Biserial Correlation Coefficients

As you know, the Pearson *r* is simply the standardized slope for an ordinary least squares linear model predicting Y from X. When thinking about *r*, most researchers think of both X and Y as being continuous variables, and often they think of at least one them as being normally distributed. The Pearson r is still useful, however, when one or both of the variables are not continuous. For example, if both are rank variables, then the Pearson *r* becomes the **Spearman rho** and is a measure of monotonicity rather than of strict linearity. If both are dichotomous, then Pearson *r* becomes the **phi coefficient** most often associated with 2 x 2 contingency tables.

What if Y is continuous and X is dichotomous? In this case, Pearson *r*, computed in the usual fashion, becomes the **point-biserial correlation coefficient**. When we were studying independent samples *t* tests, I demonstrated to you that testing the significance of the null that two population means are identical is, when using the pooled variances test, mathematically identical to testing the null hypothesis that the point biserial correlation between group membership and Y is zero.

What if the latent variables underlying X and Y are both normally distributed, but we have measured the one continuously and the other dichotomously? In this case, the point biserial *r* is likely to underestimate the value of the ρ between the two latent variables. A better estimate of that ρ can be obtained with the **biserial correlation coefficient**.

If you have statistical software that can compute Pearson *r* but not the biserial correlation coefficient, the easiest way to get the biserial coefficient is to compute the point-biserial and then transform it. Howell (1977, page 287) provided this transformation: $r_b = \frac{r_{pb}\sqrt{p_1 p_2}}{y}$, where r_{pb} is the point biserial, p_1 is the proportion of cases that are at Level 1 of the dichotomous variable, p_2 is the proportion of cases that are at Level 2 of the dichotomous variable, and *y* is the probability density (height) of the normal curve at the point where p_1 of the area under the curve is on the one side and p_2 on the other side.

To illustrate, I shall use the data described here. Variable CP_Engl is whether the student was enrolled in college prep English (1) or not (0). We are willing to assume that the underlying latent variable is Verbal Aptitude and that this latent variable is normally distributed. We are interested in how well verbal aptitude is correlated with IQ. We do have these student's IQ scores. The point biserial correlation between CP_Engl and IQ is .119, and 15.9% of the students are enrolled in the college prep English class. From our standard normal curve table we find that the value of *z* marking off the upper 15.9% of the distribution is 1.00 and the height of the curve (*y* in Howell's tables) is .242. The biserial correlation is $\frac{.119\sqrt{.159(.841)}}{.242} = .180$. [PS – the low value of this coefficient leads me to opine that placement in college prep classes was not a good proxy for verbal aptitude.]

Suppose that our observed measure of IQ is also dichotomous – the school would not give us the actual IQ scores, but we were able to determine, from public

records, for each student, whether or not the student had been enrolled in a “special education” class. We are willing to treat such enrollment as a proxy for IQ and we are confident that the underlying latent variable is normally distributed. The simple correlation between our two dichotomous variables is a **phi coefficient**, but for a better estimate of the correlation between the two latent variables we need a **tetrachoric correlation**. Don’t even think about obtaining this by hand. Find a stats program that will do it for you.

Suppose I have ranked from 1 (lowest performance) to 10 (highest performance) the term papers written by my students in Psychoscatology. I wish to see how well their performance on these term papers is correlated with previous academic performance. I am unable to get their grade point averages, but I am able to identify which students did or did not received academic honors (such as dean’s list) during the last grading period. My observed Y variable is a rank variable and my observed X variable is dichotomous, but I am willing to assume that the latent variable underlying X is also a rank variable. An appropriate statistic to estimate the (Spearman) correlation between the two underlying rank latent variables is the rank biserial correlation (Glass, 1966). See my document, Nonparametric Effect Size Estimators, for details on how to compute the rank biserial correlation.

Karl L. Wuensch
16-June-2015