Name: Dr.P K Sivakumaran,
Assistant Professor

Subject Code : 18BST46S

Subject Name :
Psychological Statistics

THE BISERIAL CORRELATION

In educational or psychological studies, we often come across situations where both the variables correlated are continuously measurable, while one of them is artificially reduced to dichotomy. In such a situation, when we try to compute correlation between a continuous variable and a variable reduced to artificial dichotomy, we always compute the coefficient of biserial correlation. At this point, the question may arise as to what do we mean by a dichotomy as also by an artificial and a natural dichotomy?

The term *dichotomous* means cut into two parts or divided into two categories. This reduction into two categories may be the consequence of the nature of the data obtained. For example, in a study to find out whether or not a student passes or fails a certain standard, we place the crucial point dividing pass and fail students anywhere we please. Hence, measurement in the variable is reduced to two categories (pass and fail). This reduction into two categories, however, is not natural as we can have the crucial or dividing point according to our convenience.

Such a reduction of the variable into two artificial categories (artificial dichotomy) may be seen in the following classifications:

1. Socially adjusted and socially maladjusted
2. Athletic and non-athletic
3. Radical and conservative
4. Poor and not poor
5. Social minded and mechanical minded
6. Drop-outs and stay-ins
7. Successful and unsuccessful
8. Moral and immoral.

If we try to analyze the nature of distributions involving these dichotomized variables (i.e. adjustment in the topmost classification), we can come to the conclusion that artificial dichotomy is based on a clear assumption that the variable underlying the dichotomy should be continuous and normal.

In the two-fold division of socially adjusted and socially maladjusted, the division is quite artificial. If sufficient data were available, we could have found the trait 'adjustment', normally distributed among the studied population and it could have been distributed equally, instead of being discrete or limited (restricted to two-fold division).

In conclusion, we may term a dichotomy (division of a variable into two categories) an artificial dichotomy, when we do not have any clear-cut crucial point or criteria for such a division. We fix the dividing point according to our own convenience. In case sufficient data were available, the continuity as well as the normality of the distribution involving this variable can be easily established. Hence the basic assumption in using biserial correlation as an estimate of the relationship between a continuous variable and a dichotomous variable is that the variable underlying the dichotomy is continuous and normal. This implies that it should be an artificial dichotomous variable rather than a natural dichotomous variable.

Computation of Biserial Coefficient of Correlation

FORMULA. The general formula for biserial coefficient of correlation (r_{bis}) is

$$r_{\text{bis}} = \frac{M_p - M_q}{\sigma_t} \times \frac{pq}{y}$$

where

p = Proportion of cases in one of the categories (group) of dichotomous variable

q = Proportion of cases in the lower group = $1 - p$

M_p = Mean (M) of the values of higher group

M_q = Mean (M) of the values of the lower group

σ_t = Standard deviation (SD) of the entire group

y = Height of the ordinate of the normal curve separating the portion p and q

Let us illustrate the use of this formula with an example.

Example 12.1: The following table shows the distribution of scores on an achievement test earned by two groups of students those who passed and those who failed in a test of Arithmetic. Compute the coefficient of biserial correlation.

Scores on a test of achievement	*Result in Arithmetic test*	
	Passed	*Failed*
185–194	7	0
175–184	16	0
165–174	10	6
155–164	35	15
145–154	24	40
135–144	15	26
125–134	10	13
115–124	3	5
105–114	0	5
	120	110

Solution. The formula for calculating r_{bis} is

$$r_{\text{bis}} = \frac{M_p - M_q}{\sigma_t} \times \frac{pq}{y}$$

Based on this, we can assign values to the above variables by using the following steps:

Step 1. Here,

$$p = \text{Proportion of cases in the higher group}$$

$$= \frac{n_1}{N} = \frac{\text{Those who passed}}{\text{Total No. of students}} = \frac{120}{120 + 110}$$

$$= \frac{120}{130} = .52$$

Step 2. $q = 1 - p = 1 - .52 = .48$

Step 3. y = Height of the normal curve ordinate separating the portions p and q

= .3984 (as read from the Table G given in th Appendix)

Step 4. For the calculation of M_p (Mean of the scores of the passed group), M_q (Mean of the scores of the failed group), and σ_t (Standard deviation of scores of total group), we shall now compute as in Table 12.1.

Table 12.1 Worksheet for Calculating M_p, M_q and σ_t

	Arithmetic Test									
Achievement test scores	*Pass students (higher group)*	*Failures (lower group)*	*Total*	*Higher group*		*Lower group*		*Entire group*		
				x'	fx'	y'	fy'	z'	fz'	fz'^2
185–194	7	0	7	4	28	4	0	4	28	112
175–184	16	0	16	3	48	3	0	3	48	144
165–174	10	6	16	2	20	2	12	2	32	64
155–164	35	15	50	1	35	1	15	1	50	50
145–154	24	40	64	0	0	0	0	0	0	0
135–144	15	26	41	–1	–15	–1	–26	–1	–41	41
125–134	10	13	23	–2	–20	–2	–26	–2	–46	92
115–124	3	5	8	–3	–9	–3	–15	–3	–24	72
105–114	0	5	5	–4	0	–4	–20	–4	–20	80
	$n_1 = 120$	$n_2 = 110$	$N = 230$	$\Sigma fx' = 87$		$\Sigma fy' = -60$		$\Sigma fz' = 27$	$\Sigma fz'^2 = 655$	

$$M_p = A + \frac{\Sigma fx'}{N} \times i$$

where

$$A = \text{Assumed mean} = \frac{145 + 154}{2} = 149.5$$

$$x' = \frac{X - A}{i} = \frac{\text{Mid-value of } X \text{ scores} - \text{Assumed mean}}{\text{Class interval}}$$

Here,

N = Total No. of passed students = 120

i = Class interval = 10

Hence

$$M_p = 149.5 + \frac{87}{120} \times 10$$

$$= 149.5 + \frac{87}{12} = 149.5 + 7.25 = 156.75$$

$$M_q = A + \frac{\Sigma fy'}{N} \times i$$

$$= 149.5 + \frac{(-60)}{110} \times 10 = 149.5 - \frac{60}{11}$$

$$= 149.5 - 5.45 = 144.05$$

$$M_p \neg M_q = 156.75 - 144.05 = 12.70$$

$$\sigma_t = i\sqrt{\frac{fz'^2}{N} - \left(\frac{fz'}{N}\right)^2}$$

$$= 10\sqrt{\frac{655}{230} - \left(\frac{27}{230}\right)^2}$$

UGC

$$= 10\sqrt{\frac{655}{230} - \frac{27 \times 27}{230 \times 230}}$$

$$= \frac{10}{230}\sqrt{655 \times 230 - 27 \times 27}$$

$$= \frac{1}{23}\sqrt{149{,}921} = \frac{387.2}{23} = 16.83$$

Step 5. Substitute the value of p, q, y, M_p, M_q and σ_t in the following formula:

$$r_{\text{bis}} = \frac{M_p - M_q}{\sigma_t} \times \frac{pq}{y}$$

$$= \frac{12.7 \times .52 \times .48}{16.83 \times .3984}$$

$$= \frac{127 \times 52 \times 480}{1683 \times 398} = 0.47$$

[**Ans.** Coefficient of biserial correlation, $r_{\text{bis}} = 0.47$.]

Alternative Formula for r_{bis}

The coefficient of biserial correlation, r_{bis}, can also be computed with the help of the following formula:

$$r_{\text{bis}} = \frac{M_p - M_t}{\sigma_t} \times \frac{p}{y}$$

In this formula, we have to compute M_t (mean of the entire group) in place of M_q.

Characteristics of Biserial Correlation

The biserial correlation coefficient, r_{bis}, is computed when one variable is continuous and the other variable is artificially reduced to two categories (dichotomy).

Assumptions. The biserial correlation coefficient, r_{bis} gives an estimate of the product moment r for the given data when the following assumptions are fulfilled:

1. Continuity in the dichotomized trait
2. Normality of the distribution underlying the dichotomy
3. A large N
4. A split near the median

Limitations

1. The biserial r cannot be used in a regression equation.
2. Does not have any standard error of estimate.
3. Is not limited unlike r to a range of ±1.00.
4. Creates problems in matching comparison with other coefficients of correlation.

THE POINT BISERIAL CORRELATION

As already discussed, we resort to the computation of point biserial correlation coefficient ($r_{p,\text{bis}}$) for estimating the relationship between two variables when one variable is in a continuous state and the other is in the state of a natural or genuine dichotomy.

We have already thrown light on the nature of genuine or natural dichotomy, by distinguishing it clearly from the artificial dichotomy. Hence, if we are sure that the dichotomized variable does not belong to the category of artificial dichotomy, then we should try to compute point biserial correlation coefficient ($r_{p,\text{ bis}}$).

Computation of Point Biserial Correlation Coefficient ($r_{p,\text{ bis}}$)

FORMULA. The general formula for $r_{p,\text{bis}}$ is

$$r_{p,\text{bis}} = \frac{M_p - M_q}{\sigma_t}\sqrt{pq}$$

and an alternative for this is

$$r_{p,\text{bis}} = \frac{M_p - M_t}{\sigma_t}\sqrt{p/q}$$

where

p = Proportion of cases in one of the categories (higher group) of dichotomous variable

q = Proportion of cases in the lower group = $1 - p$

M_p = Mean of the higher group, the first category of the dichotomous variable

M_q = Mean of the values of lower group

M_t = Mean of the entire group

σ_t = Standard deviation (SD) of the entire group

Example 12.2: The data given in the following table shows the distribution of scores of 100 students on a certain test (X) and on another test (Y) which was simply scored as right or wrong, 1 and 0. Compute the necessary coefficient of correlation.

Scores on Test X	*Those who responded rightly*	*Those who responded wrongly*
70–74	3	0
65–69	6	1
60–64	6	2
55–59	5	4
50–54	6	2
45–49	7	6
40–44	6	8
35–39	3	6
30–34	3	9
25–29	1	4
20–24	0	12
	46	54

Solution. Here we have to compute the correlation between two sets of variables, one of which is in a continuous measure and the other in a genuine dichotomy. Hence it needs the computation of coefficient of point biserial correlation ($r_{p,\text{bis}}$). Therefore, the formula for $r_{p,\text{bis}}$ is

$$r_{p,\text{bis}} = \frac{M_p - M_t}{\sigma_t}\sqrt{p/q}$$

For finding the related values of the formula, let us proceed as follows:

Step 1.

p = Proportion of cases in the first group

$$= \frac{46}{100} = 0.46$$

Step 2. $q = 1 - p = 1 - 0.46 = 0.54$

Step 3. Calculation of M_p, M_t *and* σ_t

The computation process is illustrated in Table 12.2

Table 12.2 Worksheet for Computation of M_p and σ_t

Scores on X	*Those who responded rightly*	*Those who responded wrongly*	*Total*	x'	fx'	z'	fz'	fz'^2
70–74	3	0	3	5	15	5	15	75
65–69	6	1	7	4	24	4	28	112
60–64	6	2	8	3	18	3	24	72
55–59	5	4	9	2	10	2	18	36
50–54	6	2	8	1	6	1	8	8
45–49	7	6	13	0	0	0	0	0
40–44	6	8	14	–1	–6	–1	–14	14
35–39	3	6	9	–2	–6	–2	–18	36
30–34	3	9	12	–3	–9	–3	–36	108
25–29	1	4	5	–4	–4	–4	–20	80
20–24	0	12	12	–5	0	–5	–60	300
	$n_1 = 46$	$n_2 = 54$	$N = 100$		$\Sigma fx' = 48$		$\Sigma fz' = -55$	$\Sigma fz'^2 = 841$

$$M_p = A + \frac{\Sigma fx'}{n_1} \times i = 47 + \frac{48}{46} \times 5 = 47 + 5.2 = 52.2$$

(Here, A = Assumed mean = $\dfrac{45 + 49}{2} = 47$, $i = 5$ and $n_1 = 46$)

$$M_t = A + \frac{\Sigma fz'}{N} \times \text{i} = 47 + \frac{-55}{100} \times 5 = 47 - 2.75 = 44.2$$

$$\sigma_t = i\sqrt{\frac{\Sigma fz'^2}{N} - \left(\frac{\Sigma fz'}{N}\right)^2} = 5\sqrt{\frac{841}{100} - \left(\frac{-55}{100}\right)^2} = 5\sqrt{\frac{841}{100} - \frac{55 \times 55}{100 \times 100}}$$

$$= \frac{5}{100}\sqrt{84{,}100 - 3025} = \frac{1}{20}\sqrt{81{,}075} = \frac{284.73}{20} = 14.236$$

$$r_{p,\text{bis}} = \frac{M_p - M_t}{\sigma_t}\sqrt{p/q} = \frac{52.2 - 44.2}{14.236}\sqrt{0.46/0.54}$$

$$= \frac{8}{14.236}\sqrt{0.8518518} = 0.562\sqrt{0.923} = 0.52$$

[**Ans.** Point biserial correlation coefficient $r_{p,\text{bis}} = 0.52$.]

Which One of the Correlation r_{bis} or $r_{p,\text{bis}}$ is Better and Why?

The biserial correlation coefficient, r_{bis} has an advantage over $r_{p,\text{bis}}$ in that tables are available from which we can quickly read the values of r_{bis} (with sufficient accuracy). All we need to know are the values of p and q (percentages of passing a given item in relation to the higher and the lower groups). However, as a whole $r_{p,\text{bis}}$ is always regarded as a better and a much more dependable statistics than r_{bis} on account of the following features:

- The point biserial correlation makes no assumptions regarding the form of distribution in the dichotomized variable where biserial correlation makes too many assumptions such as continuity, normality, and large N split near the median .
- It may be used in regression equation.
- In comparison to r_{bis}, it can be easily and conveniently computed.
- The point biserial r is a product moment r and can be checked against r. This is usually not possible with r_{bis}.
- Like Pearson r, the range of $r_{p,\text{bis}}$ is equal to ±1, but this is not true for r_{bis}. Due to its range, $r_{p,\text{bis}}$ can be easily compared with other measures of correlation.
- The standard error of $r_{p,\text{bis}}$ can be exactly determined and its significance can be easily tested against the null hypothesis.
- Although $r_{p,\text{bis}}$ and r_{bis} both are useful in item analysis yet r_{bis} is generally not as valid or a defensible measure as $r_{p,\text{bis}}$.
- It is always safe to compute $r_{p,\text{bis}}$ when we are not sure whether the dichotomy is natural or artificial. However, the use of r_{bis} is always restricted to the artificial dichotomy of the dichotomized variable.

THE TETRACHORIC CORRELATION

Sometimes we find situations where both, the variables are dichotomous (reduced to two categories) and none of them can be expressed in scores. In such situations, we cannot use biserial or point biserial as a measure of correlation between these variables. We can only use Tetrachoric correlation or compute the ϕ (phi) coefficient.

We make use of tetrachoric correlation when these variables have artificial dichotomy. However, the basic assumption of these variables can be stated as follows: *Both variables are continuous, normally distributed and linearly related to each other, if it were possible to obtain scores or exact measures for them.*

In practice, these variables are not expressed in scores, though they are artificially separated into the following categories:

1. To study the relationship between intelligence and emotional maturity, the first variable, "intelligence," may be dichotomized as *above average* and *below average* and the other variable "emotional maturity", as *emotionally mature* and *emotionally immature*.
2. If we want to study the relationship between "adjustment" and "success" in a job, we can dichotomize the variables as *adjusted–maladjusted* and *success–failure*.
3. If we want to seek correlation between "poverty" and "delinquency", we can dichotomize the variables as *poor–not poor,* and *delinquent–non-delinquent*.

Computation of Tetrachoric Correlation

Let us illustrate the computation of tetrachoric correlation with the help of an example.

Example 12.3: In order to seek correlation between adjustment and job success, the data were obtained in a 2 × 2 table as shown in the following representation: Compute the tetrachoric correlation.

	X-variable		
	Success	Failure	
Adjusted	25	35	60
Maladjusted	20	40	60
	45	75	120

In such problems, the contents of entries in a 2 × 2 table are denoted by *A*, *B*, *C* and *D*, and the formula for computation of tetrachoric correlation (r_t) is as follows:

(i) When $AD > BC$,

$$r_t = \cos\left(\frac{180° \times \sqrt{BC}}{\sqrt{AD} + \sqrt{BC}}\right)$$

Where *A*, *B*, *C*, and *D* are frequencies in 2 × 2 table. Here the value of r_t is always positive.

(ii) When $BC > AD$,

$$r_t = \cos\left(\frac{180° \times \sqrt{AD}}{\sqrt{AD} + \sqrt{BC}}\right)$$

Here, the value of r_t is always negative.

(iii) When $BC = AD$,

$$r_t = \cos \frac{180° \times \sqrt{AD}}{2\sqrt{AD}} = \cos 90° = 0$$

Here, the value of r_t is always 0.

THE PHI (ϕ) COEFFICIENT

In studies where we have to compute correlation between two such variables which are genuinely dichotomous, it is the ϕ coefficient that is computed. Generally, its computation may involve the following situations:

1. When the classification of the variables into two categories is entirely and truly discrete, we are not allowed to have more than two categories, i.e. living vs. dead, employed vs. not employed, blue vs. brown eyes and so on.
2. When we have test items which are scored as Pass-Fail, True-False, or opinion and attitude responses, which are available in the form of yes-no, like-dislike, agree-disagree etc., no other intermediate type of responses is allowed.
3. With such dichotomized variables which may be continuous and may even be normally distributed, but are treated in practical operations as if they were genuine dichotomies, e.g. test items that are scored as either right or wrong, 1 and 0 and the like.

Computation of Phi (ϕ) Coefficient

FORMULA. The formula for computation of ϕ coefficient is

$$\phi = \frac{AD - BC}{\sqrt{(A+\mathrm{B})(C+D)(B+D)(A+C)}}$$

where A, B, C, D represent the frequencies in the cells of the following 2×2 table:

		X-variable		
		Yes	No	Total
Y-variable	Yes	A	B	$A + B$
	No	C	D	$C + D$
	Total	$A + C$	$B + D$	$A + B + C + D$

Let us illustrate the use of forgoing formula with the help of an example.

Example 12.4: There were two items X and Y in a test which were responded by a sample of 200, given in the 2×2 table. Compute the phi coefficient of correlation between these two items.

Solution.

		Item X		
		Yes	No	Total
Item Y	Yes	55 (A)	45 (B)	100 $(A + B)$
	No	35 (C)	65 (D)	100 $(C + D)$
	Total	90 $(A + C)$	110 $(B + D)$	200 $(A + B + C + D)$

FORMULA.

$$\phi \text{ coefficient} = \frac{AD - BC}{\sqrt{(A + \mathrm{B})(C + D)(B + D)(A + C)}}$$

Substituting the respective values in the formula, we obtain

$$\phi = \frac{55 \times 65 - 45 \times 35}{\sqrt{100 \times 100 \times 110 \times 90}}$$

$$= \frac{3575 - 1575}{\sqrt{99000000}}$$

$$\phi = \frac{2000}{1000\sqrt{99}} = \frac{2}{\sqrt{99}} = \frac{2}{9.95} = .201$$

[Ans. ϕ coefficient = 0.201.]

Features and Characteristics of ϕ Coefficient

- The phi coefficient is used for measuring the correlation between two variables when both are expressed in the form of genuine or natural dichotomies.
- The phi coefficient has the same relation with tetrachoric correlation (r_t) as point biserial ($r_{p,\text{ bis}}$) has with the biserial coefficient (r_{bis}).
- It can be checked against pearson 'r' obtained from the same table.
- It is most useful in item analysis when we want to know the item to item correlation.
- It bears a relationship with χ^2 to be expressed as $\chi^2 = N\phi^2$.
- The values of phi coefficient range between –1 and +1, but these are influenced by marginal totals.
- It makes no assumptions regarding the form of distribution in dichotomized variables like r_t which needs the assumptions of large N and continuity and normality of the distributions.
- For providing a better measure like r_t, it does not require a split near median and large N rather, it proves to be a better measure when the split is away from the median.
- Standard error of ϕ can be easily computed and ϕ can be easily tested against the null hypothesis by means of its relationship to χ^2.
- When there is any doubt regarding the exact nature of the dichotomized variables, it is always safe to compute ϕ. Also, its computation is much easier and regarded as a better and a more dependable statistics than r_t.

EXERCISES

1. What is biserial correlation? How is it computed?
2. What is point biserial correlation? How is it different from biserial correlation? How is it computed?
3. Discuss, with the help of examples, the various situations when you have to compute r_{bis}, $r_{p,\text{ bis}}$, r_t and ϕ as a measure of relationship between two variables.
4. What is tetrachoric correlation? When is it computed? Discuss its computation process.
5. What is phi coefficient? Why is it regarded as a better and more dependable statistics than tetrachoric correlation? Discuss its computational process.
6. Compute the coefficient of biserial correlation from the given data to know the extent to which success on job is related to adjustment.

Scores on adjustment scale	*Success on job*	*Failure on job*
95–99	1	0
90–94	6	0
85–89	18	1
80–84	22	1
75–79	31	3
70–74	20	5
65–69	18	9
60–64	12	13
55–59	6	10
50–54	4	8
45–49	1	5
40–44	0	3
35–39	1	0
30–34	0	1
25–29	0	1
	140	60

7. (a) Find the coefficient of biserial correlation from the following data:

Performance scores	*Normal and above intelligence*	*Below normal intelligence*
130–139	5	0
120–129	7	0
110–119	21	3
100–109	26	7
90–99	30	16
80–89	27	21
70–79	10	11
60–69	3	4
50–59	1	6
40–49	0	2
	130	70

(b) A group of students, with and without training, obtained the following scores on a performance test. Find out the biserial correlation between training and performance:

Performance test scores	*Trained*	*Untrained*
90–99	6	0
80–89	19	3
70–79	31	5
60–69	58	17
50–59	40	30
40–49	18	14
30–39	9	7
20–29	5	4
	186	80

8. (a) Compute the point biserial correlation coefficient from the data given in Problem 6.

(b) A group of individuals were asked to answer 'yes' or 'no' for a particular item. Compute the point biserial correlation coefficient between the item and total score from the following data:

Total scores on the opinion scale	*Yes*	*No*
95–99	0	1
90–94	1	1
85–89	0	6
80–84	2	11
75–79	4	6
70–74	6	9
65–69	8	3
60–64	3	2
55–59	2	1
50–54	6	0
45–49	2	0
40–44	3	0
35–39	1	0
30–34	1	0
	39	40

9. (a) Compute tetrachoric correlation from the given data to find the relationship between emotional maturity and the state of married life.

	Happy married life	*Unhappy married life*
Emotionally mature	65	35
Emotionally immature	25	75
	90	110

(b) 125 students were first tested on a test of achievement motivation and then on a test of anxiety, and the results were tabulated as follows: Find the tetrachoric correlation between the level of achievement motivation and anxiety.

	High anxiety level	*Low anxiety level*
High achievement motivation	30	40
Low achievement motivation	35	20
	65	60

10. (a) 100 individuals in a survey sample responded to items Nos. 15 and 20 of an interest invertary (in "yes" or "no") as given in the following table. Compute the ϕ coefficient with the help of cell frequencies.

	Item No. 15	
Item No. 20	*No*	*Yes*
Yes	27	20
No	24	29
	51	49

(b) The number of candidates passing and failing in two items of a test are given in the following tabular representation.

	Item No. 1	
Item No. 2	*Pass*	*Fail*
Pass	80	55
Fail	20	70
	100	125

Compute phi coefficient between these two items.