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.