THE POINT BISERIAL COEFFICIENT OF CORRELATION

By Joseph Lev

New York State Department of Civil Service

The product moment coefficient of correlation between a continuous variate y and a variate x which takes the values 1 and 0 only, is known in psychological statistics as the point biserial coefficient of correlation. Let y_i, $i = 1, \cdots, n$, be observations on y; y_{1i}, $i = 1, \cdots, n_1$, be y values which are paired with the value $x = 1$; y_{0i}, $i = 1, \cdots, n_0$, be values paired with $x = 0$; $\bar{y}$, $\bar{y}_1$, and $\bar{y}_0$ be the corresponding means; and $n = n_1 + n_0$. Then the point biserial coefficient of correlation may be written

$$r = \frac{\sqrt{\frac{n_1 n_0}{n}}\,(\bar{y}_1 - \bar{y}_0)}{\left[\sum_{i=0}^{1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2\right]^{\frac{1}{2}}}. \tag{1}$$

The distribution of r is readily obtained when the y_i, $i = 1, \cdots, n$, are distributed as

$$\frac{1}{\sqrt{2\pi}\,\sigma\sqrt{1 - \rho^2}} \exp\left[\frac{-1}{2\sigma^2(1 - \rho^2)}(y_i - \alpha - \rho\sigma z_i)^2\right] \tag{2}$$

where

$$z_i = \frac{x_i - \bar{x}}{\sigma_x} = \begin{cases} \sqrt{\frac{n_0}{n_1}}, & i = 1, 2, \cdots, n_1, \\ -\sqrt{\frac{n_1}{n_0}}, & i = n_1 + 1, n_1 + 2, \cdots, n, \end{cases}$$

σ^2 is the variance of the y_i about the common mean α, and ρ is the parameter which represents the correlation between the y_i and the x_i. It is easy to verify that the statistic in (1) is a maximum likelihood estimate of ρ.

It will be convenient to express the two population means in (2) as μ_1 and μ_0 so that

$$\begin{aligned} \mu_1 &= \alpha + \rho\sigma\sqrt{\frac{n_0}{n_1}}, \\ \mu_0 &= \alpha - \rho\sigma\sqrt{\frac{n_1}{n_0}}. \end{aligned} \tag{3}$$

Hence

$$\rho = \sqrt{\frac{n_1 n_0}{n}}\,\frac{\mu_1 - \mu_0}{\sigma}. \tag{4}$$

Now write

$$(5) \qquad t = \frac{\sqrt{\frac{n_1 n_0}{n}}\,(\bar{y}_1 - \bar{y}_0)\sqrt{n-2}}{\left[\sum_{i=0}^{1}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2\right]^{\frac{1}{2}}} = \frac{\sqrt{n-2}\,r}{\sqrt{1-r^2}},$$

where r is obtained from (1).

Using (5) we may write t as

$$t = \frac{\dfrac{(\bar{y}_1 - \bar{y}_0) - (\mu_1 - \mu_0)}{\sqrt{\frac{n}{n_1 n_0}}\,\sigma\sqrt{1-\rho^2}} + \dfrac{\mu_1 - \mu_0}{\sqrt{\frac{n}{n_1 n_0}}\,\sigma\sqrt{1-\rho^2}}}{\dfrac{\left[\dfrac{\sum_{i=0}^{1}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}{n-2}\right]^{\frac{1}{2}}}{\sigma\sqrt{1-\rho^2}}}.$$

Therefore t has non-central t distribution [1] with

$$(6) \qquad \delta = \frac{\mu_1 - \mu_0}{\sqrt{\frac{n}{n_1 n_0}}\,\sigma\sqrt{1-\rho^2}} = \sqrt{n}\,\frac{\rho}{\sqrt{1-\rho^2}}.$$

The methods and tables given in [1] may be used to calculate tests of significance and confidence limits for ρ.

When $\rho = 0$, t has Student's distribution, and the statistic $t = \sqrt{n-2}\,r/\sqrt{1-r^2}$ may be used to test the hypothesis, $\rho = 0$, by means of the t tables with $n - 2$ degrees of freedom. The non-central t distribution then determines the power function of this test.

Table IV of [1] can be used to calculate confidence limits for ρ. If the confidence interval is to be based on equal tails of the distribution choose a confidence coefficient $1 - 2\epsilon$. Then compute $\delta(f, t_0, \epsilon)$ and $\delta(f, t_0, 1 - \epsilon)$, where $f = n - 2$, and $t_0 = \sqrt{n-2}\,r/\sqrt{1-r^2}$.

A lower limit for ρ is given by

$$\frac{\delta(f, t_0, \epsilon)}{[n + \delta^2(f, t_0, \epsilon)]^{\frac{1}{2}}},$$

and an upper limit by

$$\frac{\delta(f, t_0, 1-\epsilon)}{[n + \delta^2(f, t_0, 1-\epsilon)]^{\frac{1}{2}}}.$$

REFERENCE

[1] N. L. Johnson and B. L. Welch, "Applications of non-central t-distribution," *Biometrika*, Vol. 31 (1940), pp. 362–389.