ON THE SAMPLING DISTRIBUTION OF THE MULTIPLE CORRELATION COEFFICIENT

By S. S. Wilks*

The problem of finding the distribution of the multiple correlation coefficient in samples from a normal population with a non-zero multiple correlation coefficient was solved in 1928 by Fisher[1] by the application of geometrical methods. In his derivation he used the facts that the population value ρ of the multiple correlation coefficient is invariant under linear transformations of the independent variates, and that the distribution of the multiple correlation coefficient is independent of all population parameters except ρ.

In this paper it will be shown that the distribution of the multiple correlation coefficient can be derived directly from Wishart's[2] generalized product moment distribution without making use of geometrical notions and the property of the invariance of ρ under linear transformations of the independent variates. Furthermore, it will not be necessary to show that the distribution will be independent of all population parameters except ρ.

The population value of the multiple correlation coefficient between a variate x_1 and a set of variates $x_2, x_3, \cdots x_n$ is the ordinary correlation coefficient between x_1 and that linear function of the variates $x_2, x_3, \cdots x_n$ which will make this correlation a maximum. It can be expressed as $\rho^2 = 1 - \frac{\Delta}{\Delta_1}$ where Δ is the determinant of the correlations among all of the

*National Research Fellow in Mathematics.

[1]R. A. Fisher, The general sampling distribution of the multiple correlation coefficient, Proceedings of the Royal Society of London, series A, vol. 121 (1928), pp. 654-73.

[2]John Wishart, The generalized product moment distribution in samples from a normal multivariate population, Biometrika, vol. 20A (1928) pp. 32-52.

variates $x_1, x_2, \cdots x_n$ and Δ_1 is the determinant of correlations among the independent variates $x_2, x_3, \cdots x_n$. Denoting the sample value of ρ^2 by R^2 it is well known that $R^2 = 1 - \frac{D}{D_1}$, where D and D_1 are the determinants of sample correlations among the sets of variates $x_1, x_2, \cdots x_n$ and $x_2, x_3, \cdots x_n$ respectively.

Let us suppose a sample of N items to be drawn at random from the normal n-variate population whose distribution is

$$(1) \qquad \frac{\sqrt{A}}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i,j=1}^{n} A_{ij}(x_i - m_i)(x_j - m_j)}$$

where $A_{ij} = \frac{\Delta_{ij}}{\sigma_i \sigma_j \Delta}$, $\Delta = \left| \rho_{ij} \right|$ the determinant of correlations among the n variates, Δ_{ij} is the cofactor of ρ_{ij} in Δ, σ_i is the standard deviation of x_i and $A = \left| A_{ij} \right|$.

In the sample, let

$$\bar{x}_i = \frac{1}{N} \sum_{\alpha=1}^{N} x_{i\alpha},$$

and

$$a_{ij} = \frac{1}{N} \sum_{\alpha=1}^{N} (x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j),$$

where $x_{i\alpha}$ is the value of x_i for the α-th individual of the sample. Wishart[3] has proved that the simultaneous distribution function of the set $\{a_{ij}\}$, $(i, j = 1, 2, \cdots n)$ is

$$(2) \quad f(a) = \frac{\left(\frac{N}{2}\right)^{\frac{n(N-1)}{2}} A^{\frac{N-1}{2}}}{\pi^{\frac{n(n-1)}{4}} \Gamma\left(\frac{N-1}{2}\right) \Gamma\left(\frac{N-2}{2}\right) \cdots \Gamma\left(\frac{N-n}{2}\right)} e^{-\frac{N}{2}\sum_{i,j=1}^{n} A_{ij} a_{ij}} \left| a_{ij} \right|^{\frac{N-n-2}{2}}$$

---

[3] J. Wishart, loc. cit.

where $|a_{ij}|$ is the determinant of the a's.

We shall define a moment-generating function $\phi(\alpha,k)$ as

$$(3) \qquad \phi(\alpha,k) = \int e^{\alpha a_{11}} |a_{ij}|^{k} |a_{pq}|^{-k} f(\bar{a})\, d\bar{a},$$

where the integration is to be taken over the field of all possible values of the a's and $|a_{pq}|$ is the cofactor of a_{11} in $|a_{ij}|$. From this definition of $\phi(\alpha,k)$, it is clear that $\frac{\partial^h}{\partial\alpha^h}\phi(\alpha,k)\Big|_{\alpha=0}$ is the product moment $E\left[a_{11}^{h+k}(1-R^2)^k\right]$. It will be shown that this expectation exists for $h=-k$ which will yield the k-th moment of $(1-R^2)$, from which the distribution of R^2 can be found.

To find $\phi(\alpha,k)$ we observe that since (2) is a probability function, its value over the field of all possible values of the a's is unity. Hence, we must have

$$(4) \qquad \int e^{-\frac{N}{2}\sum_{i,j=1}^{n} A_{ij}a_{ij}} |a_{ij}|^{\frac{N-n-2}{2}} d\bar{a} = G,$$

where $G = \dfrac{\pi^{\frac{n(n-1)}{4}}\Gamma\left(\frac{N-1}{2}\right)\Gamma\left(\frac{N-2}{2}\right)\cdots\Gamma\left(\frac{N-n}{2}\right)}{\left(\frac{N}{2}\right)^{\frac{n(N-1)}{2}} A^{\frac{N-1}{2}}}$. This relation holds for all positive values of $N>n$ and for all values of A_{ij} which will make the matrix $\|A_{ij}\|$ positive definite.

If $f(\bar{a})$ be integrated with respect to $a_{11}, a_{12}, \cdots a_{1n}$, the resulting form will clearly be the distribution of the set of a's contained in $|a_{pq}|$ and will be

$$(5)\ \frac{\left(\frac{N}{2}\right)^{\frac{(n-1)(N-1)}{2}} B^{\frac{N-1}{2}}}{\pi^{\frac{(n-1)(n-2)}{4}}\,\Gamma\left(\frac{N-1}{2}\right)\cdots\Gamma\left(\frac{N-n+1}{2}\right)}\, e^{-\frac{N}{2}\sum_{p,q=2}^{n} B_{pq}a_{pq}}\, |a_{pq}|^{\frac{N-n-1}{2}},$$

where B_{pq} is the element in the p-th row and q-th column of the reciprocal form[4] of the determinant which is the co factor of the term in the first row and first column of the reciprocal form of $|A_{ij}|$. The value of B_{pq} in terms of correlation coefficients and standard deviations is $\frac{\Delta^{(1)}_{pq}}{\sigma_p\sigma_q\Delta^{(1)}}$, where $\Delta^{(1)}=\Delta_{11}$, and $\Delta^{(1)}_{pq}$ is the cofactor of ρ_{pq} in Δ_{11}. Furthermore, $B=|B_{pq}|$. Hence

$$(6)\quad \int e^{-\frac{N}{2}\sum_{i,j=1}^{n} A_{ij}a_{ij}} |a_{ij}|^{\frac{N-n-2}{2}} da_1\, d[a-a_1]$$
$$= \pi^{\frac{n-1}{2}}\left(\frac{N}{2}\right)^{-\frac{N-1}{2}}\left(\frac{B}{A}\right)^{\frac{N-1}{2}}\Gamma\left(\frac{N-n}{2}\right)e^{-\frac{N}{2}\sum_{p,q=2}^{n} B_{pq}a_{pq}}|a_{pq}|^{\frac{N-n-1}{2}}d[a-a_1],$$

where $da_1 = da_{11}da_{12}\cdots da_{1n}$ and $d[a-a_1]$ is the product of the differentials of all a's in $|a_{pq}|$ $(p,q=2,3,\cdots n)$.

Now, it is clear that (6) is an identity for all values of N and the population parameters σ_i and ρ_{ij} $(i,j=1,2,\cdots m;$ $i\neq j)$, for which both sides of (6) exist. Thus, we can perform the following operations on (6):

(a). Replace N by $N+2k$.

(b). Replace σ_i by $\sigma_i\sqrt{\frac{N+2k}{N}}$, $(i=1,2,\cdots n)$

---

[4]By the reciprocal form of a determinant $|c_{ij}|$ we mean the determinant formed by replacing each element c_{ij} by the ratio $\frac{C_{ij}}{C}$ where C_{ij} is the cofactor of c_{ij} and $C=|c_{ij}|$.

(c). Replace A_{11} by $A_{11} - \frac{2\alpha}{N}$.

(d). Multiply both sides of the identity by $\frac{1}{G}$.

(e). Multiply both sides by $|a_{pq}|^{-k}$.

Accordingly, we find that the integral of the left side of (6) over all possible values of the a's is the definition of $\Phi(\alpha,k)$, which must be equal to the integral of the right side over the field of all possible values of the a's in $|a_{pq}|$ But the value of the integral of the right side can be deduced at once from (4). Hence, we finally obtain,

$$(7) \qquad \Phi(\alpha,k) = \left(\frac{N}{2}\right)^{-k} A^{\frac{N-1}{2}} A_\alpha^{-\frac{N-1}{2}-k} B_\alpha^{k} \frac{\Gamma\left(\frac{N-n}{2}+k\right)}{\Gamma\left(\frac{N-n}{2}\right)},$$

where A_α is the determinant A with A_{11} replaced by $A_{11} - \frac{2\alpha}{N}$, and B_α is the reciprocal of the cofactor of the element in the first row and first column of the reciprocal form of A_α.

That is,

$$(8) \qquad B_\alpha = \frac{A_\alpha^{n-1}}{|\bar{A}_{\alpha pq}|},$$

where $\bar{A}_{\alpha pq}$ is the cofactor of the element in the p-th row and q-th column of A_α, $(p,q=2,3,\cdots n)$. The value of $|\bar{A}_{\alpha pq}|$ can be readily found by writing

$$|\bar{A}_{\alpha pq}| = \frac{|\bar{A}_{\alpha pq}|\cdot|\bar{A}_{\alpha ij}|}{A_\alpha},$$

where $A_{\alpha ij} \equiv A_{ij}$ except for $i=j=1$ and $A_{\alpha 11} = A_{11} - \frac{2\alpha}{N}$. Increasing $|\bar{A}_{\alpha pq}|$ to an n-th order determinant by inserting, as first row and first column, an additional row and column which will not change the value of the determinant, and multiplying it by $|A_{\alpha iq}|$ we find

$$|\bar{A}_{\alpha pq}| = A_\alpha^{n-2}\left(A_{11} - \frac{2\alpha}{N}\right).$$

Therefore,

$$B_\alpha = \frac{A_\alpha}{\left(A_{11} - \frac{2\alpha}{N}\right)} \, .$$

Substituting this for B_α in (7) and using the fact that

$A_\alpha = A - \frac{2\alpha}{N}\bar{A}_{11}$, we finally obtain

$$(9) \quad \phi(\alpha,k) = \left(\frac{NA}{2\bar{A}_{11}}\right)^{\frac{N-1}{2}} \left(\frac{NA}{2\bar{A}_{11}} - \alpha\right)^{-\frac{N-1}{2}} \left(\frac{NA_{11}}{2} - \alpha\right)^{-k} \frac{\Gamma\left(\frac{N-n}{2}+k\right)}{\Gamma\left(\frac{N-n}{2}\right)} .$$

Thus, it is evident that $\phi(\alpha,k)$ exists for sufficiently small values of α. Let us write

$$\left(\frac{NA}{2\bar{A}_{11}} - \alpha\right)^{-\frac{N-1}{2}} = \left(\frac{NA_{11}}{2} - \alpha\right)^{-\frac{N-1}{2}} \left[1 - \frac{\frac{N}{2}\left(A_{11} - \frac{A}{\bar{A}_{11}}\right)}{\left(\frac{NA_{11}}{2} - \alpha\right)}\right]^{-\frac{N-1}{2}},$$

and expand the second factor on the right into a Taylor series. Substituting in (9), we have the convergent series

$$(10) \quad \begin{aligned} \phi(\alpha,k) &= \left(\frac{NA}{2\bar{A}_{11}}\right)^{\frac{N-1}{2}} \frac{\Gamma\left(\frac{N-n}{2}+k\right)}{\Gamma\left(\frac{N-n}{2}\right)} \\ &\times \sum_{i=0}^{\infty} \frac{\left(\frac{NA_{11}}{2} - \alpha\right)^{-k-\frac{N-1}{2}-i} \left(\frac{N}{2}\right)^i \left(A_{11} - \frac{A}{\bar{A}_{11}}\right)^i \Gamma\left(\frac{N-1}{2}+i\right)}{i!\,\Gamma\left(\frac{N-1}{2}\right)} . \end{aligned}$$

For the coefficient of $\frac{\alpha^h}{h!}$ in the expansion of the right side of (10) in powers of α, we find

$$(11) \quad \begin{aligned} &\left(\frac{NA_{11}}{2}\right)^{-k-h} \left(\frac{A}{\bar{A}_{11}A_{11}}\right)^{\frac{N-1}{2}} \frac{\Gamma\left(\frac{N-n}{2}+k\right)}{\Gamma\left(\frac{N-n}{2}\right)} \\ &\times \sum_{i=0}^{\infty} \frac{\left(1 - \frac{A}{A_{11}\bar{A}_{11}}\right)^i \Gamma\left(\frac{N-1}{2}+i\right) \Gamma\left(\frac{N-1}{2}+k+h+i\right)}{i!\,\Gamma\left(\frac{N-1}{2}\right) \Gamma\left(\frac{N-1}{2}+k+i\right)} \end{aligned}$$

which is the definition of $E\left[a_{11}^{h+k}(1-R^2)^k\right]$. We observe that (11) exists for all values of k and h for which $\frac{N-n}{2}+k>0$ and $\frac{N-1}{2}+h+k>0$. Placing $h=-k$ and pointing out that $\frac{A}{A_{11}\bar{A}_{11}}=1-\rho^2$, we have as the k-th moment of $1-R^2$,

$$M_k\left[(1-R^2)\right]=E\left[(1-R^2)^k\right]= \tag{12}$$

$$\frac{(1-\rho^2)^{\frac{N-1}{2}}}{\Gamma\left(\frac{N-n}{2}\right)\Gamma\left(\frac{N-1}{2}\right)}\sum_{i=0}^{\infty}\frac{\rho^{2i}\Gamma^2\left(\frac{N-1}{2}+i\right)\Gamma\left(\frac{N-n}{2}+k\right)}{i!\,\Gamma\left(\frac{N-1}{2}+k+i\right)}.$$

By using the relation

$$\frac{\Gamma\left(\frac{N-n}{2}+k\right)}{\Gamma\left(\frac{N-1}{2}+k+i\right)}=\frac{1}{\Gamma\left(\frac{n-1}{2}+i\right)}\int_0^1(1-\theta)^{\frac{N-n}{2}+k-1}\theta^{\frac{n-1}{2}+i-1}\,d\theta$$

we can write (12) in the form

$$E\left[(1-R^2)^k\right]=\frac{(1-\rho^2)^{\frac{N-1}{2}}}{\Gamma\left(\frac{N-n}{2}\right)\Gamma\left(\frac{N-1}{2}\right)} \tag{13}$$

$$\times\sum_{i=0}^{\infty}\int_0^1\frac{\rho^{2i}(1-\theta)^{\frac{N-n}{2}+k-1}\theta^{\frac{n-1}{2}+i-1}\Gamma^2\left(\frac{N-1}{2}+i\right)}{i!\,\Gamma\left(\frac{n-1}{2}+i\right)}d\theta.$$

The series in (13) is uniformly convergent in θ for $0\le\theta\le 1$ and therefore, we can interchange the order of summation and integration and write

$$E\left[(1-R^2)^k\right]=\int_0^1(1-\theta)^k\phi(\theta)\,d\theta, \tag{14}$$

where
$$\phi(\theta) = \frac{(1-\rho^2)^{\frac{N-1}{2}}(1-\theta)^{\frac{N-n}{2}-1}\,\theta^{\frac{n-1}{2}-1}}{\Gamma\left(\frac{N-1}{2}\right)\Gamma\left(\frac{N-n}{2}\right)}$$

(15)
$$\times \sum_{i=0}^{\infty} \frac{\rho^{2i}\theta^{i}\Gamma^{2}\left(\frac{N-1}{2}+i\right)}{i!\,\Gamma\left(\frac{n-1}{2}+i\right)}.$$

Thus, we have a distribution function of a variable θ such that the k-th moment of θ is identical with the k-th moment of R^2 for all positive values of k. It follows from Stekloff's[5] theory of closure that $\phi(\theta)$ must be the only continuous solution of (14), where $E\left[(1-R^2)^k\right]$ is defined as (12). Therefore, the distribution of R^2 is identical with that of θ and can be written finally as

(16)
$$df = \frac{\Gamma\left(\frac{N-1}{2}\right)}{\Gamma\left(\frac{N-n}{2}\right)\Gamma\left(\frac{n-1}{2}\right)}$$
$$\times (1-\rho^2)^{\frac{N-1}{2}}(1-R^2)^{\frac{N-n}{2}-1}(R^2)^{\frac{n-3}{2}}F\left[\tfrac{N-1}{2}, \tfrac{N-1}{2}, \tfrac{n-1}{2}, \rho^2R^2\right]d(R^2).$$

which is the distribution found by Fisher except that he uses the notation $n_1 = n-1$, the number of independent variates, and $n_1+n_2+1=N$, the sample number.

---

[5]W. Stekloff: Quelques applications nouvelles de la thêorie de fermé-ture au probleme de representation approchéc des functiones et au probleme des moments, Memoire de l'Academie Imperial des Sciences de St. Petersburg, vol. 32, no. 4, (1914).

S. S. Wilks