}}\right)\right), \tag{1.2}$$

where u_α is $\left(1-\frac{\alpha}{2}\right)$-th percentile of standard normal distribution and $\tanh(x)=\frac{e^{2x}-1}{e^{2x}+1}$.

The ends ρ_1 and ρ_2 of confidence interval based on pdf (1.1) can be calculated (see for example Cramer (1958)) as solutions of integral equations:

$$\int_{-1}^{r_0} f(r;\rho_1)dr=\frac{\alpha}{2} \text{ and } \int_{-1}^{r_0} f(r;\rho_2)dr=1-\frac{\alpha}{2}\ , \tag{1.3}$$

where r_0 is the observed sample correlation coefficient.

2. Results

10 000 random samples, each of size 5, from bivariate normal distribution $N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right)$ with $\rho=-0.9$, -0.5, 0, 0.5 and 0.9 were generated. The coverage probability and average length of 95%, 90% and 99% confidence intervals (1.2) and (1.3) were calculated. In case of (1.3) four terms in the series $G\left(\frac{1}{2},\frac{1}{2};n-\frac{1}{2},\frac{1+\rho r}{2}\right)$ were taken. It gives very accurate results for pdf as Hotelling (1953) proved that the series in $f(r;\rho)$ converges very rapidly and that the first term alone is often a sufficient approximation. Additionally, as sample size $n=5$ is very small, it was checked that for four terms

$\int_{-1}^{1} f(r;\rho) \approx 0.9995$ for all considered ρ. Thus approximation with four terms turned out to be very accurate. All calculations were made by using an own programme in the Maple application. Solutions of (1.3) were found by the bisection method.

The standard errors of coverage probability are 0.0022 ; 0.003 and 0.001 for $\alpha = 0.05$; 0.1 and 0.01 respectively.

The results are given in Table 1. Method *z* in it denotes c.i. calculated by (1.2) whereas *f* –by (1.3). It can be seen that coverage probabilities for both methods are very close to nominal 0.95 except the case $\rho = -0.9$ for the method based on pdf. The reason for that is presented in Fig.1 which gives the plot of function $g(r_0) = \int_{-1}^{r_0} f(r;\rho)dr$ for $n = 5$, $\rho = -0.9$. The plot is very "flat" so the solution of the equation $\int_{-1}^{r_0} f(r;\rho)dr = 1 - \frac{\alpha}{2}$ can not be sufficiently accurate. The situation is even worse for 99% c.i. and quite good for 90%. For greater sample size this effect is negligible. For example if *n*=10 then coverage probability for method *f* , confidence level 99% and $\rho = -0.9$ is 0.988. It should be noted that c.i. based on pdf are shorter than the one obtained by means of *z* transformation.

Table 1. Coverage probabilities and average lengths of $(1-\alpha)100\%$ c.i. based on Fisher's z transformation and on pdf

ρ	method	Coverage probability			Average length		
		$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$
-0.9	z	0.958	0.923	0.988	0.800	0.628	1.133
	f	0.933	0.905	0.865	0.702	0.568	0.922
-0.5	z	0.957	0.915	0.988	1.443	1.272	1.691
	f	0.948	0.903	0.981	1.342	1.150	1.626
0	z	0.954	0.908	0.985	1.585	1.427	1.787
	f	0.950	0.900	0.989	1.470	1.277	1.741
0.5	z	0.952	0.913	0.987	1.440	1.270	1.687
	f	0.950	0.897	0.991	1.347	1.149	1.654
0.9	z	0.955	0.922	0.988	0.793	0.628	1.132
	f	0.952	0.901	0.991	0.788	0.605	1.196

But also this transformation gives quite good results (coverage probability) in spite of small sample size and the fact that some handbooks recommend it for $n \geq 10$ (Krysicki et al. , 1986). Newerthless for 90% c.i. this method gives a little too much coverage probability.

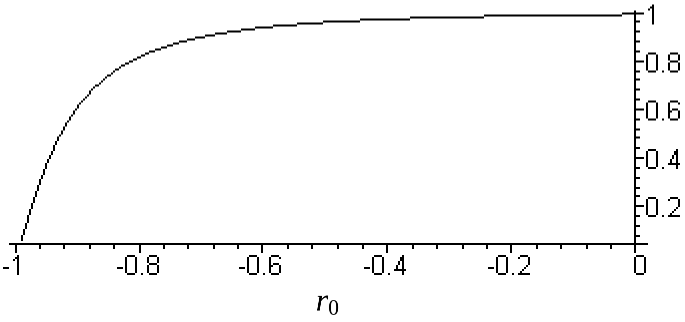

Fig. 1. The plot of $g(r_0)$, $\rho = -0.9$

References

Cramer H. (1958). *Metody matematyczne w statystyce*. PWN, Warszawa.

Fisher R.A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* 10, 507-521.

Fisher R.A. (1921). On the “probable error” of a coefficient of correlation deduced from a small sample. *Metron* 1, 3-32

Hotelling H. (1953). New light on the correlation coefficient and its transform. *J. Roy. Statist. Soc.* Ser. B 15,193-232.

Krysicki W., Bartos J., Dyczka W., Królikowska K., Wasilewski M.J. (1986). *Rachunek prawdo-podobieństwa i statystyka matematyczna w zadaniach*. PWN, Warszawa.

Sun Y., Wong A.C.M. (2007). Interval estimation for the normal correlation coefficient. *Statistics & Probability Letters* 77, 1652-1661.

O DWÓCH PRZEDZIAŁACH UFNOŚCI DLA WSPÓŁCZYNNIKA KORELACJI

Streszczenie

Rozważa się dwa przedziały ufności dla współczynnika korelacji w dwuwymiarowym rozkładzie normalnym. Bierze się pod uwagę próby o bardzo małej liczebności $n = 5$. Przy pomocy symulacji wyznacza się prawdopodobieństwo pokrycia oraz średnią długość przedziałów obliczonych na podstawie przekształcenia *z* Fishera oraz wyznaczonych na podstawie funkcji gęstości współczynnika korelacji z próby.

Słowa kluczowe: współczynnik korelacji, estymacja przedziałowa, prawdopodobieństwo pokrycia, długość przedziału ufności

Klasyfikacja AMS 2000: 62F25, 62H20