Biserial Correlation

Biserial correlation coefficients are measures of bivariate association that arise when one of the observed variables is on a measurement scale and the other variable takes on two values. There are several biserial coefficients, with appropriate choices based on the nature of the underlying bivariate population. Two common forms are the Pearson biserial correlation (hereafter referred to as the biserial correlation) and the point biserial correlation.

Pearson [9] developed the biserial correlation to estimate the product moment **correlation** ρ_{YZ} between two measurements Y and Z using data where Z is not directly observed. Instead of Z, data are collected on a categorical variable X which takes on the values $X = 1$ if Z exceeds a threshold level, and $X = 0$ otherwise. In many applications, the latent variable Z is conceptual rather than observable. The actual values used to code X do not matter, provided the larger value of X is obtained when Z exceeds the threshold. The point biserial correlation is the product moment correlation ρ_{YX} between Y and X.

We use data adapted from the study by Karelitz et al. [7] of 38 infants to illustrate ideas. Table 1 gives a listing of the data. The categorical variable X corresponds to whether the child's speech developmental level at age three is high ($X = 1$) or low ($X = 0$). The child's IQ score at age three is Y. The biserial correlation ρ_{YZ} is a reasonable measure of association when X can be viewed as a surrogate for an underlying continuum Z of speech levels. The point biserial correlation ρ_{YX} might be considered when the scientist is uninterested in the relationship between IQ and the underlying Z scale, or cannot justify the existence of such a scale.

The remainder of this article discusses methods for estimating the point biserial and the biserial correlation. Other forms of biserial correlation are briefly mentioned.

The Point Biserial Correlation

Suppose that a sample $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$ is selected from the (Y, X) population. Let s_{YX} be the sample covariance between the $y_i s$ and the $x_i s$, and let s_Y^2 and s_X^2 be the sample variances of the $y_i s$ and the $x_i s$, respectively. The population correlation ρ_{YX} is estimated consistently by the sample point biserial correlation

$$r_{YX} = \frac{s_{YX}}{s_Y s_X} = \frac{(\overline{y}_1 - \overline{y}_0)}{s_Y}[\hat{p}(1 - \hat{p})]^{1/2}, \qquad (1)$$

where $\overline{y}_1$ and $\overline{y}_0$ are the average y values from sampled pairs having $x_i = 1$ and $x_i = 0$, respectively, and $\hat{p}$ is the observed proportion of pairs with $x_i = 1$.

The sampling distribution of r_{YX} is known only for certain models. Tate [12] derived the distribution of $T = (n - 2)^{1/2} r_{YX}/(1 - r_{YX}^2)^{1/2}$ under the assumption that the conditional distributions of Y given $X = 1$ and given $X = 0$ are normal with identical variances. Tate [12] noted that T is equal to the usual two-sample Student's t statistic for comparing the means of the y samples having $x_i = 1$ and $x_i = 0$. The hypothesis $\rho_{YX} = 0$ is usually tested using this standard t test. Tate's [12] results are more complex for testing nonzero values of ρ_{YX}. In large samples, hypothesis tests and confidence intervals for ρ_{YX} can be based on a normal approximation to r_{YX} with mean ρ_{YX} and estimated variance

$$\widehat{\text{var}}(r_{YX}) = \frac{(1 - r_{YX}^2)^2}{n}\left\{1 - 1.5 r_{YX}^2 + \frac{r_{YX}^2}{4\hat{p}(1 - \hat{p})}\right\}. \qquad (2)$$

Das Gupta [5] generalized Tate's [12] results to nonnormal populations.

For the IQ data, the distributions of the IQ scores for the samples with $X = 0$ and $X = 1$ are slightly skewed to the right and have similar spreads. The mean IQ scores in the two samples are $\overline{y}_1 = 2779/22 = 126.318$ and $\overline{y}_0 = 1676/16 = 104.750$. With $\hat{p} = 22/38 = 0.579$ and $s_Y = 19.383$, we obtain

Table 1 IQ data for a sample of 38 children: X = speech developmental level (0 = low; 1 = high) and Y = IQ score

$X = 0$	Y:	87	90	94	94	97	103	103	104	106	108	109	109
		109	112	119	132								
$X = 1$	Y:	100	103	103	106	112	113	114	114	118	119	120	120
		124	133	135	135	136	141	155	157	159	162		

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a10007

$r_{YX} = 0.557$, $\widehat{\mathrm{sd}}(r_{YX}) = 0.103$, and $T = 4.024$ under Tate's assumptions.

A limiting feature of the point biserial correlation is that the range of ρ_{YX} is smaller than the usual reference range of -1.0 to 1.0. For example, the magnitude of ρ_{YX} cannot exceed 0.798 when Y is normally distributed. This restriction can lead to misinterpreting the strength of the sample correlation. Shih & Huang [10] examined this problem in a general setting, and offer a useful method to calibrate point biserial correlations.

Pearson's Biserial Correlation and Generalizations

Suppose that X is obtained by **categorizing a continuous variable** Z with $X = 1$ if $Z > \omega$ and $X = 0$ otherwise, where ω is a fixed but possibly unknown threshold. Let $f(t)$ and $F(t)$ be the pdf for Z and the cdf for Z, respectively. We assume without loss of generality that $\mathrm{E}(Z) = 0$ and $\mathrm{var}(Z) = 1$. The threshold ω is the upper pth percentile of Z, i.e. $\omega = F^{-1}(1 - p)$, where

$$p = \Pr(X = 1) = \Pr(Z > \omega) = 1 - F(\omega). \tag{3}$$

If the regression of Y on Z is linear, then the biserial correlation and the point biserial correlation are related by

$$\rho_{YZ} = \rho_{YX}\frac{[p(1-p)]^{1/2}}{\lambda(\omega, F)}, \tag{4}$$

where

$$\lambda(\omega, F) = \mathrm{E}(XZ) = \int_{\omega}^{\infty} tf(t)\,\mathrm{d}t = \int_{\omega}^{\infty} t\,\mathrm{d}F(t). \tag{5}$$

The linear regression assumption is satisfied when (Y, Z) has a bivariate normal distribution, a common assumption, but holds for other elliptically symmetrical bivariate distributions as well.

Eq. (4) provides a way to estimate the biserial correlation from a sample of $(y_i, x_i)s$ when the cdf of Z is known. Bedrick [2] proposed a simple method-of-moments estimator,

$$\tilde{r}_{YZ} = r_{YX}\frac{[\hat{p}(1-\hat{p})]^{1/2}}{\lambda(\hat{\omega}, F)}, \tag{6}$$

where $\hat{\omega} = F^{-1}(1 - \hat{p})$ is the estimated threshold based on the proportion $\hat{p}$ of sampled pairs with $x_i = 1$. If Z is normally distributed, then (6) is Pearson's biserial estimator

$$r_{\mathrm{Pb}} = \frac{r_{YX}}{\phi(\hat{\omega})}[\hat{p}(1-\hat{p})]^{1/2} = \frac{(\overline{y}_1 - \overline{y}_0)}{s_Y\phi(\hat{\omega})}\hat{p}(1-\hat{p}), \tag{7}$$

where $\phi(t)$ is the standard normal pdf.

Bedrick [2] showed that the asymptotic distribution of $\tilde{r}_{YZ}$ is normal with mean ρ_{YZ} and gave an expression for the large sample $\mathrm{var}(\tilde{r}_{YZ})$. In earlier work, Soper [11] gave an estimator for $\mathrm{var}(r_{\mathrm{Pb}})$ when (Y, Z) is normal:

$$\begin{aligned}\widehat{\mathrm{var}}(r_{\mathrm{Pb}}) = \frac{1}{n}\Bigg\{r_{\mathrm{Pb}}^4 &+ \frac{r_{\mathrm{Pb}}^2}{\phi^2(\hat{\omega})}[\hat{p}(1-\hat{p})\hat{\omega}^2 \\ &+ (2\hat{p} - 1)\hat{\omega}\phi(\hat{\omega}) - 2.5\phi^2(\hat{\omega})] \\ &+ \frac{\hat{p}(1-\hat{p})}{\phi^2(\hat{\omega})}\Bigg\}.\end{aligned} \tag{8}$$

Unlike the point biserial estimator, the magnitudes of r_{Pb} and $\tilde{r}_{YZ}$ can exceed 1.0. For the IQ data, $r_{\mathrm{Pb}} = 0.694$ and $\widehat{\mathrm{sd}}(r_{\mathrm{Pb}}) = 0.135$.

Brogden [3] and Lord [8] generalized Pearson's estimator by relaxing the assumption that the distribution of Z is known. Bedrick [1, 2] gave a detailed study of Brogden and Lord's estimators. Cureton [4] and Glass [6] proposed versions of Brogden's estimator that are based on the ranks of the y sample.

As a final point, note that a **maximum likelihood estimator** (MLE) of ρ_{YZ} can be computed iteratively whenever a joint distribution for (Y, Z) can be specified. Tate [13] proposed the MLE of ρ_{YZ} as an alternative to Pearson's biserial estimator with bivariate normal populations. Although MLEs are fully efficient, Bedrick's [1, 2] results show that the asymptotic variances of Lord's estimator and the MLE are often close in normal and nonnormal populations.

References

[1] Bedrick, E.J. (1990). On the large sample distributions of modified sample biserial correlation coefficients, *Psychometrika* **55**, 217–228.

[2] Bedrick, E.J. (1992). A comparison of modified and generalized sample biserial correlation estimators, *Psychometrika* **57**, 183–201.

DOI: 10.1002/0470011815.b2a10007

[3] Brogden, H.E. (1949). A new coefficient: application to biserial correlation and to estimation of selective inefficiency, *Psychometrika* **14**, 169–182.

[4] Cureton, E.E. (1956). Rank-biserial correlation, *Psychometrika* **21**, 287–290.

[5] Das Gupta, S. (1960). Point biserial correlation and its generalization, *Psychometrika* **25**, 393–408.

[6] Glass, G.W. (1966). Note on rank biserial correlation, *Educational and Psychological Measurement* **26**, 623–631.

[7] Karelitz, S., Fisichelli, V.R., Costa, J., Karelitz, R. & Rosenfeld, L. (1964). Relation of crying activity in early infancy to speech and intellectual development at age three years, *Child Development* **35**, 769–777.

[8] Lord, F.M. (1963). Biserial estimates of correlation, *Psychometrika* **28**, 81–85.

[9] Pearson, K. (1909). On a new method of determining the correlation between a measured character A and a character B, *Biometrika* **7**, 96–105.

[10] Shih, W.J. & Huang, W.-H. (1992). Evaluating correlation with proper bounds, *Biometrics* **48**, 1207–1213.

[11] Soper, H.E. (1914). On the probable error for the biserial expression for the correlation coefficient, *Biometrika* **10**, 384–390.

[12] Tate, R.F. (1954). Correlation between a discrete and a continuous variable. Point-biserial correlation, *Annals of Mathematical Statistics* **25**, 603–607.

[13] Tate, R.F. (1955). The theory of correlation between two continuous variables when one is dichotomized, *Biometrika* **42**, 205–216.

(*See also* **Association, Measures of**; **Pearson, Karl**)

EDWARD J. BEDRICK

Encyclopedia of Biostatistics, Online © 2005 John Wiley & Sons, Ltd.

This article was published in the *Encyclopedia of Biostatistics* in 2005 by John Wiley & Sons, Ltd.
DOI: 10.1002/0470011815.b2a10007