NOTE ON CORRELATIONS

By D. B. De Lury

When the value of a correlation coefficient is to be estimated from a set of N pairs of observations, (x_i, y_i), $i = 1, 2, \cdots N$, the statistic ordinarily computed is, of course, the product-moment correlation coefficient,

$$r = s_{12}/(s_1 s_2), \quad \text{where}$$

$$ns_1^2 = \sum_{i=1}^{N} (x_i - \bar{x})^2, \qquad ns_2^2 = \sum_{i=1}^{N} (y_i - \bar{y})^2, \qquad ns_{12} = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}),$$

$$N\bar{x} = \sum_{i=1}^{N} x_i, \qquad N\bar{y} = \sum_{i=1}^{N} y_i, \qquad n = N - 1.$$

However, when x and y are known to have the same population mean and variance, the precision of the estimate may be improved slightly by using the intraclass correlation coefficient,

$$r' = \frac{2\sum_{1}^{N} (x_i - \xi)(y_i - \xi)}{\sum_{1}^{N} \{(x_i - \xi)^2 + (y_i - \xi)^2\}}, \qquad 2N\xi = \sum_{1}^{N} (x_i + y_i).$$

It may be of interest to inquire into the properties of an analogous coefficient, appropriate to the case of equal variances and different means. This coefficient would naturally be chosen to be

$$u = 2s_{12}/(s_1^2 + s_2^2) = \{2s_1 s_2/(s_1^2 + s_2^2)\}r.$$

Obviously, $| u | \leq | r |$.

The probability distribution of u is easily determined, under the assumption that x and y obey a bivariate normal distribution. If σ^2 is their common variance, no restriction is introduced by taking $\sigma = 1$. Then the probability element of s_1, s_2, r, is known to be[1]

$$\frac{n^n}{\pi(n-2)!(1-\rho^2)^{n/2}} (s_1 s_2)^{n-1} e^{-\frac{n}{2(1-\rho^2)}(s_1^2 - 2\rho r s_1 s_2 + s_2^2)} (1 - r^2)^{\frac{n-3}{2}} ds_1 ds_2 dr,$$

where ρ is the correlation of x and y. From this, the distribution of u can be obtained by making the transformation

$$u = \{2s_1 s_2/(s_1^2 + s_2^2)\}r, \qquad v = 2s_1 s_2/(s_1^2 + s_2^2), \qquad w = s_1^2 + s_2^2.$$

[1] R. A. Fisher, Biometrika, Vol. 10, p. 510.

Under this transformation, the range of s_1, s_2, r, determined by the inequalities $0 \leq s_i \leq \infty$, $i = 1, 2$, $-1 \leq r \leq 1$, is mapped in a two-fold manner upon the space $-v \leq u \leq v$, $0 \leq v \leq 1$, $0 \leq w \leq \infty$. For fixed u, v ranges from u to 1 or from $-u$ to 1, according as u is positive or negative, and w runs from 0 to ∞. The probability element of u, v, w, is found to be

$$\frac{(n/2)^n}{\pi(n-2)!(1-\rho^2)^{n/2}} \frac{v}{\sqrt{1-v^2}} (v^2-u^2)^{(n-3)/2} w^{n-1} e^{-\frac{n}{2(1-\rho^2)}(1-\rho u)w} \, du\, dv\, dw,$$

and the distribution of u, obtained by integrating with respect to v and w, is

$$K(1-\rho^2)^{n/2}(1-\rho u)^{-n}(1-u^2)^{(n-2)/2} du, \qquad K = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n}{2}\right)}.$$

If $\rho = 0$, the distribution of u is identical with that of r, the product-moment correlation coefficient (for $\rho = 0$), in samples of $(N + 1)$ pairs of observations. Therefore, to test the hypothesis of independence, using the coefficient u, the methods and tables appropriate to testing the same hypothesis, using the coefficient r, are available. The precision gained by using u rather than r is equivalent to that supplied by another pair of observations.

In the general case, the transformation introduced by R. A. Fisher,[2]

$$u = \tanh z, \quad \rho = \tanh \zeta,$$

leads to the distribution element[3]

$$K \operatorname{sech}^n(z - \zeta) dz.$$

This distribution is invariant in form under varying ζ, and is effectively normal for samples of any size. In all cases, z is an unbiased estimate of ζ.

The variance of z can be obtained by the following device. Denote by $I(2p, n)$ the 2 p-th moment of z about the mean,

$$I(2p, n) = K \int_{-\infty}^{\infty} x^{2p} \operatorname{sech}^n x \, dx.$$

Integration by parts gives the recurrence formula,

$$I(2p, n) = \frac{n^2}{(2p+1)(2p+2)} \{I(2p+2, n) - I(2p+2, n+2)\}, \qquad p \geq 0.$$

---

[2] Metron, Vol. 1, N. 4, p. 7.

[3] The distributions of u and z for $n = 1$ have been given by R. A. Fisher, Metron, Vol. 1, N. 4, p. 8.

From this follows at once the relation

$$I(2p+2, n+2) = I(2p+2, 1) - (2p+1)(2p+2)\left\{\frac{I(2p,1)}{1^2} + \frac{I(2p,3)}{3^2} + \cdots + \frac{I(2p,n)}{n^2}\right\}, \; n \text{ odd},$$

$$= I(2p+2, 2) - (2p+1)(2p+2)\left\{\frac{I(2p,2)}{2^2} + \frac{I(2p,4)}{4^2} + \cdots + \frac{I(2p,n)}{n^2}\right\}, \; n \text{ even}.$$

The values of $I(2p+2, 1)$ and $I(2p+2, 2)$ can be found without evaluating the integrals, by letting $n \to \infty$. It can be shown that $I(2p, n) = O(n^{-p})$, and hence $\lim_{n\to\infty} I(2p, n) = 0$ for $p > 0$. We obtain

$$I(2p+2, 1) = (2p+1)(2p+2)\left\{\frac{I(2p,1)}{1^2} + \frac{I(2p,3)}{3^2} + \frac{I(2p,5)}{5^2} + \cdots\right\},$$

$$I(2p+2, 2) = (2p+1)(2p+2)\left\{\frac{I(2p,2)}{2^2} + \frac{I(2p,4)}{4^2} + \frac{I(2p,6)}{6^2} + \cdots\right\}.$$

Hence, for all values of n and p, (replacing $n+2$ by n),

$$I(2p+2, n) = (2p+1)(2p+2)\left\{\frac{I(2p,n)}{n^2} + \frac{I(2p,n+2)}{(n+2)^2} + \frac{I(2p,n+4)}{(n+4)^2} + \cdots\right\}.$$

Setting $p = 0$ to get the variance,

$$\mu_2 = I(2, n) = 2\left\{\frac{1}{n^2} + \frac{1}{(n+2)^2} + \frac{1}{(n+4)^2} + \cdots\right\}.$$

Therefore, making use of the fact that $\int_m^\infty x^{-2}\,dx < \sum_{i=m}^{\infty} i^{-2} < \int_{m-1}^\infty x^{-2}\,dx$, we find that

$$1/n < \mu_2 < 1/(n-2),$$

and from the numerical values of μ_2 for small values of n, it appears that the approximation $\mu_2 \sim 1/(n-1)$ is satisfactory in all cases.

In the same way, it can be shown that

$$3/n^2 < \mu_4 < 3/(n-2)^2.$$

Thus the method of transforming correlations to test for significance, used by R. A. Fisher in connection with both interclass and intraclass correlations, is available here also, and is, in fact, slightly simpler, owing to the absence of bias.

The coefficient u can, of course, be used in all situations where the intraclass coefficient is appropriate, (when the number of observations in each class is two), and conceivably in a small class of other cases as well. The test of significance is simpler using u instead of r', and the loss of precision is negligible.

UNIVERSITY OF TORONTO.