LETTERS TO THE EDITOR

On the K-Statistic.

In an earlier paper[1] worked out by the author jointly with Mr R. C. Bose the problem of discrimination between and classification of different multi variate normal populations (with the same sets of variances and covariances but with different sets of means) was sought to be tackled by defining the Studentised D^2-statistic and obtaining its sampling distribution. But the problem yet remained to be tackled when the populations differ in the sets of variances and covariances and when it is sought to discriminate between and classify them in the light of their variances and covariances. In a paper which is to shortly appear in *Sankhyā, the Indian Journal of Statistics*, a first step is taken towards answering the following question. Can the two multivariate normal populations from which two samples have been drawn be reasonably (*i.e.* on a given probability level) assumed to have the same set of variances and covariances? The problem of analysis of variance is, of course, intimately connected with the above and the two are really solved together. For the univariate case the problem is complely solved by the sampling distribution of s_1/s_2 and Fisher's χ

In the earlier paper[1] by the author already quoted it was noted that if following Fisher[2] one started from a linear compound with arbitrary coefficients of the variates and so chose the compounding coefficients as to make a maximum the ratio between the square of the difference of the sample means (for the compound character) and the within variance of the samples for the same character, one could easily see after maximisation that the ratio spoken of above is proportional to the Studentised D^2-statistic of that paper. In the coming paper, starting from a similar linear compound of the variates and so choosing the compounding coefficients as to make a maximum the ratio of variances of the 1st and 2nd samples (for the compound character), one finds that the quest for the maximised value of the ratio leads one, in the case of p-variate populations, to as many as p statistics instead of to merely one as in the previous case. These p statistics are given by the roots of the following pth degree determinantal equation in K^2

$$|a_{ij} - K^2 a'_{ij}| = 0 \qquad \ldots \qquad \ldots \qquad (1)$$

where a_{ij} is the co-variance of the 1st sample for the i-th and j-th characters and a'_{ij} is the covariance of the 2nd sample for the same characters. Of these p roots $(K^2_1, K^2_2, \ldots K_p^2)$ one, say, K^2_1, is a true maximum, another, say, K^2_p is a true minimum, while the rest are stationary values. In the case when the populations sampled are identical in sets of variances and covariances (which is the hypothesis to be tested) and when we go on drawing two samples of sizes n and n' respectively from the two populations, the joint sampling distribution of the p functions $(K_1, K_2, \ldots\ldots K_p)$ has come out in the form

$$\text{Const.} \prod_{i=1}^{p} \frac{K_i^{\,n-p-1}}{(1+c^2K_i^2)^{\frac{n+n'-2}{2}}} . (K^2_1 - K^2_2) \ldots . (K^2_1 - K_p^2)$$

$$(K^2_2 - K^2_3) \ldots (K^2_2 - K_p^2) \ldots (K^2_{p-1} - K_p^2) \times \prod_{i=1}^{p} dK_i \qquad (2)$$

where the symbol Π stands for the product of a certain number of terms, and $c^2 = n'/n$.

The K's vary each from 0 to ∞; suppose that *in any particular case* K_1 is the maximum and K_p the minimum. Then if $K_1 \geq K_p$ and also $\geq 1/K_p$, then *in such a case* K_1 is the statistic which serves best for purposes of discrimination and whose sampling distribution we should seek. If on the other hand $K_1 \leq 1/K_p$ then K_p is the statistic best suited for discrimination and one whose sampling disribution should be obtained. In the first case integrating out (2) successively over $K_p, K_{p-1}, \ldots K_2$, each from 0 to K_1 we have the distribution of K_1 (the maximum statistic) in the form, say,

$$\text{Const. } F(K_1)\, dK_1 \qquad \ldots \qquad \ldots \qquad (3)$$

In the latter case integrating out (2) successively over $K_1, K_2, \ldots K_{p-1}$, each from K_p to ∞, the distribution of K_p (the minimum statistic) is obtained, say, in the form

$$\text{Const. } f(K_p)\, dK_p \qquad \ldots \qquad \ldots \qquad (4)$$

It should be noted from (2) that both $F(K_1)$ and $f(K_p)$ of (3) and (4) are the sums of a certain number of terms each of which is the product of incomplete Γ functions

In a paper which is to appear in *Sankhyā* very shortly after the next it is proposed to obtain (3) and (4) in a form suitable for purposes of statistical application, *i.e.* for numerical computations. Once we have prepared tables for the incomplete probability integrals of (3) and (4) we can immediately use K_1 or K_p for purposes of discrimination. In any given problem K_1 and K_p can, of course, be obtained numerically to any desired degree of approximation from the sample readings by solving the determinantal equation (1) by, say, Horner's method.

Statistical Laboratory,
Presidency College,
Calcutta, 18-7-39.

Samarendra Nath Ray.

[1] Bose, R. C. and Ray, S. N.: The Distribution of the Studentised D^2-Statistic, *Sankhyā*, Vol. 4 (1), pp. 19-38.

[2] Fisher, R. A.: The use of Multiple Measurements in Taxonomic Problems, *Annals of Eugenics*, Vol. 7 (2), pp. 179-188, 1936.