Proc. Estonian Acad. Sci. Phys. Math., 1996, **45**, 2/3, 259–266
https://doi.org/10.3176/phys.math.1996.2/3.17

THE ASYMPTOTIC DISTRIBUTION OF THE SAMPLE CORRELATION COEFFICIENT

Olga RYKUNOVA and Anne-Mai PARRING

Tartu Ülikooli matemaatilise statistika instituut (Institute of Mathematical Statistics, University of Tartu), Liivi 2, EE-2400 Tartu, Eesti (Estonia)

Received 14 December 1995, accepted 4 June 1996

Abstract. For large samples the asymptotic distribution of the sample correlation coefficient is a normal distribution. The variance of that asymptotic distribution is calculated using the approximate linearization. It depends on the fourth-order central moments of the general population. An example is given illustrating usefulness of that result.

Key words: approximate linearization, asymptotic normality, correlation coefficient, matrix derivative.

1. THE PROBLEM

The asymptotic normality of statistics is very often used for making statistical inferences. In simple cases the parameters of the asymptotic distribution do not depend on the distribution of the general population, but usually they do. In this paper the asymptotic distribution of the well-known statistic, the sample correlation coefficient, is considered. Its asymptotic distribution is normal; the parameters of that distribution are calculated.

2. THE APPROXIMATE LINEARIZATION

One possible method to find the asymptotic distribution is the approximate linearization [1], pp. 33–34.

Let $\mathbf{V_n}$ be a random vector converging in probability to a same dimensional constant vector $\mathbf{c}$,

$$\mathbf{V_n} \xrightarrow{p} \mathbf{c}$$

and

$$\sqrt{n}(\mathbf{V_n} - \mathbf{c}) \xrightarrow{\ell} N(\mathbf{0}, \Xi),$$

where $N(\mathbf{0}, \Xi)$ is a normal distribution with the mean vector $\mathbf{0}$ and the variance matrix Ξ. Here $\xrightarrow{\ell}$ marks converging in the distribution.

Denoting by $\mathbf{V}_*$ a random vector with the distribution $N(\mathbf{0}, \Xi)$, $\mathbf{V}_* \sim N(\mathbf{0}, \Xi)$, we have

$$\sqrt{n}(\mathbf{V_n} - \mathbf{c}) \sim \mathbf{V}_* + o(1),$$

where $o(1) \xrightarrow{p} 0$. Let now $\mathrm{h}(\mathbf{V})$ be a twice differentiable function with a nonvanishing first derivative at $\mathbf{c}$,

$$\frac{\partial h}{\partial \mathbf{V}}(\mathbf{c}) \neq 0,$$

and with a bounded second derivative. From Taylor's formula it follows that

$$h(\mathbf{V_n}) = h(\mathbf{c}) + \frac{\partial h}{\partial \mathbf{V}}(\mathbf{c})'(\mathbf{V_n} - \mathbf{c}) + \ldots$$

or

$$\sqrt{n}(h(\mathbf{V_n}) - h(\mathbf{c})) = \frac{\partial h}{\partial \mathbf{V}}(\mathbf{c})\mathbf{V}_* + o(1).$$

As the variance matrix of the vector $\mathbf{V}_*$ is Ξ, $D\mathbf{V}_* = \Xi$, we get

$$\sqrt{n}(h(\mathbf{V_n}) - h(\mathbf{c})) \xrightarrow{\ell} N(0, \Pi),$$

where

$$\Pi = \frac{\partial h}{\partial \mathbf{V}}(\mathbf{c})\Xi\frac{\partial h}{\partial \mathbf{V}}(\mathbf{c})'. \tag{1}$$

So the approximate linearization gives us the possibility of calculating the asymptotic parameters for a wide range of functions of asymptotically normal statistics.

3. THE ASYMPTOTIC VARIANCE OF THE SAMPLE CORRELATION COEFFICIENT

Let us consider a two-dimensional random vector $\mathbf{X}$,

$$\mathbf{X} = (X_1, X_2)'$$

with the variance matrix Σ,

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}.$$

The correlation coefficient for this vector is ρ,

$$\rho = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}.$$

In real problems the variance matrix Σ is unknown. We can observe the random vector $\mathbf{X}$ and have as a sample its values

$$\mathbf{x_i} = (x_{i1},\ x_{i2}),\quad i = 1, 2, \ldots, n.$$

Instead of the matrix Σ we have to use its sample estimation $\mathbf{S_n}$,

$$\mathbf{S_n} = \begin{pmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{pmatrix},$$

where

$$\mathbf{S_n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x_i}\mathbf{x_i}' - \bar{\mathbf{x}}_\mathbf{n}\bar{\mathbf{x}}_\mathbf{n}'$$

and

$$\bar{\mathbf{x}}_\mathbf{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x_i}.$$

As we are looking for the asymptotic variance of the sample correlation coefficient, it is obvious [2] that as the vector $\mathbf{V_n}$ we have to use the vector $\mathbf{s_n}$,

$$\mathbf{s_n} = (s_{11},\ s_{12},\ s_{22})'.$$

Since $\mathbf{s_n} \xrightarrow{p} \boldsymbol{\sigma}$, where

$$\boldsymbol{\sigma} = (\sigma_{11},\ \sigma_{12},\ \sigma_{22})',$$

it is obvious that as the vector $\mathbf{c}$ we have to use the vector $\boldsymbol{\sigma}$. In such a choice

$$\mathbf{V}_* \sim N(0, \Xi),$$

where the matrix Ξ is ([3], p. 104)

$$\Xi = \mathbf{M_4} - \sigma\sigma' \tag{2}$$

and $\mathbf{M_4}$ is the matrix, composed of the fourth-order central moments,

$$\mathbf{M_4} = \begin{pmatrix} m_{1111} & m_{1112} & m_{1122} \\ m_{1112} & m_{1122} & m_{1222} \\ m_{1122} & m_{1222} & m_{2222} \end{pmatrix}$$

with elements

$$m_{ijkl} = E[(X_i - EX_i)(X_j - EX_j)(X_k - EX_k)(X_l - EX_l)],$$

$i, j, k, l = 1, 2$.

As ρ is a function of $\boldsymbol{\sigma}$, we must calculate the derivative

$$\frac{\partial \mathbf{h}}{\partial \boldsymbol{\sigma}}(\boldsymbol{\sigma}) = \rho\left(-\frac{1}{2\sigma_{11}},\ \frac{1}{\sigma_{12}},\ -\frac{1}{2\sigma_{22}}\right)'. \tag{3}$$

Following the formulae (1), (2), and (3), we get

$$\Pi = \rho^2 \left(\frac{m_{1111}}{4\sigma_{11}^2} + \frac{m_{1122}}{2\sigma_{11}\sigma_{22}} + \frac{m_{2222}}{4\sigma_{22}^2} \right) - \frac{\rho}{\sqrt{\sigma_{11}\sigma_{22}}} \left(\frac{m_{1112}}{\sigma_{11}} + \frac{m_{1222}}{\sigma_{22}} \right) + \frac{m_{1122}}{\sigma_{11}\sigma_{22}}. \tag{4}$$

So for a large n we may suppose that the sample correlation coefficient is asymptotically normal,

$$r \sim N\left(\rho, \sqrt{\frac{\Pi}{n}}\right).$$

4. SOME PROPERTIES OF THE ASYMPTOTIC VARIANCE OF THE SAMPLE CORRELATION COEFFICIENT

In the following we list some properties of the asymptotic variance of the sample correlation coefficient.

(1) The asymptotic variance Π depends only on the correlation coefficient ρ and the fourth-order central moments of the distribution of the general population.

(2) If X_1 and X_2 are independent, then the asymptotic variance Π is equal to one, $\Pi = 1$. Indeed, if X_1 and X_2 are independent, then $\rho = 0$, $m_{1122} = \sigma_{11}\sigma_{22}$, and from (4) we get $\Pi = 1$.

(3) If $X_2 = a + bX_1$ then the asymptotic variance Π is equal to zero, $\Pi = 0$. Indeed, if X_2 is a linear function of X_1, then $\rho = 1$ and denoting $m_{1111} = m$, we get the following equations:

$$\sigma_{22} = b^2\sigma_{11}, \quad \sigma_{12} = b\sigma_{11},$$

and

$$m_{1112} = bm, \quad m_{1122} = b^2m,$$
$$m_{1222} = b^3m, \quad m_{2222} = b^4m.$$

Now from (4) we get $\Pi = 0$.

5. THE BASIC IDEA FOR THE COMPARISON OF NORMAL AND NON-NORMAL CASES

If we consider a population with a normal distribution and a population with a non-normal distribution, the asymptotic variance of the sample correlation coefficient may be different. To see the differences caused by the type of distribution, we must choose the parameters of comparable distributions as similar as possible. As the value of the correlation

coefficient depends only on elements of the variance matrix, we must choose the distributions which have the same variance matrices. Then the value of the correlation coefficient for the chosen distributions is the same and the value of the asymptotic variance depends only on the fourth-order central moments.

Let us calculate the asymptotic variance Π for a normal population. It is known ([3], p. 106) that for a p-dimensional normal distribution the $p^2 \times p^2$ matrix of the fourth-order central moments is given by the variance matrix and is calculated by the formula

$$\mathbf{M_4} = (\mathbf{I}_{p^2} + \mathbf{I}_{p,p})(\Sigma \otimes \Sigma) + \operatorname{vec} \Sigma \operatorname{vec}' \Sigma,$$

where

$\mathbf{I}_{p^2}$ – the $p^2 \times p^2$ unit matrix,

$\mathbf{I}_{p,p}$ – the $p^2 \times p^2$ permutation matrix ([3], p. 22),

vec – operator which transforms a matrix into a vector by stacking the columns of the matrix one underneath the other (see [4], p. 30).

Hence the fourth-order central moment for the normal distribution is calculated by the formula

$$m_{ijgh} = \sigma_{jh}\sigma_{ig} + \sigma_{ih}\sigma_{jg} + \sigma_{ij}\sigma_{gh} \tag{5}$$

and we get

$$\begin{aligned} m_{1111} &= 3\sigma_{11}^2, \\ m_{1112} &= 3\sigma_{12}\sigma_{11}, \\ m_{1122} &= 2\sigma_{12}^2 + \sigma_{11}\sigma_{22}, \\ m_{1222} &= 3\sigma_{12}\sigma_{22}, \\ m_{2222} &= 3\sigma_{22}^2. \end{aligned}$$

Using these values in the formula (4), we get the asymptotic variance Π for a normal distribution

$$\Pi = (1 - \rho^2)^2. \tag{6}$$

The asymptotic variance of the sample correlation coefficient in case of a normal distribution depends only on the value of the correlation coefficient.

6. THE FAMILY OF DISCRETE NON-NORMAL DISTRIBUTION

For the planned comparison it would be good to have a family of discrete two-dimensional distributions for which the correlation coefficient may freely change in an interval from 0 to 1. To construct such a family, let us start from a very simple marginal distribution

x_i	-1	0	1
p_i	p	kp	kp^2,

where the parameters of the distribution, p $(0 < p < 1)$ and k $(k > 0)$, must satisfy the condition

$$p(1 + k + k^2) = 1.$$

By the proper choice of the parameters we can deform the distribution from a uniform distribution to a very asymmetric one. In practice that kind of distribution may be useful for describing the distribution of attitudes which have three possible values: negative, neutral, and positive.

Using such marginal distributions, we may construct different two-dimensional distributions. Really, defining the two-dimensional distribution by table P_1

x_{1i}/x_{2i}	-1	0	1
-1	p	0	0
0	0	kp	0
1	0	0	k^2p,

we get the functionally dependent components, $X_1 = X_2$. Defining the two-dimensional distribution by table P_2

x_{1i}/x_{2i}	-1	0	1
-1	p^2	kp^2	k^2p^2
0	kp^2	k^2p^2	k^3p^2
1	k^2p^2	k^3p^2	k^4p^2,

we get independent components.

Following now the results of Tiit [5], for the mixture of these two distributions $cP_1+(1-c)P_2$, $(0 \leq c \leq 1)$, we get the correlation coefficient equal to c. Such mixture is defined by the table

x_{1i}/x_{2i}	-1	0	1
-1	$p^2 + p^2ck(1+k)$	$kp^2(1-c)$	$k^2p^2(1-c)$
0	$kp^2(1-c)$	$k^2p^2 + p^2ck(1+k^2)$	$k^3p^2(1-c)$
1	$k^2p^2(1-c)$	$k^3p^2(1-c)$	$k^4p^2 + p^2ck^2(1+k)$.

The family of discrete distributions defined by the last table has the following properties:

(1) The marginal distributions and all the marginal moments of the components X_1 and X_2 are equal.

(2) Let a_k denote the marginal moment of the order k. Then we have

$$\begin{aligned} a_1 &= p(k^2 - 1) = 1 - p(k + 2), \\ a_2 &= p(k^2 + 1) = 1 - pk, \end{aligned}$$

and

$$\begin{aligned} a_{2k+1} &= a_1, \\ a_{2k} &= a_2, \end{aligned}$$

$k = 1, 2, \ldots$

Hence the variance matrix of that distribution is

$$\Sigma = \begin{pmatrix} pk(1+3kp) & cpk(1+3kp) \\ cpk(1+3kp) & pk(1+3kp) \end{pmatrix}.$$

(3) The fourth-order central moments are the following:

$$\begin{aligned} m_{1111} &= m_{2222} = m \\ &= kp(1+12kp-9k^2p^2(2+3kp)), \\ m_{1112} &= m_{1222} = cm, \\ m_{1122} &= cm + \sigma^4(1-c), \end{aligned}$$

where σ^2 denotes the variance of a component,

$$\sigma^2 = pk(1+3kp).$$

Using these properties, it is easy to get an expression for the asymptotic variance Π:

$$\Pi = \frac{1}{2\sigma^4}[c^3(m-\sigma^4) + c^2(\sigma^4 - 3m) + 2c(m-\sigma^4) + 2\sigma^4]. \qquad (7)$$

7. THE RESULTS

Comparing formulae (6) and (7), it is obvious that for a given value of the correlation coefficient we can get different values of the asymptotic variance of the sample correlation coefficient depending on the distribution of the general population. For the considered special family the differences depend on the value of the correlation coefficient of the general population and on the parameters of the discrete distribution p and k.

Figure 1 illustrates these differences in a special case. The parameters p and k of the discrete distribution are fixed: $k = 9$ and $p = 1/91$. Then the asymptotic variance Π of this discrete distribution is computed by the formula (7) for all possible values of the correlation coefficient c, $0 \leq c \leq 1$. The graph of the asymptotic variance has been drawn with the dashed line.

In addition, the asymptotic variance of a normal distribution is computed by the formula (6) for all possible values of the correlation coefficient c, $0 \leq c \leq 1$. The graph of that asymptotic variance has been drawn with the continuous line.

It is supposed that different distributions cause remarkable differences in the asymptotic variance. This fact opens a possibility of obtaining useful results by working with the general expression for the asymptotic variance of the sample correlation coefficient.

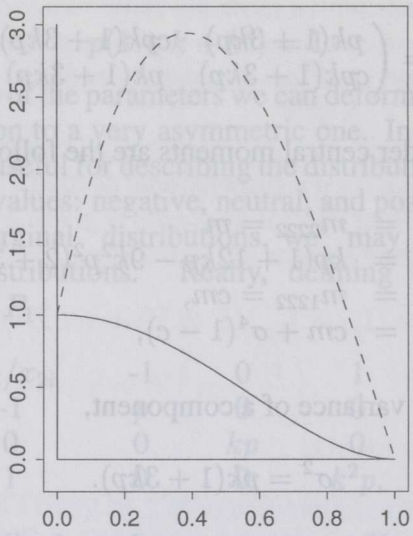

Fig. 1. Dependence of the asymptotic variance Π on the correlation coefficient of the general population.

REFERENCES

1. Barndorff-Nielsen, O. E. and Cox, D. R. *Asymptotic Techniques for Use in Statistics.* Chapman and Hall, London, New York, 1989.
2. Rykunova, O. *Korrelatsioonikordaja asümptootiline jaotus.* Bachelor work, 1995 (manuscript, in Estonian).
3. Kollo, T. *Matrichnaya proizvodnaya dlya mnogomernoj statistiki.* TU, Tartu, 1991 (in Russian).
4. Magnus, J. R. and Neudecker, H. *Matrix Differential Calculus with Applications in Statistics and Econometrics.* John Wiley & Sons, Chichester, 1988.
5. Tiit, E. Postroeniye diskretnykh mnogomernykh raspredelenij s zadannymi momentami. Diskretnyj analog normal'nogo raspredeleniya. *Trudy vychislitel'nogo tsentra. TGU*, 1984, **51**, 142–165 (in Russian).

VALIMI KORRELATSIOONIKORDAJA ASÜMPTOOTILINE JAOTUS

Olga RÕKUNOVA, Anne-Mai PARRING

Suurte valimite korral on valimi korrelatsioonikordaja asümptootiliseks jaotuseks normaaljaotus. Selle jaotuse dispersioon sõltub üldkogumi jaotusest. Kasutades lineaarset lähendamist on leitud eeskiri asümptootilise jaotuse dispersiooni arvutamiseks üldkogumi jaotuse neljandat järku tsentraalsete momentide kaudu. On toodud näide, mis kirjeldab saadud tulemuse kasulikkust.