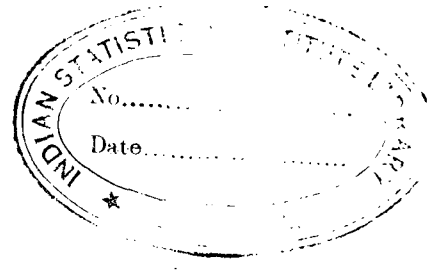

A note on non-normal correlation

By J. B. S. HALDANE

The product-moment correlation ρ is frequently estimated for two variates which are not normally distributed. There are, however, no general expressions for the effect of this non-normal distribution on the precision of the estimate of ρ. They may be obtained in one special case which is of biological importance. Suppose X and Y are two correlated variates. Then if

$$X = a+\frac{\sigma}{\sqrt{2}}[(1+\rho)^{\frac{1}{2}}x+(1-\rho)^{\frac{1}{2}}y], \quad Y = b+\frac{\tau}{\sqrt{2}}[(1+\rho)^{\frac{1}{2}}x-(1-\rho)^{\frac{1}{2}}y],$$

and further if $\bar{x} = \bar{y} = 0, \bar{x}^2 = \bar{y}^2 = 1$, and x and y are independent, the variance of X is σ^2, that of Y is τ^2, and their covariance is $\rho\sigma\tau$, regardless of the distributions of x and y. Hence the correlation of X and Y is ρ. If x and y are normally distributed the correlation is of course normal. Now in biological statistics X and Y may be measurements of two organs in the same individual, or of their logarithms. x depends on the sum of causes which affect X and Y alike, y on the sum of causes which affect them oppositely. For example, in any series of specimens, not all of which are fully grown, x will increase with age up to a certain point; and in a population containing a minority of juvenile members the distribution of x will probably be negatively skew. But y may be quite independent of age if the variability of the organs measured is uncorrelated with age, and may well be normally distributed when x is not.

Let κ_{rs} be the cumulants of the joint distribution of x and y. Then since they are independent, $\kappa_{rs} = 0$ unless r or $s = 0$, $\kappa_{10} = \kappa_{01} = 0$, $\kappa_{20} = \kappa_{02} = 1$; and let $\kappa_{30} = \gamma_1$, $\kappa_{03} = \gamma_1'$, $\kappa_{40} = \gamma_2$, $\kappa_{04} = \gamma_2'$, etc., these being measures of the deviations from normality of the distributions of x and y.

Our estimate of ρ on a sample of n members is thus

$$r = \frac{n\Sigma X_r Y_r - \Sigma X_r \Sigma Y_r}{[n\Sigma X_r^2-(\Sigma X_r)^2]^{\frac{1}{2}}[n\Sigma Y_r^2-(\Sigma Y_r)^2]^{\frac{1}{2}}}$$

$$= [n(1+\rho)\Sigma x_r^2 - n(1-\rho)\Sigma y_r^2 - (1+\rho)(\Sigma x_r)^2 + (1-\rho)(\Sigma y_r)^2]$$
$$\times [n(1+\rho)\Sigma x_r^2 + n(1-\rho)\Sigma y_r^2 + 2n(1-\rho^2)^{\frac{1}{2}}\Sigma x_r y_r - (1+\rho)(\Sigma x_r)^2 - (1-\rho)(\Sigma y_r)^2 - 2(1-\rho^2)^{\frac{1}{2}}\Sigma x_r \Sigma y_r]^{-\frac{1}{2}}$$
$$\times [n(1+\rho)\Sigma x_r^2 + n(1-\rho)\Sigma y_r^2 - 2n(1-\rho^2)^{\frac{1}{2}}\Sigma x_r y_r - (1+\rho)(\Sigma x_r)^2 - (1+\rho)(\Sigma y_r)^2 + 2(1-\rho^2)^{\frac{1}{2}}\Sigma x_r \Sigma y_r]^{-\frac{1}{2}}.$$

So
$$r^2 = \frac{[(1+\rho)\{n\Sigma x_r^2-(\Sigma x_r)^2\}-(1-\rho)\{n\Sigma y_r^2-(\Sigma y_r)^2\}]^2}{[(1+\rho)\{n\Sigma x_r^2-(\Sigma x_r)^2\}+(1-\rho)\{n\Sigma y_r^2-(\Sigma y_r)^2\}]^2-4(1-\rho^2)(n\Sigma x_r y_r-\Sigma x_r \Sigma y_r)^2}$$

$$= \frac{[(1+\rho)k_{20}-(1-\rho)k_{02}]^2}{[(1+\rho)k_{20}+(1-\rho)k_{02}]^2-4(1-\rho^2)k_{11}^2}, \qquad (1)$$

where k_{rs} is the unbiased estimate of κ_{rs} from the moments of the variates in the sample. For example, $k_{20} = \dfrac{n\Sigma x_r^2 - (\Sigma x_r)^2}{n(n-1)}$. We can now ask how the mean value of r^2 will be affected by deviations of the distributions of x and y from normality. $\overline{k_{20}^2}$ exceeds $(\overline{k_{20}})^2$, or unity, by the sampling variance of k_{20} which is $\dfrac{2\kappa_{20}^2}{n-1} + \dfrac{\kappa_{40}}{n}$, or $\dfrac{2}{n-1} + \dfrac{\gamma_2}{n}$. The effect of non-normality in the distribution of x is therefore to increase the mean value of k_{20}^2 by γ_2/n. Similarly, $\overline{k_{02}^2}$ is increased by γ_2'/n. $\overline{k_{20}k_{02}}$ is not increased, since x and y are independent. k_{11}^2 does not include terms with zero suffixes, so it is also unaltered. In fact, both numerator and denominator of (1) are increased by $n^{-1}[(1+\rho)^2\gamma_2 + (1-\rho)^2\gamma_2']$.

We cannot calculate the variance of r directly from (1) since $\bar{r}$ differs from ρ by a quantity of order n^{-1}. But since both the numerator and denominator are increased by

$$(1+\rho)^2\frac{\kappa_{40}}{n} + (1-\rho)^2\frac{\kappa_{04}}{n} \quad \text{or} \quad n^{-1}[(1+\rho)^2\gamma_2 + (1-\rho)^2\gamma_2']$$

above the values found when the distributions of x and y are normal, we have in the normal case

$$r_0^2 = \frac{4\rho^2 + 4n^{-1}P}{4 + 4n^{-1}Q},$$

where P and Q are independent of n to order n^{-2}, and in general

$$\overline{r^2} = \frac{4\rho^2 + n^{-1}[4P + (1+\rho)^2\gamma_2 + (1-\rho)^2\gamma_2']}{4 + n^{-1}[4Q + (1+\rho)^2\gamma_2 + (1-\rho)^2\gamma_2']}.$$

So
$$\overline{r^2} - \overline{r_0^2} = \frac{1-\rho^2}{4n}[(1+\rho)^2\gamma_2 + (1-\rho)^2\gamma_2'] + O(n^{-2}).$$

The variance of r is therefore increased by this quantity. The precision of the estimate of ρ does not therefore depend on the skewness of the distributions of x and y, provided they are mesokurtic. And since

$$\gamma_1 = \sqrt{2}(1+\rho)^{-\frac{3}{2}}[\gamma_1(X) + \gamma_1(Y)], \quad \gamma_1' = \sqrt{2}(1-\rho)^{-\frac{3}{2}}[\gamma_1(X) - \gamma_1(Y)],$$

it follows that skew distribution of X and Y will not affect the precision of r. On the other hand, the distributions of X and Y have the same value of γ_2 or $\beta_2 - 3$, namely,

$$\Gamma_2 = \tfrac{1}{2}[(1+\rho)^2\gamma_1 + (1-\rho)^2\gamma_2].$$

Hence
$$\operatorname{var}(r) = \frac{1-\rho^2}{n}(1-\rho^2+\Gamma_2) + O(n^{-2}). \tag{2}$$

If we employ Fisher's transformation $z = \frac{1}{2}\log\left(\dfrac{1+r}{1-r}\right)$, we find

$$\operatorname{var}(z) = n^{-1}\left(1 + \frac{\Gamma_2}{1-\rho^2}\right) + O(n^{-2}). \tag{3}$$

The variance of z is thus no longer almost independent of ρ. But the precision of r is increased if the distributions of X and Y are platykurtic, and decreased if they are leptokurtic. Clearly, however, (2) and (3) are inapplicable when $|\rho|$ is near unity, terms of at least order n^{-2} being required.

On empirical grounds, E. S. Pearson (1931, 1932) stated that 'the normal bivariate surface may be distorted and mutilated to a remarkable degree without affecting the frequency distribution of r'. This would seem to be true when $|\rho|$ is not near unity. It is also true for 'mutilations' which affect skewness without doing a great deal to kurtosis. However, when correlation is high it would seem that a relatively slight change in kurtosis may have a large effect on the variance of r.

It is possible that the formula (1) might serve as a basis for a new development of the theory of the distribution of r in the normal case, and further information could certainly be obtained from it concerning the more general case here considered. In the most general case x and y, though they have a zero coefficient of correlation, are not independent, so such cumulants as κ_{22} would not in general be zero, and it is doubtful whether the method would be of value. On the other hand, if the distributions of X and Y, though having different values of β_1, have insignificantly different values of β_2, equation (2) or (3) may be used with some confidence.

I have to thank Mr K. A. Kermack for useful criticism.

REFERENCE

PEARSON, E. S. (1931, 1932). The test of significance for the correlation coefficient. *J. Amer. Statist. Soc.* **26**, 128; **28**, 424.