NOTES

CORRELATION BETWEEN A DISCRETE AND A CONTINUOUS VARIABLE. POINT-BISERIAL CORRELATION

By Robert F. Tate

University of Washington[1]

1. Introduction and Summary. A problem of some importance in statistical applications, especially in the field of psychology, is that of finding a measure of association between a discrete random variable X, which takes the values 0 and 1, and a continuous random variable Y. The ordinary product-moment correlation coefficient $\rho(X, Y)$ is used for this purpose. It has received the name point-biserial correlation coefficient because of its relation to the biserial correlation coefficient proposed by Karl Pearson for a similar problem. The usual estimator r, based on a random sample (X_i, Y_i), $i = 1, 2, \cdots, n$, is referred to as the sample point-biserial correlation coefficient.

The psychological value of ρ (and hence of r) is that it affords a measure of the degree of association between a trait and a measurable characteristic, usually an ability of some kind. For the ith individual in a random sample of n individuals, X_i has the value 1 if the trait is possessed and Y_i is a measure of the ability in question.

We shall give in Section 2 the appropriate mathematical model, based on normal theory, and the asymptotic distribution of r (Theorem 1), the derivation of which is an elementary application of a well known theorem of Cramér. An important special case of this distribution will be discussed in Section 3, namely that in which X takes the values 0 and 1 with equal probabilities. In this connection a variance-stabilizing transformation will be given (Theorem 2). Numerical work based on this transformation may be carried out with the use of existing tables. In particular, the calculation of confidence limits for ρ is immediate. Theorem 2 is especially useful in investigating the association between sex and some other characteristic, since animal populations consist of approximately half males and half females. As an illustration of the ease with which calculations may be carried out, a problem is considered in which the trait is male and the characteristic is IQ.

The small-sample distribution of r is quite easily found, although it is difficult to deal with when n is even moderately large, asymptotic methods appearing to be more desirable. This is discussed in Section 4.

Received 6/29/53, revised 1/26/54.

[1] This research was performed while the author was at the Statistical Laboratory, University of California, Berkeley, and was sponsored in part by the Office of Naval Research.

2. Model and asymptotic distribution. Consider (X_i, Y_i), $i = 1, 2, \cdots, n$, a sequence of independent random vectors. Let X_i have the Bernoulli distribution:

$$P(X_i = 1) = p, \qquad P(X_i = 0) = q, \qquad 0 < p < 1, p + q = 1.$$

Let each Y_i have the mixed distribution with distribution function $F(y) = pF_1(y) + qF_0(y)$, where

$$F_j(y) = P(Y \leqq y \mid X = j) = \int_{-\infty}^{y} \frac{1}{\tau\sqrt{2\pi}} e^{-[(z-\mu_j)^2/2\tau^2]}\, dz, \quad j = 0, 1.$$

The following notation will be used. The standardized difference of means, $(\mu_1 - \mu_0)/\tau$, will be denoted by Δ. The variance of the random variable Z will be denoted by $V(Z)$. The fact that Z is asymptotically normal with mean a and variance b^2 will be denoted by $Z \sim \mathfrak{N}(a, b^2)$. Of course,

$$r = \{\sum(X_iY_i) - n\bar{X}\bar{Y}\}/\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}.$$

Throughout the paper the indices for all summations run from 1 to n. Easy calculation shows that

$$E(X) = p, \qquad V(X) = pq, \qquad E(Y) = p\mu_1 + q\mu_0, \qquad V(Y) = \tau^2(1 + pq\Delta^2),$$
$$E(XY) = p\mu_1, \qquad \rho(X, Y) = \Delta\sqrt{pq/(1 + pq\Delta^2)}.$$

THEOREM 1.

$$r \sim \mathfrak{N}\left[\rho, \frac{4pq - \rho^2(6pq - 1)}{4npq} \cdot (1 - \rho^2)^2\right].$$

As r is a function of sample means which has a total differential at the point given by the expectations of these means, this result is obtainable by a lengthy, but elementary, calculation from a theorem of Cramér [1]. Moments $\mu_{\kappa\lambda} = E\{(X - p)^\kappa[Y - E(Y)]^\lambda\}$ are needed. In this case

$$\mu_{\kappa\lambda} = \tau^\lambda\{pq^\kappa a(\lambda) + (-p)^\kappa qb(\lambda)\},$$

where

$$a(\lambda) = \int_{-\infty}^{+\infty} (t + q\Delta)^\lambda \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt, \qquad b(\lambda) = \int_{-\infty}^{+\infty} (t - p\Delta)^\lambda \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt.$$

Straightforward analysis leads to Theorem 1.

3. A special case. It may easily be shown that the asymptotic variance of r has a minimum for each ρ when $p = q = \frac{1}{2}$, since, for each ρ,

$$V(r \mid P) - V(r \mid \tfrac{1}{2}) = \frac{1}{n}\rho^2(1 - \rho^2)^2\left(\frac{1}{4pq} - 1\right) \geqq 0.$$

In this event we obtain the greatest precision (in terms of the smallest confidence interval for ρ). This we should expect intuitively because of the obvious analogy between our set-up and that of the ordinary two-sample t test, since in the case of that test it is well known that equal sample sizes yield greatest power. For the case $p = \frac{1}{2}$ we have from Theorem 1

$$r \sim \mathfrak{N}\left[\rho, \frac{1}{2n}(2 - \rho^2)(1 - \rho^2)^2\right].$$

The parameter ρ may be removed from the variance by the variance stabilizing transformation $\phi(x)$ which satisfies the equation

$$\phi'(x) = \sqrt{2}/(1 - x^2)\sqrt{2 - x^2}.$$

The desired solution is $\phi(x) = \sqrt{\tfrac{1}{2}}\,\mathrm{sgn}(x)\cdot\mathrm{sech}^{-1}(1 - x^2)$. The function $\tanh^{-1}(x)$ is given in Table VB of Fisher [2]. Hence it is convenient to express sech^{-1} in terms of $\tanh^{-1}$ and express our final result as follows.

THEOREM 2.

$$\mathrm{sgn}(r)\cdot\tanh^{-1}\sqrt{1 - (1 - r^2)^2} \sim \mathfrak{N}[\mathrm{sgn}(\rho)\cdot\tanh^{-1}\sqrt{1 - (1 - \rho^2)^2},\ 2/n].,$$

EXAMPLE. Calculations associated with Theorem 2 are quite easy. Consider, for example, that we are sampling school children of some fixed age. Let $X_i = 1$ if the ith child is a boy and 0 if the ith child is a girl. Let Y_i be the IQ of the ith child. Assuming variability of IQ to be the same for the two sexes, we shall use the following data for 25 school children in order to determine 95 per cent confidence limits for ρ.

Boys				Girls				
106	143	109	109	93	115	105	107	111
98	114	98	85	119	113	117	111	92
109	107	96		91	104	89	85	

$$\sum x_i = 11 \qquad \sum y_i = 2626 \qquad \sum x_i y_i = 1174 \qquad 1.96\sqrt{(2/n)} = .5542$$

$$\sum x_i^2 = 11 \qquad \sum y_i^2 = 279738 \qquad r = +.120$$

From Table VIII of Pearson [5], $1 - (1 - r^2)^2 = .02891$.

From Table VB of Fisher [2], $\mathrm{sgn}(r)\tanh^{-1}\sqrt{1 - (1 - r^2)^2} = .1718$.

Therefore, 95 per cent confidence limits for $\mathrm{sgn}(\rho)\tanh^{-1}\sqrt{1 - (1 - \rho^2)^2}$ are $.1718 \pm .5542$. To find confidence limits for ρ, we use the same tables in reverse order to solve the equations

$$\tanh^{-1}\sqrt{1 - (1 - \rho^2)^2} = .7260, \quad \text{and} \quad = .3824.$$

Taking the positive solution for the first and the negative solution for the second, we obtain $(-.263, .465)$ as 95 per cent confidence limits for ρ.

4. The small sample distribution of r. Let $T = r\sqrt{(n - 2)/(1 - r^2)}$. The small sample distribution of r can be obtained easily by making use of a relation given by Lev [3]. Lev considered not the distribution of T, but only the conditional distribution of T. More precisely, let $\sum X_i = n_1$ and $n_0 = n - n_1$. Then

$$r(n_0, n_1) = \sqrt{\frac{n_0 n_1}{n}} (\bar{Y}_1 - \bar{Y}_0) \Bigg/ \sqrt{\sum_{i=0}^{1} \sum_{j=1}^{n\,i} (Y_{ij} - \bar{Y})^2}, \tag{4.1}$$

where $\bar{Y}_0 = \sum(1 - X_i)Y_i/n_0$, $\bar{Y}_1 = \sum(X_iY_i)/n_1$, $\bar{Y} = \sum Y_i/n$. Expression (4.1) denotes the conditional value of r given $\sum X_i = n_1$. It can be shown (Lev) that

$$\frac{\sqrt{n - 2}\, r(n_0, n_1)}{\sqrt{1 - r^2(n_0, n_1)}} = \sqrt{\frac{n_0 n_1}{n}} (\bar{Y}_1 - \bar{Y}_0) \Bigg/ \sqrt{\frac{1}{n - 2} \sum_{i=0}^{1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2}.$$

Denoting this quantity by $t(n_0, n_1)$, we have (Lev)

$$t(n_0, n_1) = \left\{ \frac{(\bar{Y}_1 - \bar{Y}_0) - (\mu_1 - \mu_0)}{\tau\sqrt{n/n_0 n_1}} + \frac{(\mu_1 - \mu_0)}{\tau\sqrt{n/n_0 n_1}} \right\} \Bigg/ \sqrt{\frac{1}{\tau^2(n - 2)} \sum_{i=0}^{1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2}.$$

The random variable $t(n_0, n_1)$ has the noncentral t distribution with $n - 2$ degrees of freedom, and parameter of noncentrality

$$\delta = \Delta\sqrt{n_0 n_1/n} = \rho\sqrt{n_0 n_1/npq(1 - \rho^2)}.$$

Denoting the density of $t(n_0, n_1)$ by $f(t; n_1, n, p, \rho)$ and the density of T by $g(t; n, p, \rho)$, we have, by the definition of $t(n_0, n_1)$,

$$g(t; n, p, \rho) = \sum_{n_1=0}^{n} \binom{n}{n_1} p^{n_1} q^{n-n_1} f(t; n_1, n, p, \rho).$$

When $\rho = 0$, we see that $f(t; n_1, n, p, 0)$ is independent of n_1 and p, since $\delta = 0$. Hence, $g(t; n, p, 0)$ is also independent of p and is the density of the ordinary t distribution with $n - 2$ degrees of freedom. Thus, to test the hypothesis $H: \rho = 0$ at level of significance α, we reject H when $|T| \geqq k_\alpha$, where k_α is obtained from a table of the t distribution. The power against an alternative of the form (p, ρ) can be computed from the expression

$$\beta(p, \rho) = 1 - \sum_{n_1=0}^{n} \binom{n}{n_1} p^{n_1} q^{n-n_1} \int_{-k_\alpha}^{+k_\alpha} f(t; n_1, n, p, \rho)\, dt.$$

The integrals may be evaluated directly from the tables of Neyman and Tokarska [4] for level of significance $\alpha = .05$ or $\alpha = .01$. They use the symbol ρ for parameter of noncentrality, instead of our δ.

If a large sample is available, quicker and sufficiently accurate results may be obtained from Theorem I. Then the above procedure, which amounts to looking up n values in the Neyman-Tokarska table, can be avoided.

REFERENCES

[1] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, 1946, pp. 359, 366–367.

[2] R. A. Fisher, *Statistical Methods for Research Workers*, Oliver and Boyd (1946), p. 210.

[3] J. LEV, "The point-biserial coefficient of correlation," *Ann. Math. Stat.*, Vol. 20 (1949), pp. 125–126.

[4] J. NEYMAN AND B. TOKARSKA, "Errors of the second kind in testing 'Student's' hypothesis," *J. Amer. Stat. Assn.*, Vol. 31 (1936), pp. 318–326.

[5] K. PEARSON, *Tables for Statisticians and Biometricians*, Part I, Cambridge University Press, 1945, p. 20.

A COMPUTING FORMULA FOR THE POWER OF THE ANALYSIS OF VARIANCE TEST

BY W. L. NICHOLSON

University of Oregon[1] *and University of Illinois*[2]

1. Summary. A formula for the power of the analysis of variance test is derived for the case when the denominator of the F ratio has an even number of degrees of freedom. The form employed is particularly adapted to computation of the power as a function of the alternative hypothesis with arbitrary fixed level of significance and fixed degrees of freedom. For m degrees of freedom in the numerator and 2, 4, 6, 8 and 10 in the denominator, the power functions are deduced from the general formula, with an indication of their use.

2. The power function. In the classical analysis of variance test we are interested in a ratio of the form

$$F = n \sum_{i=1}^{m} x_i^2 \Big/ m \sum_{j=1}^{n} y_j^2, \tag{1}$$

where x_i $(i = 1, 2, \cdots, m)$ and y_j $(j = 1, 2, \cdots, n)$ are distributed $N(\theta_i, \sigma^2)$ and $N(0, \sigma^2)$, respectively. If the null hypothesis, $\theta_i = 0$ $(i = 1, 2, \cdots, m)$, is false, it is well known that the distribution of F is completely specified by m, n, and the single additional parameter

$$\lambda = \frac{1}{2\sigma^2} \sum_{i=1}^{m} \theta_i^2. \tag{2}$$

Therefore, for a predetermined level of significance α, the power of the test is a function of m, n, and λ. It is [1]

$$P(\lambda \mid a, b; \alpha) = 1 - \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} I_x(a + k, b), \tag{3}$$

where $m = 2a$ and $n = 2b$, and

$$I_x(a + k, b) = \frac{1}{\beta(a + k, b)} \int_0^x t^{a+k-1}(1 - t)^{b-1}\, dt \tag{4}$$

Received 2/24/52, revised 2/12/53.

[1] Research under contract with the Office of Naval Research, U. S. Navy.

[2] Research under contract with the Office of Ordnance Research, U. S. Army.