Point-biserial correlation coefficient

From Wikipedia, the free encyclopedia

The **point biserial correlation coefficient** (r_{pb}) is a correlation coefficient used when one variable (e.g. Y) is dichotomous; Y can either be "naturally" dichotomous, like gender, or an artificially dichotomized variable. In most situations it is not advisable to dichotomize variables artificially. When you artificially dichotomize a variable the new dichotomous variable may be conceptualized as having an underlying continuity. If this is the case, a biserial correlation would be the more appropriate calculation.

The point-biserial correlation is mathematically equivalent to the Pearson (product moment) correlation, that is, if we have one continuously measured variable X and a dichotomous variable Y, $r_{XY} = r_{pb}$. This can be shown by assigning two distinct numerical values to the dichotomous variable.

To calculate r_{pb}, assume that the dichotomous variable Y has the two values 0 and 1. If we divide the data set into two groups, group 1 which received the value "1" on Y and group 2 which received the value "0" on Y, then the point-biserial correlation coefficient is calculated as follows:

$$r_{pb} = \frac{M_1 - M_0}{s_n}\sqrt{\frac{n_1 n_0}{n^2}},$$

where s_n is the standard deviation used when you have data for every member of the population:

$$s_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2},$$

M_1 being the mean value on the continuous variable X for all data points in group 1, and M_0 the mean value on the continuous variable X for all data points in group 2. Further, n_1 is the number of data points in group 1, n_0 is the number of data points in group 2 and n is the total sample size. This formula is a computational formula that has been derived from the formula for r_{XY} in order to reduce steps in the calculation; it is easier to compute than r_{XY}.

There is an equivalent formula that uses s_{n-1}:

$$r_{pb} = \frac{M_1 - M_0}{s_{n-1}}\sqrt{\frac{n_1 n_0}{n(n-1)}},$$

where s_{n-1} is the standard deviation used when you only have data for a sample of the population:

$$s_{n-1} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

It's important to note that this is merely an equivalent formula. It is not a formula for use in the case where you only have sample data. There is no version of the formula for a case where you only have sample data. The version of the formula using s_{n-1} is useful if you are calculating point-biserial correlation coefficients in a programming language or other development environment where you have a function available for calculating s_{n-1}, but don't have a function available for calculating s_n.

To clarify:

$$r_{pb} = \frac{M_1 - M_0}{s_n}\sqrt{\frac{n_1 n_0}{n^2}} = \frac{M_1 - M_0}{s_{n-1}}\sqrt{\frac{n_1 n_0}{n(n-1)}}.$$

Glass and Hopkins' book *Statistical Methods in Education and Psychology*, (3rd Edition)[1] contains a correct version of point biserial formula.

Also the square of the point biserial correlation coefficient can be written:

$$\frac{(M_1 - M_0)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\left(\frac{n_1 n_0}{n}\right).$$

We can test the null hypothesis that the correlation is zero in the population. A little algebra shows that the usual formula for assessing the significance of a correlation coefficient, when applied to r_{pb}, is the same as the formula for an unpaired *t*-test and so

$$r_{pb}\sqrt{\frac{n_1 + n_0 - 2}{1 - r_{pb}^2}}$$

follows Student's t-distribution with (n_1+n_0 - 2) degrees of freedom when the null hypothesis is true.

One disadvantage of the point biserial coefficient is that the further the distribution of *Y* is from 50/50, the more constrained will be the range of values which the coefficient can take. If *X* can be assumed to be normally distributed, a better descriptive index is given by the biserial coefficient

$$r_b = \frac{M_1 - M_0}{s_n}\frac{n_1 n_0}{n^2 u},$$

where *u* is the ordinate of the normal distribution with zero mean and unit variance at the point which divides the distribution into proportions n_0/n and n_1/n. As you might imagine, this is not the easiest thing in the world to calculate and the biserial coefficient is not widely used in practice.

A specific case of biserial correlation occurs where *X* is the sum of a number of dichotomous variables of which *Y* is one. An example of this is where *X* is a person's total score on a test composed of *n* dichotomously scored items. A statistic of interest (which is a discrimination index) is the correlation between responses to a given item and the corresponding total test scores. There are three computations in wide use,[2] all called the *point-biserial correlation*: (i) the Pearson correlation between item scores and total test scores including the

item scores, (ii) the Pearson correlation between item scores and total test scores excluding the item scores, and (iii) a correlation adjusted for the bias caused by the inclusion of item scores in the test scores. Correlation (iii) is

$$r_{upb} = \frac{M_1 - M_0 - 1}{\sqrt{\frac{n^2 s_n^2}{n_1 n_0} - 2(M_1 - M_0) + 1}}.$$

A slightly different version of the point biserial coefficient is the rank biserial which occurs where the variable X consists of ranks while Y is dichotomous. We could calculate the coefficient in the same way as where X is continuous but it would have the same disadvantage that the range of values it can take on becomes more constrained as the distribution of Y becomes more unequal. To get round this, we note that the coefficient will have its largest value where the smallest ranks are all opposite the 0s and the largest ranks are opposite the 1s. Its smallest value occurs where the reverse is the case. These values are respectively plus and minus $(n_1 + n_0)/2$. We can therefore use the reciprocal of this value to rescale the difference between the observed mean ranks on to the interval from plus one to minus one. The result is

$$r_{rb} = 2\frac{M_1 - M_0}{n_1 + n_0},$$

where M_1 and M_0 are respectively the means of the ranks corresponding to the 1 and 0 scores of the dichotomous variable. This formula, which simplifies the calculation from the counting of agreements and inversions, is due to Gene V Glass (1966).

It is possible to use this to test the null hypothesis of zero correlation in the population from which the sample was drawn. If r_{rb} is calculated as above then the smaller of

$$(1 + r_{rb})\frac{n_1 n_0}{2}$$

and

$$(1 - r_{rb})\frac{n_1 n_0}{2}$$

is distributed as Mann–Whitney U with sample sizes n_1 and n_0 when the null hypothesis is true.

External links

- Point Biserial Coefficient (http://www.andrews.edu/~calkins/math/edrm611/edrm13.htm#POINTB) (Keith Calkins, 2005)

Notes

1. ^ Gene V. Glass and Kenneth D. Hopkins (1995). *Statistical Methods in Education and Psychology* (3rd edition ed.). Allyn & Bacon. ISBN 0-205-14212-5.

2. ^ Linacre, John (2008). "The Expected Value of a Point-Biserial (or Similar) Correlation (http://www.rasch.org/rmt/rmt221e.htm)". *Rasch Measurement Transactions* **22** (1): 1154.