A NOTE ON THE SIGNIFICANCE OF r_{pb}

LEONEL CAMPOS
Ateneo de Manila University

It is suggested here that the point-biserial coefficient of correlation, r_{pb}, is the square root of the ratio of the "Between" Sum of Squares to the "Total" Sum of Squares, of the one-factor, completely randomized analysis of variance design, and that the extrapolation, $t = r_{pb}\sqrt{(N\text{-}2)\ \ (1 - r_{pb}^2)}$, is distributed as Student's t with degrees of freedom, $df = N - 2$, and is therefore an appropriate test of significance for r_{pb}.

The point-biserial product-moment coefficient of correlation, r_{pb}, is an algebraic hybrid related to Pearson's well known statistic, r. It is obtained by formula, as follows,

$$r_{pb} = \frac{(m_1 - m_2)\sqrt{n_1 n_2}}{S_t \cdot N} \qquad (1)$$

where

m_1 and m_2, the arithmetic means of each of two groups within a larger, dichotomized sample;

n_1 and n_2, the number of cases in each of the groups mentioned above, respectively $(n_1 + n_2 = N)$, and

S_t, the standard deviation of the whole undichotomized sample.

Among the problems complicating the interpretation of r_{pb} is the fact that it does not seem to be amenable to a definite test of significance. Some authors (e.g., McNemar, 1956, p. 195) emphasize the dependency of r_{pb} on the difference between two means, and suggest an ordinary t test as the appropriate test of significance. Other authors (e.g., Peatman, 1963, p. 312), prescribe the use of the extrapolation

$$t = r_{pb}\sqrt{\frac{N - 2}{1 - r_{pb}}} \qquad (2)$$

since the quantity yielded by (2) is known to be distributed as Student's t with degrees of freedom, $df = N - 2$, when Pearson's r is substituted by r_{pb}; however, it is not clear why formula (2) is recommended for use with r_{pb}, unless one makes the dubious assumption that if (2) works with r it must also work with r_{pb}, as well as with any one of the other members of the family of product-moment correlations, i.e., the biserial r, Spearman's *Rho*, etc. To confound matters further, one author (Guilford, 1965, p. 323) argues that "the hypothesis of zero correlation [with r_{pb}] can be tested *in two ways*" (Italics mine). The *two ways* listed by Guilford are the two alternatives already cited above.

It turns out, however—and that is the main argument of this communication—that these so-called alternatives simply represent two equivalent ways of doing exactly the same thing. The proof is straightforward and is given next.

Consider the one-factor, completely randomized design in the Analysis of Variance. Here it is well known that

$$\sum_j \sum_i (X_{ij} - M)^2 = \sum_j \sum_i (X_{ij} - m_j)^2 + \sum_j n_j (m_j - M)^2 \quad (3)$$

where

X_{ij}, an individual score i, found in group group j (i: 1, 2, 3, . . . , n)

m_j, the arithmetic mean of group j (j: 1, 2, 3, . . . , k);

M, the mean of the total, undichotomized sample.

$\sum_j$ the summation over group means, or group totals, up to the kth case;

$\sum_i$, the summation over individual observations up to the nth case, and,

n_j, the number of observations in the jth group.

In expression (3), the term on the left side of the equality is known as the "Total" Sum of Squares (SS_T); the first term on the right side of the equation is known as the "Within Groups" Sum of Squares" sum of Squares (SS_W).

If we take $\sum_j n_j (m_j - M)^2$ and expand the binomial, when K = 2, i.e., when only two groups are being considered, we obtain

$$\sum_j n_j (m_j - M)^2 = n_1 m_1^2 + n_2 m_2^2 - \frac{(n_1 m_1 + n_2 m_2)^2}{(n_1 + n_2)} \quad (4)$$

$$= \frac{(n_1 m_1)^2 + (n_2 m_2)^2 + n_1 n_2 (m_1^2 + m_2^2) - (n_1 m_1 + n_2 m_2)^2}{(n_1 + n_2)}$$

Since,

$$(n_1 m_1)^2 + (n_2 m_2)^2 = (n_1 m_1 + n_2 m_2)^2 - 2 n_1 n_2 m_1 m_2$$

equation (4) becomes, after substitution and simplification,

$$\sum_j n_j (m_j - M)^2 = \frac{n_1 n_2 (m_1 - m_2)^2}{(n_1 + n_2)} \quad (5)$$

If we now take square root of both sides of (5), we get

$$\sqrt{\sum_j n_j (m_j - M)^2} = \sqrt{SS_B} = (m_1 - m_2) \sqrt{\frac{n_1 n_2}{N}} \quad (5a)$$

Take next the "Total" Sum of Squares" and multiply it by N/N, that is, by unity. The result is

$$\frac{N}{N} \sum_j \sum_i (X_{ij} - M)^2 = NV_t \quad (6)$$

where V_t is the total variance of the sample in question. Again, square root of both sides of (6) yields

$$\sqrt{N} \sqrt{\frac{SS_T}{N}} = S_t \sqrt{N} \quad (6a)$$

where S_t is the Standard deviation of the total, undichotomized sample. Finally, if equation (5a) is divided by (6a), we get

$$\sqrt{\frac{SS_B}{SS_T}} = \frac{(m_1 - m_2)\sqrt{n_1 n_2}}{S_t N} = r_{pb} \quad (7)$$

the formula of the point-biserial r. Formula (7) shows that the point-biserial r is a trivial instance of the general case

$$R = \sqrt{\frac{SS_B}{SS_T}} \quad (8)$$

well known to statisticians, where R is a general measure of correlation and symbols inside the radical are as defined earlier. Formula (7) also indicates that the r_{pb} cannot reach unity as long as there is some residual variability within the groups.

This fact suggests that r_{pb} underestimates correlations present in the data from which it is obtained. Coincidentally, a rank-analogue of the point-biserial r described by Campos & Santos (1968) consistently gave higher values than r_{pb} when both statistics were computed from the same data. r_u, the statistic of Campos and Santos, can have values of —1 and 1, where r_{pb} cannot have the same values.

Further, if $r_{pb} = \sqrt{SS_B / SS_T}$, this identity may be substituted into equation (2) to obtain

$$t = \sqrt{\frac{SS_B}{SS_T}} \Bigg/ \sqrt{\frac{1 - \frac{SS_B}{SS_T}}{(N-2)}}$$

$$= \sqrt{\frac{SS_B\ (N-2)}{SS_T - SS_B}} \qquad (9)$$

It is enough to recall that $SS_T - SS_B = SS_w$ and that, when $k=2$, the "Between Groups" Means Square, MS_B, is $MS_B = SS_B/(k-1) = SS_B$; Also, the "Within Groups" Mean Square is $MS_w = SS_w/(N-2)$. Therefore, formula (9) can be rewritten as

$$t = \sqrt{\frac{MS_B}{MS_w}} \qquad (10)$$

which is the square root of Fisher's ratio between two variances and which is distributed as Student's t with degrees of freedom $df = N - 2$. This is a happy coincidence. Formula (10) shows that equation (2) contains a *bonafide* t test. On the other hand, expression (10) also suggests that (2) is not distributed as Student's t when it is used to evaluate the biserial r; neverthelss, the biserial r is related to r_{pb} in a very definite manner and the problem is solved by transforming the biserial r into r_{pb}, and then finding t.

In any case, the main point made here is that equation (2) is a direct test for significance of the difference between two independent means, and an appropriate test for the hypothesis that r_{pb} is zero.

One numerical example may further dramatize what has already been said. Suppose we had the following sets of numbers:

Set 1				Set 2		
59	99	91	63	31	57	25
84	75	48	54	96	14	56
41	85	74	98	45	25	38
62	48	59	37	12	43	46
35	61	33	49	27	17	21
98	32	85	85	54	42	54
77	54	67	76	32	33	19

TABLE 1

ANALYSIS OF VARIANCE OF THE DATA OF SETS 1 AND 2

Source of Variation	df	SS	MS	F
Between Sets	1	9304.3081	9304.3081	
				22.7887
Within Sets	47	19189.4687	408.2865	
Total	48	28493.7768		

For these data,
$M_1 = 62.2500$; $n_1 = 28$; $M_2 = 37.4761$; $n_2 = 21$;
$S_t = 24.0794$, and $N = 49$.

Computation of r_{pb} by formula (1) gives

$$r_{pb} = \frac{(62.3214) - 37.4762) \sqrt{(28)\ (21)}}{(24.1144)\ (49)}$$
$$= 0.5714$$

and the t test as obtained by formula (2),

$$t = (0.5714) \sqrt{\frac{47}{1 - (0.5714)^2}}$$
$$= 4.7733.$$

On the other hand, an Analysis of Variance of the same data yields results as shown in table 1.

Finally, by formula (7) we get

$$r_{pb} = \sqrt{\frac{9304 \cdot 3081}{28493 \cdot 7768}}$$
$$= 0.5714$$

and, since $t = \sqrt{F}$

$$\therefore t = \sqrt{22.7132} = 4.75658.$$

These two latest values match the values obtained earlier through conventional methods.

REFERENCES

CAMPOS, L. & SANTOS, J. Mann-Whitney's *U* as an indicator of relationship. *Philippine Journal of Psychology*, 1969, 2, 31-33.

GUILFORD, J. P. *Fundamental Statistics for Psychology and Education.* (4th ed.) New York: McGraw-Hill, 1965.

MCNEMAR, Q. *Psychological Statistics.* (3rd ed.) New York: Wiley, 1953.

PEATMAN, J. G. *Inroduction to Applied Statistics.* New York: Harper, 1963.

SIEGEL, S. *Nonparametric Statistics for the Behavioral Sciences.* New York: McGraw-Hill, 1956.