288

THE SIMULTANEOUS DISTRIBUTION OF CORRELATION COEFFICIENTS

By SIR RONALD A. FISHER

Division of Mathematical Statistics, C.S.I.R.O.

SUMMARY. From any sample in t variables an inter-related system of $\frac{1}{2}t(t-1)$ correlations r_{ij} can be calculated. Their simultaneous distribution depends only on the corresponding system of true correlations in the hypothetical multinormal population sampled, and in the following paper is expressed explicitly in terms of them.

1. If a sample of N be available from the simultaneous distribution of t correlated variates, direct multiplication of the N independent elements of frequency gives the expression

$$(2\pi)^{-\frac{1}{2}tN}(\sigma_1\,\sigma_2\ldots\,\sigma_t)^{-N}|\rho_{ij}|^{-\frac{1}{2}N}$$

$$\exp\left[-\tfrac{1}{2}S\left\{\frac{(x_i-\mu_i)^2}{\sigma_i^2}\,\rho^{ii}+\ldots+\frac{2(x_i-\mu_i)(x_j-\mu_j)}{\sigma_i\sigma_j}\,\rho^{ij}+\ldots\right\}\right]$$

$$dx_{11}\ldots dx_{1N}dx_{21}\ldots dx_{2N}dx_{t1}\ldots dx_{tN} \qquad \ldots \quad (1)$$

where S stands for summation over the N sets of observations; $\sigma_1, \sigma_2, \ldots, \sigma_t$ for the standard deviations in the population sampled, $|\rho_{ij}|$ for the determinant of the population correlations, and ρ^{ij} for the elements of the reciprocal of the matrix of population correlations ρ_{ij}.

From this primitive product, the simultaneous sampling distributions are to be inferred of (a) the sample means,

$$\bar{x} = \frac{1}{N}\,S(x), \qquad \ldots \quad (2)$$

of each variate, (b) of the estimates of the variances,

$$s^2 = \frac{1}{N-1} S(x-\bar{x})^2 \qquad \ldots \quad (3)$$

and finally, of the estimated coefficients of correlation

$$r_{ij} = \frac{S(x_i-\bar{x}_i)(x_j-\bar{x}_j)}{s_i s_j(N-1)} \qquad \ldots \quad (4)$$

and from this the marginal simultaneous distribution of the r_{ij}.

This program was carried out in 1915 for the case of two variates, and the distribution of the estimate r in terms of the true value ρ only, has been available in practical use for many years. For more than two variables the problem seems to have been left in a quite incomplete form.

As in the bivariate case it is sufficiently clear that the estimates s_i and r_{ij} are distributed independently of the empirical means, $\bar{x}_i$, in a distribution which does not involve the true means, μ_i, so that on integrating out all means, the marginal distribution is the same as that conditional on specific values. Equally, the $t(t-1)/2$ statistics r_{ij} will be distributed independently not only of the means, but also of the standard deviations.

The factor representing the simultaneous distribution of the $\bar{x}_i$ is easily found and removed by considering the transformation (rotation) of each variate effected by the orthogonal matrix G,

where $\qquad X'_i = GX_i$

and X_i stands for the vector $\qquad (x_{i1}, x_{i2}, \ldots, x_{iN})$

for if $\qquad X'_{i1} = S_j(x_{ij})/\sqrt{N} = \bar{x}_i\sqrt{N}$

the t variables $\qquad X'_{i1}$,

will be distributed in exactly the same simultaneous distribution as the original t variables, so that the omission of this one factor leaves a product equivalent to that appropriate to a simple sample of $(N-1)$ i.e.,

$$(2\pi)^{-\frac{1}{2}t(N-1)}(\sigma_1\sigma_2\ldots\sigma_t)^{-(N-1)}|\rho_{ij}|^{-\frac{1}{2}(N-1)}$$

$$\exp\left[-\frac{N-1}{2}\left\{\frac{s_i^2}{\sigma_i^2}\rho^{ii}+\ldots+\frac{2r_{ij}s_is_j}{\sigma_i\sigma_j}\rho^{ij}+\ldots\right\}\right]dv \qquad \ldots \quad (5)$$

in which s_i^2 has been substituted for

$$\left.\begin{array}{ll} & \dfrac{1}{N-1}\{S(x_i^2)-N\bar{x}_i^2\} \\ \text{and} \quad r_{ij}s_is_j \ \text{ for} & \dfrac{1}{N-1}\{S(x_ix_j)-N\bar{x}_i\bar{x}_j\} \end{array}\right\} \qquad \ldots \quad (6)$$

and finally, dv for the element of volume in $t(N-1)$ dimensions.

2. The second important step was taken by Wishart (1928) by expressing the element of volume in terms of the quantities

$$\xi_{ii} = S(x_i - \bar{x}_i)^2$$
$$\xi_{ij} = S(x_i - \bar{x}_i)(x_j - \bar{x}_j).$$

Regarding these quadratic forms as coordinates, Wishart showed that

$$dv = \frac{\pi^{t(2N-t-1)/4}}{\left(\frac{N-3}{2}\right)! \ldots \left(\frac{N-t-2}{2}\right)!} V^{(N-t-2)/2}\, d\xi_{11} \ldots d\xi_{tt} \qquad \ldots \quad (7)$$

where V stands for the determinant of the ξ_{ij}.

To obtain the distribution explicitly in terms of the s_i and the r_{ij}, by substituting

$$\left.\begin{aligned} \xi_{ii} &= (N-1)s_i^2 \\ \xi_{ij} &= (N-1)r_{ij}s_is_j, \end{aligned}\right\} \qquad \ldots \quad (8)$$

we may note $\qquad V = s_1^2 \ldots s_t^2 (N-1)^t |r_{ij}|$

where $|r_{ij}|$ stands for determinant of the r_{ij}, and

$$\frac{\partial(\xi_{11} \cdots \xi_{tt})}{\partial(r_{ij},\ s_i)} = 2^t (s_1 \ldots s_t)^t\, (N-1)^{t(t+1)/2},$$

giving

$$\frac{(N-1)^{\frac{1}{2}t(N-1)} \pi^{-t(t-1)/4} |\rho_{ij}|^{-\frac{1}{2}(N-1)}}{\left(\frac{N-3}{2}\right)! \ldots \left(\frac{N-t-2}{2}\right)! \quad 2^{\frac{1}{2}t(N-3)}} \left(\frac{s_1 \ldots s_t}{\sigma_1 \ldots \sigma_t}\right)^{N-2}$$

$$\exp\left[-\frac{N-1}{2}\left\{\frac{s_i^2}{\sigma_i^2}\rho^{ii} + \ldots + \frac{2r_{ij}s_is_j}{\sigma_i\sigma_j}\rho^{ij} + \ldots\right\}\right]$$

$$|r_{ij}|^{\frac{1}{2}(N-t-2)} \frac{ds_1}{\sigma_1} \ldots \frac{ds_t}{\sigma_t} \Pi(dr_{ij}). \qquad \ldots \quad (9)$$

3. So far we have the direct result of substituting the s_i and the r_{ij} in Wishart's distribution when the numerical factors are restored, which is not quite easy with the form as left by Wishart. Further simplification flows from the substitution

$$u_i = \frac{s_i}{\sigma_i}\sqrt{(N-1)\rho^{ii}} \qquad \ldots \quad (10)$$

which leads to

$$\frac{\pi^{-t(t-1)/4}\, 2^{-t(N-3)/2}}{\left(\frac{N-3}{2}\right)! \ldots \left(\frac{N-t-2}{2}\right)!} \frac{|\rho_{ij}|^{-\frac{1}{2}(N-1)}}{(\rho^{11}\rho^{22} \ldots \rho^{tt})^{\frac{1}{2}(N-1)}} |r_{ij}|^{(N-t-2)/2}\, \Pi(dr_{ij})$$

$$\exp\left\{-\frac{1}{2}\left(u_1^2 + \ldots + u_t^2 - 2\gamma_{12}u_1u_2 - \ldots\right)\right\}$$

$$(u_1 \ldots u_t)^{N-2} du_1 \ldots du_t \qquad \ldots \quad (11)$$

where $$\gamma_{ij} = -\frac{\rho^{ij}r_{ij}}{\sqrt{\rho^{ii}\rho^{jj}}}\,. \qquad \ldots \quad (12)$$

It may be observed that the range of s_i is from 0 to ∞, that u_i is s_i multiplied by a positive constant, and that

$$\int_0^\infty \ldots \int_0^\infty \exp\{-\tfrac{1}{2}(u_1^2+\ldots+u_i^2-2\gamma_{12}u_1u_2-\ldots\}(u_1\ldots u_t)^{N-2}du_1\ldots du_t \qquad \ldots \quad (13)$$

depends only on the coefficients γ_{ij}, and on N.

Writing $$F_{N-2}(\gamma_{ij})$$

for this integral, the simultaneous distribution of the values r_{ij}, in number $t(t-1)/2$, is

$$\frac{\pi^{-t(t-1)/4}\,2^{-t(N-3)/2}}{\left(\frac{N-3}{2}\right)!\ldots\left(\frac{N-t-2}{2}\right)!}\;\frac{|\rho_{ij}|^{-\frac{1}{2}(N-1)}}{(\rho^{11}\rho^{22}\ldots\rho^{tt})^{\frac{1}{2}(N-1)}}$$

$$|r_{ij}|^{(N-t-2)/2}\,F_{n-2}\,(\gamma_{ij})\,\Pi(dr_{ij}) \qquad \ldots \quad (14)$$

when the variation of $s_1, s_2, \ldots, s_t$ is eliminated.

Denoting $$\rho^*_{ij} = \frac{\rho^{ij}}{\sqrt{\rho^{ii}\rho^{jj}}}$$

and therefore $$\gamma_{ij} = -r_{ij}\rho^*_{ij}$$

the expression (14) may be written

$$\frac{\pi^{-t(t-1)/4}\;2^{-t(N-3)/2}}{\left(\frac{N-3}{2}\right)!\ldots\left(\frac{N-t-2}{2}\right)!}\,|\rho^*_{ij}|^{(N-1)/2}|r_{ij}|^{(N-t-2)/2}F_{N-2}(\gamma_{ij})\,\Pi(dr_{ij}) \ldots \quad (15)$$

Geometrically, if ρ_{ij} is the cosine of the angular distance between two points i and j on a hypersphere, then ρ^*_{ij} is the cosine with reversed sign of the dihedral angle opposite to these, or to the cosine of the angular distance between corresponding points of the polar figure.

Wilks (1944, p. 120) proposed that this elimination should be carried out by expanding the exponential as in (9) in powers of its exponent, a quadratic in s_i, and by eliminating these variates by integration. This, however, is not a feasible path for more than two variates. The case of two variates had been solved nearly 30 years earlier by quite a different approach. Wilks makes no specific proposal for more than two variables.

4. *Analytic notes.* (a) *Certain definite integrals.*

From equation (15) can be derived a number of cognate identities. If, for any values of N and t all the true correlations ρ_{ij} are zero, it follows that all γ_{ij} are zero also. Consequently the determinant

$$|\rho_{ij}^{\bullet}|$$

is replaced by unity, and the function F by

$$2^{t(N-3)/2}\left(\frac{N-3}{2}\,!\right)^t$$

then putting

$$N = t+2$$

we find the generalised volume obtained by integration with respect to r_{ij} over all possible values i.e., over the closed region within which the determinant remains positive, in the expression

$$\frac{0\,!\,\frac{1}{2}!\ldots\frac{t-1}{2}!}{\left(\frac{t-1}{2}!\right)^t}\,\pi^{t(t-1)/4}. \qquad \ldots \quad (16)$$

For any value of t the number of dimensions of the generalised volume is $\frac{1}{2}t(t-1)$; the following table gives algebraic and numerical value for moderate values of t.

TABLE. GENERALISED VOLUME OF REGIONS OF INTEGRATION FOR CORRELATION COEFFICIENTS

t	dimensions	generalised volume	numerical value
2	1	2	2.
3	3	$\pi^2/2$	4.9348
4	6	$32\pi^2/27$	11.6973
5	10	$3\pi^6/128$	22.5326
6	15	$2^{13}\pi^6/3^4 5^5$	31.114
7	21	$5\pi^{12}/3^4 2^{11}$	27.858
8	28	$2^{24}\pi^{12}/3^4 5^6 7^7$	14.877
9	36	$5^2 7\pi^{20}/3^4 2^{32}$	4.4115
10	45	$2^{40}\pi^{20}/3^{22} 5^7 7^8$	.68227

More generally if $N = t+2+k$ and $|r_{ij}| = D^2$,

$$\int \ldots \int D^k dr_{12} \ldots dr_{t-1,t} = \frac{\frac{k}{2}!\,\frac{k+1}{2}!\ldots\frac{k+t-1}{2}!}{\left(\frac{k+t-1}{2}\;!\right)^t}\,\pi^{t(t-1)/4} \qquad \ldots \quad (17)$$

giving the integrals of all powers of D, and the means of all powers over this region.

(b) *Derivatives of* $F_{N-2}(\gamma_{ij} = 0)$.

The function $F_{N-2}(\gamma_{ij})$ when γ_{ij} are all zero may be regarded as a special case of the more general function

$$\int_0^\infty {u_1}^{N-2+s_1}\; e^{-u_1^2/2}\, du_1 \int_0^\infty {u_2}^{N-2+s_2}\; e^{-u_2^2/2}\, du_2 \ldots$$

for all t variables u; this product is easily evaluated as the product of

$$2^{(N+s_1-3)/2}\; \frac{N+s_1-3}{2}!\;\; 2^{(N+s_2-3)/2}\; \frac{N+s_2-3}{2}!\; \ldots \qquad \ldots \quad (18)$$

Consequently, if a_{ij} are any positive integers, or zero, and if

$$\sum_j a_{ij} = A_i$$

so that
$$\sum_i A_i = 2 \sum_{i>j} \sum a_{ij},$$

then the differential coefficient

$$\Pi\, \frac{\partial^{a_{ij}}}{\partial \gamma_{ij}^{a_{ij}}}\, F_{N-2}(\gamma_{ij} = 0)$$

will be
$$2^{(N+A_1-2)/2}\; \frac{N+A_1-2}{2}\;!\;\; 2^{(N+A_2-2)/2}\; \frac{N+A_2-2}{2}\;! \ldots \qquad \ldots \quad (19)$$

so giving completely the differential coefficients of $F_{N-2}(\gamma_{ij} = 0)$ with respect to the $\frac{1}{2}t(t-1)$ variables γ_{ij}.

In particular the first differential coefficients are all equal, namely,

$$2^{N-2} \left(\frac{N-2}{2}\;! \right)^2 2^{(t-2)(N-3)/2} \left(\frac{N-3}{2}\;! \right)^{t-2}$$

or
$$2.\; \frac{\left(\dfrac{N-2}{2}\;! \right)^2}{\left(\dfrac{N-3}{2}\;! \right)^2}\; F$$

so that
$$\frac{\partial}{\partial(\gamma_{ij})} \log F = 2\, \frac{\left(\dfrac{N-2}{2}\;! \right)^2}{\left(\dfrac{N-3}{2}\;! \right)^2} \text{ for all } i, j. \qquad \ldots \quad (20)$$

Similarly
$$\left.\begin{aligned}
&\frac{\partial^2}{\partial \gamma_{ij}^2} \log F = (N-1)^2 - 4\, \frac{\left(\dfrac{N-2}{2}\;! \right)^4}{\left(\dfrac{N-3}{2}\;! \right)^4} \\
&\frac{\partial^2}{\partial \gamma_{ij} \partial \gamma_{ik}} \log F = 2(N-1)\, \frac{\left(\dfrac{N-2}{2}\;! \right)^2}{\left(\dfrac{N-3}{2}\;! \right)^2} - 4\, \frac{\left(\dfrac{N-2}{2}\;! \right)^4}{\left(\dfrac{N-3}{2}\;! \right)^4} \\
\text{and}\quad &\frac{\partial^2}{\partial \gamma_{ij}\, \partial \gamma_{kl}} \log F = 0
\end{aligned}\right\} \ldots \quad (21)$$

where separate terms refer to different partitions of the aggregate a_{ij}. Log F is then expressible as a multiple power series in γ_{ij}. In connection with these ratios of factorials, it may be useful to note that

$$2\left(\frac{N-2}{2}\,!\right)^2 \div \left(\frac{N-3}{2}\,!\right)^2$$

may be expanded in the form

$$N-\frac{3}{2}+\frac{1}{8(N-3/2)}-\frac{9}{128(N-3/2)^3}+\frac{153}{1024(N-3/2)^5}-\dots \qquad \dots \quad (22)$$

so that

$$\frac{\partial}{\partial\gamma_{ij}}\log F \sim N-\frac{3}{2}$$

$$\frac{\partial^2}{\partial\gamma_{ij}^2}\log F \sim N-\frac{3}{2}$$

$$\frac{\partial^2}{\partial\gamma_{ij}\partial\gamma_{ik}}\log F \sim \frac{1}{2}\left(N-\frac{7}{4}\right)$$

all being nearly equal, apart from a simple coefficient.

(c) *Geometrical interpretation.*

The expression $F_0(\gamma_{ij})$ has a rather simple geometrical interpretation. When, for example, $t = 3$ we may consider the transformation of the original coordinates u, v, w into new coordinates u', v', w', according to homogeneous linear equations

$$u' = \lambda_1 u+\mu_1 v+\nu_1 w$$
$$v' = \lambda_2 u+\mu_2 v+\nu_2 w$$
$$w' = \lambda_3 u+\mu_3 v+\nu_3 w.$$

Then

$$u'^2+v'^2+w'^2 = u^2+v^2+w^2-2\gamma_{12}uv-2\gamma_{13}uw-2\gamma_{23}vw$$

if

$$\begin{aligned} \lambda_1^2+\lambda_2^2+\lambda_3^2 &= 1, & \mu_1\nu_1+\mu_2\nu_2+\mu_3\nu_3 &= -\gamma_{23} \\ \mu_1^2+\mu_2^2+\mu_3^2 &= 1, & \lambda_1\nu_1+\lambda_2\nu_2+\lambda_3\nu_3 &= -\gamma_{13} \\ \nu_1^2+\nu_2^2+\nu_3^2 &= 1, & \lambda_1\mu_1+\lambda_2\mu_2+\lambda_3\mu_3 &= -\gamma_{12}, \end{aligned}$$

and the bounding edge $v = 0$, $w = 0$ has direction cosines

$$\lambda_1, \lambda_2, \lambda_3,$$

and so with the others, so that the cosines of the angles between pairs of such edges are $\mu_1\nu_1+\mu_2\nu_2+\mu_3\nu_3 = -\gamma_{23}$, etc.

Noting that

$$\int_0^\infty e^{-r^2/2}r^2dr = \sqrt{2}\,(\tfrac{1}{2})! = \sqrt{\pi/2}$$

or for t dimensions

$$\int_0^\infty e^{-r^2/2}r^{t-1}dr = 2^{(t-2)/2}\,\frac{t-2}{2}\,! \qquad \dots \quad (23)$$

it is seen that $F_0(\gamma_{ij})$ is for three dimensions, the area of the spherical triangle the sides of which have cosines $(-\gamma_{ij})$, multiplied by $\sqrt{\pi/2}$, and for t dimensions the generalised volume of the corresponding hyperspherical figure, multiplied by

$$2^{(t-2)/2}\,.\,\frac{t-2}{2}\,!$$

and in each case divided by

$$\partial(u', v', w')/\partial(u, v, w),$$

or by the determinant

$$\begin{vmatrix} \lambda_1 & \mu_1 & \nu_1 \\ \lambda_2 & \mu_2 & \nu_2 \\ \lambda_3 & \mu_3 & \nu_3 \end{vmatrix}$$

of which the square is

$$\begin{vmatrix} 1 & -\gamma_{12} & -\gamma_{13} \\ -\gamma_{12} & 1 & -\gamma_{23} \\ -\gamma_{13} & -\gamma_{23} & 1 \end{vmatrix}$$

If therefore this is written D^2, then in general

$$F_0(\gamma_{ij}) = 2^{(t-2)/2}\, \frac{t-2}{2}\, !\, \frac{V}{D} \qquad \ldots \quad (24)$$

where V is the generalised volume of the figure defined by $-\gamma_{ij}$, and D is the 'generalised sine' of the solid angle it subtends at the centre of the hypersphere, or the volume defined by the t unit vectors.

When t is even F_1 (γ_{ij}) can be derived from F_0 directly by differentiation with respect to chosen variables, e.g., for $t = 4$, as

$$\frac{\partial^2}{\partial\gamma_{12}\partial\gamma_{34}} F_0(\gamma_{ij}) \qquad \ldots \quad (25)$$

If t is odd, the suffix can be increased in this way only by steps of two; so when t=3

$$F_2(\gamma_{ij}) = \frac{\partial^3}{\partial\gamma_{12}\partial\gamma_{23}\partial\gamma_{31}} F_0(\gamma_{ij}) \qquad \ldots \quad (26)$$

A general form for F_1, when t is odd, is thus to be desired. In the case $t = 3$, fairly direct integration may be used to show that, if $\cos\theta_1 = \gamma_{23}$, etc., $0 \leqslant \theta \leqslant \pi$, then

$$F_1 D^2 = 1 + \frac{1}{D^2}\, \Sigma\, \frac{\theta_1}{\sin\theta_1} \{\gamma_{23}(1-\gamma_{23}^2)+\gamma_{31}(\gamma_{12}+\gamma_{23}\gamma_{31})+\gamma_{12}(\gamma_{31}+\gamma_{12}\gamma_{23})\}. \qquad \ldots \quad (27)$$

References

Fisher, R. A. (1915): The frequency distribution of the correlation coefficient in samples from an indefinitely large population. *Biom.*, **10**, 507-521.

Wishart, J. (1928): The generalised product-moment distribution in samples from a normal multivariate population. *Biom.*, **20a**, 32-52.

Wilks, S. S. (1944): *Mathematical Statistics.* Princeton University Press.

Paper received : January, 1962.