A NOTE ON THE DISTRIBUTION OF DIFFERENCES IN MEAN VALUES OF TWO SAMPLES DRAWN FROM TWO MULTIVARIATE NORMALLY DISTRIBUTED POPULATIONS, AND THE DEFINITION OF THE D^2-STATISTIC.

BY RAJ CHANDRA BOSE

STATISTICAL LABORATORY, CALCUTTA.

1. I am indebted to Professor S. N. Bose of Dacca for pointing out a mistake which occurred in my paper "*On the Exact Distribution and Moment-coefficients of the D^2-statistic*" published in Volume 2, Part 2 of this journal, in writing down the distribution of the set of mean differences of two samples drawn from two multi-variate normally distributed populations. I obtain in this note the correct distribution which necessitates a small alteration in the definition of the D^2-statistic for the case of correlated variates. The net result is that the formulæ (1·6), (1·7), (2·2) and (3·1) of my previous paper need modifications, which I give here ; but the main investigation remains completely valid.

It is also shown that if we assume the two populations to have the same set of variances and covariances, no modifications in the results given previously are necessary. As in practical applications of the D^2-statistic, it is usually assumed that the variances and covariances are identical in the two populations, the results obtained in the previous paper can be used legitimately.

2. Let Σ and Σ' be two random samples of sizes n and n' drawn respectively from two normal populations π and π' of p-variates which are linearly correlated. We shall write down the population statistics in the following way :

$$\left.\begin{aligned} \alpha_i &= \text{mean value of the } i\text{-th character in population } \pi \\ \alpha'_i &= \text{mean value of the } i\text{-th character in population } \pi' \end{aligned}\right\} \quad \dots \quad (2\cdot1)$$

$$\left.\begin{aligned} \sigma_i &= \text{standard deviation of the } i\text{-th character in population } \pi \\ \sigma'_i &= \text{standard deviation of the } i\text{-th character in population } \pi' \end{aligned}\right\} \quad \dots \quad (2\cdot2)$$

$$\left.\begin{aligned} \rho_{ij} &= \text{coefficient of correlation between the } i\text{-th and } j\text{-th character in population } \pi \\ \rho'_{ij} &= \text{coefficient of correlation between the } i\text{-th and } j\text{-th character in population } \pi' \end{aligned}\right\} (2\cdot3)$$

We also write $\qquad \alpha_{ij} = \sigma_i \sigma_j \rho_{ij}, \qquad \alpha'_{ij} = \sigma_i' \sigma_j' \rho'_{ij} \qquad \dots \quad (2\cdot4)$

*The D^2-statistic was intended to be and was defined as a quantity determined entirely by the sample values of the variates. Raj Chandra Bose has investigated the exact distribution of a modified form of the D^2-statistic in which the population values of the variances and co-variances have been substituted for the corresponding sample estimates.—*Editor, Sankhyā.*

For the two samples Σ and Σ' we write

$$\left.\begin{aligned} a_i &= \text{observed mean value of the } i\text{-th character in sample } \Sigma \\ a'_i &= \text{observed mean value of the } i\text{-th character in sample } \Sigma' \end{aligned}\right\} \qquad \ldots \quad (2\cdot5)$$

Let Λ_{pq} be the co-factor of α_{pq} in the determinant

$$\Lambda \equiv \begin{vmatrix} \alpha_{11} & \alpha_{12} & \ldots\ldots\ldots\ldots & \alpha_{1p} \\ \alpha_{21} & \alpha_{22} & \ldots\ldots\ldots\ldots & \alpha_{2p} \\ \ldots & \ldots & \ldots\ldots\ldots\ldots & \ldots \\ \alpha_{p1} & \alpha_{p2} & \ldots\ldots\ldots\ldots & \alpha_{pp} \end{vmatrix} \equiv |\alpha_{pq}| \qquad \ldots \qquad \ldots \quad (2\cdot6)$$

and Λ'_{pq} the co-factor of α'_{pq} in the determinant

$$\Lambda' \equiv \begin{vmatrix} \alpha'_{11} & \alpha'_{12} & \ldots\ldots\ldots\ldots & \alpha'_{1p} \\ \alpha'_{21} & \alpha'_{22} & \ldots\ldots\ldots\ldots & \alpha'_{2p} \\ \ldots & \ldots & \ldots\ldots\ldots\ldots & \ldots \\ \alpha'_{p1} & \alpha'_{p2} & \ldots\ldots\ldots\ldots & \alpha'_{pp} \end{vmatrix} \qquad \ldots \qquad \ldots \quad (2\cdot7)$$

Then the distribution of the set of mean differences $(a_1, a_2 \ldots\ldots\ldots a_p)$ in repeated samples of size n drawn from the population π may be written as

$$\text{Constant} \times e^{-(n/2\Lambda)\{\Lambda_{11}(a_1-\alpha_1)^2 + \ldots + 2\Lambda_{12}(a_1-\alpha_1)(a_2-\alpha_2) + \ldots\}} . da_1 . da_2 \ldots da_p \quad \ldots \quad (2\cdot8)$$

and likewise the distribution of the set of mean differences $(a_1', a_2', \ldots\ldots a_p')$ in repeated samples of size n' drawn from the population π' may be written as

$$\text{Const.} \times e^{-(n/2\Lambda')\{\Lambda'_{11}(a_1'-\alpha_1')^2 + \ldots\ldots + 2\Lambda_{12}'(a_1'-\alpha_1')(a_2'-\alpha_2') + \ldots\ldots\}} . da_1' \, da_2' \ldots\ldots da_p' \quad (2\cdot9)$$

Our immediate object is to write down the distribution of the set of mean differences $(a_1 - a_1', \; a_2 - a_2', \ldots\ldots\ldots, \; a_p - a_p')$.

3. Let us set

$$\left.\begin{aligned} (a_i - \alpha_i) + (a_i' - \alpha_i') &= 2x_i \\ (a_i - \alpha_i) - (a_i' - \alpha_i') &= 2y_i \end{aligned}\right\} \quad (i = 1, 2, \ldots\ldots p) \qquad \ldots \quad (3\cdot1)$$

$$\left.\begin{aligned} (a_i - \alpha_i) &= (x_i + y_i) \\ (a_i' - \alpha_i') &= (x_i - y_i) \end{aligned}\right\} \quad (i = 1, 2, \ldots\ldots p) \qquad \ldots \quad (3\cdot2)$$

We shall further set $\gamma_{ij} = \dfrac{n\Lambda_{ij}}{2\Lambda}$, $\gamma'_{ij} = \dfrac{n'\Lambda'_{ij}}{2\Lambda'}$ $\qquad \ldots \quad (3\cdot3)$

Then from (1·7) and (1·8) the joint distribution of $x_1, x_2, \ldots\ldots x_p$; $y_1, y_2, \ldots\ldots y_p$ may be written as

$$\text{Const.} \times exp \{ -\gamma_{11}(x_1+y_1)^2 - \ldots\ldots - 2\gamma_{12}(x_1+y_1)(x_2+y_2) - \ldots\ldots - \gamma_{11}'(x_1-y_1)^2 - \ldots\ldots - 2\gamma_{12}'(x_1-y_1)(x_2-y_2) - \ldots\ldots\} \, dx_1 \, dx_2 \ldots\ldots dx_p \, dy_1 \, dy_2 \ldots\ldots dy_p \quad (3\cdot4)$$

or $$\text{Const.} \times exp \, [-\{(\gamma_{11}+\gamma_{11}')x_1^2 + \ldots\ldots + 2(\gamma_{12}+\gamma_{12}')x_1x_2 \ldots\ldots\} - 2\{x_1(\gamma_{11}y_1 + \gamma_{12}y_2 + \ldots\ldots \gamma_{1p}y_p) + \ldots\ldots - x_1(\gamma_{11}'y_1 + \gamma_{12}'y_2 + \ldots\ldots + \gamma_{1p}'y_p) + \ldots\ldots\} - \{(\gamma_{11}+\gamma_{11}')y_1^2 + \ldots\ldots + 2(\gamma_{12}+\gamma_{12}')y_1y_2 + \ldots\ldots\}] dx_1 \, dx_2 \ldots\ldots dx_p \, dy_1 \, dy_2 \ldots\ldots dy_p \quad \ldots \quad (3\cdot5)$$

4. The expression within the squared brackets in (3·5), when equated to zero, may be looked upon as representing a hyperquadric in a space of p-dimensions ; $x_1, x_2, \ldots\ldots x_p$ being regarded as the variables, and $y_1, y_2, \ldots\ldots y_p$ being momentarily regarded as constants.

Hence there exists a linear transformation by which the origin is transformed to the centre of this quadric. This transformation is of the type $x_i = x_i' + \text{const}$ ... (4·1)

Making this transformation, the joint distribution of $x_1', x_2', \ldots x_p'$; $y_1, y_2, \ldots\ldots y_p$ can be written

$$\text{Const.} \times exp \, [-\{(\gamma_{11}+\gamma_{11}')x_1'^2 + \ldots\ldots + 2(\gamma_{12}+\gamma_{12}')x_1'x_2' + \ldots\ldots\} - \phi(y_1, y_2, \ldots\ldots y_p)/k] \times dx_1' \, dx_2' \ldots\ldots dx_p' \, dy_1 \, dy_2 \ldots\ldots dy_p \quad \ldots \quad (4\cdot2)$$

where $$k \equiv \begin{vmatrix} \gamma_{11}+\gamma_{11}' & \gamma_{12}+\gamma_{12}' & \cdot & \gamma_{1p}+\gamma_{1p}' \\ \gamma_{21}+\gamma_{21}' & \gamma_{22}+\gamma_{22}' & \cdot & \gamma_{2p}+\gamma_{2p}' \\ \ldots & \ldots & & \ldots \\ \gamma_{p1}+\gamma_{p1}' & \gamma_{p2}+\gamma_{p2}' & \cdot & \gamma_{pp}+\gamma_{pp}' \end{vmatrix} \quad \ldots \quad (4\cdot3)$$

and $$\phi(y_1, y_2, \ldots y_p) \equiv \begin{vmatrix} \gamma_{11}+\gamma_{11}' & \gamma_{12}+\gamma_{12}' & \cdot\cdot & \gamma_{1p}+\gamma_{1p}' & \frac{1}{2}\left(\frac{\partial\psi}{\partial y_1} - \frac{\partial\psi'}{\partial y_1}\right) \\ \gamma_{21}+\gamma_{21}' & \gamma_{22}+\gamma_{22}' & \cdot\cdot & \gamma_{2p}+\gamma_{2p}' & \frac{1}{2}\left(\frac{\partial\psi}{\partial y_2} - \frac{\partial\psi'}{\partial y_2}\right) \\ \ldots & \ldots & \cdot\cdot & \ldots & \ldots \\ \gamma_{p1}+\gamma_{p1}' & \gamma_{p2}+\gamma_{p2}' & \cdot\cdot & \gamma_{pp}+\gamma_{pp}' & \frac{1}{2}\left(\frac{\partial\psi}{\partial y_p} - \frac{\partial\psi'}{\partial y_p}\right) \\ \frac{1}{2}\left(\frac{\partial\psi}{\partial y_1} - \frac{\partial\psi'}{\partial y_1}\right) & \frac{1}{2}\left(\frac{\partial\psi}{\partial y_2} - \frac{\partial\psi'}{\partial y_2}\right) & \cdot\cdot & \frac{1}{2}\left(\frac{\partial\psi}{\partial y_p} - \frac{\partial\psi'}{\partial y_p}\right) & \psi+\psi' \end{vmatrix} \quad (4\cdot4)$$

$$\text{where } \psi \equiv \gamma_{11} y_1^2 + \ldots\ldots \; 2\gamma_{12} y_1 y_2 + \ldots\ldots = \sum_{i,j=1}^{p} \gamma_{ij} y_i y_j \quad \ldots \quad (4\cdot5)$$

$$\psi' \equiv \gamma_{11}' y_1^2 + \ldots\ldots \; 2\gamma_{12}' y_1 y_2 + \ldots\ldots = \sum_{i,j=1}^{p} \gamma_{ij}' y_i' y_j \quad \ldots \quad (4\cdot6)$$

To simplify (4·4) we multiply, the 1st, 2nd, ... pth column of the determinant by $y_1, y_2, \ldots\ldots y_p$ and add to the $(p+1)$th column, then from Euler's theorem

$$\varphi(y_1, y_2, \ldots\ldots y_p) \equiv \begin{vmatrix} \gamma_{11}+\gamma_{11}' & \gamma_{12}+\gamma_{12}' & \ldots & \gamma_{1p}+\gamma_{1p}' & \frac{\partial\psi}{\partial y_1} \\ \gamma_{21}+\gamma_{21}' & \gamma_{22}+\gamma_{22}' & \ldots & \gamma_{2p}+\gamma_{2p}' & \frac{\partial\psi}{\partial y_2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \gamma_{p1}+\gamma_{p1}' & \gamma_{p2}+\gamma_{p2}' & \ldots & \gamma_{pp}+\gamma_{pp}' & \frac{\partial\psi}{\partial y_p} \\ \frac{1}{2}\left(\frac{\partial\psi}{\partial y_1}-\frac{\partial\psi'}{\partial y_1}\right) & \frac{1}{2}\left(\frac{\partial\psi}{\partial y_2}-\frac{\partial\psi'}{\partial y_2}\right) & \ldots & \frac{1}{2}\left(\frac{\partial\psi}{\partial y_p}-\frac{\partial\psi}{\partial y_p'}\right) & 2\psi \end{vmatrix} \quad (4·7)$$

Again multiplying the 1st, 2nd, ... pth row by $y_1, y_2, \ldots\ldots y_p$, and subtracting from the last row we get on again applying Euler's theorem

$$\varphi(y_1, y_2, \ldots y_p) \equiv - \begin{vmatrix} \gamma_{11}+\gamma_{11}' & \gamma_{12}+\gamma_{12}' & \ldots & \gamma_{1p}+\gamma_{1p}' & \frac{\partial\psi}{\partial y_1} \\ \gamma_{21}+\gamma_{21}' & \gamma_{22}+\gamma_{22}' & \ldots & \gamma_{2p}+\gamma_{2p}' & \frac{\partial\psi}{\partial y_2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \gamma_{p1}+\gamma_{p1}' & \gamma_{p2}+\gamma_{p2}' & \ldots & \gamma_{pp}+\gamma_{pp} & \frac{\partial\psi}{\partial y_p} \\ \frac{\partial\psi}{\partial y_1} & \frac{\partial\psi'}{\partial y_2} & \ldots & \frac{\partial\psi'}{\partial y_p} & 0 \end{vmatrix} \quad (4·8)$$

In (4·2) we can integrate out for $x_1', x_2' \ldots\ldots x_p'$. Hence we get as the distribution of $y_1, y_2, \ldots\ldots y_p$

$$\text{Const.} \times e^{-\varphi(y_1, y_2, \ldots y_p)/k} \, dy_1\, dy_2 \ldots\ldots dy_p$$

5. If k_{ij} denotes the cofactor of the element in the i-th row and the j-th column of the determinant k given by (4·3) we have

$$\varphi(y_1, y_2, \ldots y_p) = \sum_{i,j=1}^{p} k_{ij} \frac{\partial\psi}{\partial y_i} \frac{\partial\psi}{\partial y_j} = 4 \sum_{i,j=1}^{p} l_{ij}\, y_i\, y_j \quad \ldots \quad (5·1)$$

where

$$l_{mn} = l_{nm} = \sum_{i,j=1}^{p} k_{ij}\, \gamma_{mi}\, \gamma'_{nj} \quad \ldots \quad (5·2)$$

Substituting for $y_1, y_2, \ldots\ldots y_p$, from (3·1), and using (4·9) and (5·1), we can write the distribution of $(a_1 - a_1', a_2 - a_2', \ldots\ldots a_p - a_p')$ in the form

$$e^{(-1/k)[l_{11}\{(a_1-a_1')-(\alpha_1-\alpha_1')\}^2+\ldots+2l_{12}\{(a_1-a_1')-(\alpha_1-\alpha_1')\}\{(a_2-a_2')-(\alpha_2-\alpha_2')\}\ldots]}$$

$$\times\; d(a_1-a_1')\,d(a_2-a_2')\ldots\ldots d(a_p-a_p') \quad \ldots \quad (5·3)$$

6. Hence the formula (2·2) of my paper "*On the Exact Distribution and Moment Coefficients of the D^2-statistic*"; will be valid provided that we set

$$\beta = k, \qquad \beta_{ij} = \frac{4l_{ij}}{\bar{n}} \qquad \dots \quad (6\cdot1)$$

where k is defined by the relation (4·3) of this note, l_{ij} by the relation (5·2), and

$$\frac{2}{\bar{n}} = \frac{1}{n} + \frac{1}{n'} \qquad \dots \quad (6\cdot2)$$

The formulæ (1·6), (1·7) on page 145, *Sankhyā* vol. 2, part 2 should be dropped. The definition (3·1), on page 146 for the D^2-statistic remains valid, provided that by β and β_{ij} we understand the constants given by the relation (6·1) of the present note. No further corrections are necessary in the remainder of the paper.

7. Let us now go on to consider the special case

$$\alpha_{ij} = \alpha'_{ij} \qquad i, j = 1, 2, \dots\dots\dots\dots p \qquad \dots \quad (7\cdot1)$$

i.e. when the two populations π and π' have the same set of variances and covariances.

From (3·3) we see that $\qquad n\gamma'_{ij} = n'\gamma_{ij} \qquad \dots \quad (7\cdot2)$

and hence from (4·5) and (4·6) $\qquad n'\psi = n\psi' \qquad \dots \quad (7\cdot3)$

and also $$n'\frac{\partial\psi}{\partial y_i} = n\frac{\partial\psi'}{\partial y_i} \quad (i = 1, 2, \dots\dots p) \qquad \dots \quad (7\cdot4)$$

If we substitute for γ'_{ij} and $\partial\psi'/\partial y_i$ from (7·3) and (7·4) in (4·8); then multiply the 1st, 2nd, ... pth column of the determinant by

$$\frac{2n}{n+n_1}\, y_1, \; \frac{2n}{n+n_1}\, y_2, \; \dots\dots \frac{2n}{n+n_1}\, y_p \qquad \dots \quad (7\cdot41)$$

and subtract their sum from the last column, we see that

$$\phi(y_1, y_2, \dots\dots y_p) = \frac{4n'}{n+n'}\cdot\psi \cdot \left(\frac{n+n'}{n}\right)^p \; |\gamma_{ij}| \qquad \dots \quad (7\cdot42)$$

where $$|\gamma_{ij}| = \begin{vmatrix} \gamma_{11} & \gamma_{12} & \cdot & \gamma_{1p} \\ \gamma_{21} & \gamma_{22} & \cdot & \gamma_{2p} \\ \dots & \dots & \dots & \dots \\ \gamma_{p1} & \gamma_{p2} & \cdot & \gamma_{pp} \end{vmatrix} \qquad \dots \quad (7\cdot43)$$

Again $$K = \left(\frac{n+n'}{n}\right)^p |\gamma_{ij}| \qquad \dots \quad (7\cdot44)$$

Hence . (4·0) can be written as $\quad \text{Const.} \times e^{-4n'\psi/(n+n')} dy_1\, dy_2 \dots\dots dy_p \qquad \dots \quad (7\cdot45)$

Hence from (4·5), (3·3) and (6·2), we can write the distribution of $y_1, y_2 \ldots\ldots y_p$ in the form

$$\text{Const.} \times e^{-(\bar{n}/\Lambda)\{\Lambda_{11}y_1^2 + \ldots\ldots 2\Lambda_{12}y_1y_2 + \ldots\ldots\}} . dy_1, dy_2 \ldots \ldots dy_p \qquad \ldots \quad (7·5)$$

Finally from (3·1), we see that the distribution of $(a_1 - a_1', a_2 - a_2', \ldots\ldots a_p - a_p')$ can be written in the following form

$$\text{C}.e^{-(\bar{n}/4\Lambda)[\Lambda_{11}\{(a_1 - a'_1) - (\alpha_1 - \bar{\alpha}_1')\}^2 + \ldots + 2\Lambda_{12}\{(a_1 - a_1') - (\alpha_1 - \alpha_1')\}\{(a_2 - a_2') - (\alpha_2 - \alpha_2')\} +]}$$

$$\times\, d(a_1 - a_1')d(a_2 - a_2') \ldots\ldots d(a_p - a_p) \qquad \ldots \quad (7·6)$$

In this special case therefore, the results of my previous paper remain valid, without any correction ; since in this case $\bar{\alpha}_{pq}$ as defined on page 145, formula (1·6) of paper, is simply equal to α_{pq}, and consequently $\beta \equiv \Lambda$ and $\beta_{ij} \equiv \Lambda_{ij}$.

Thus if two populations π and π' have the same set of variances and covariances, then we can define D^2 by

$$D^2 = D_1^2 - 2/\bar{n} \qquad \ldots \quad (7·7)$$

where

$$D_1^2 = \frac{1}{P\Lambda} \{\Lambda_{11}(a_1 - a_1')^2 + \ldots\ldots 2\Lambda_{12}(a_1 - a_1')(a_2 - a_2') + \ldots\ldots\} \quad \ldots \quad (7·8)$$

8. If the variables in both the populations are independent. Then

$$\gamma_{ij} = \gamma'_{ij} = 0 \quad \text{when } i \neq j \,;\ \gamma_{ii} = \frac{n}{2} \; \frac{\Lambda_{ii}}{\Lambda} = \frac{n}{2\alpha_{ii}}; \quad \gamma'_{ii} = \frac{n}{2} \; \frac{\Lambda_{ii}'}{\Lambda'} = \frac{n}{2\alpha_{ii}'}$$

$$k_{ij} = 0 \text{ when } i \neq j,\ k_{ii} = \frac{k}{\gamma_{ii} + \gamma_{ii}'},\ l_{ij} = 0 \text{ when } i \neq j,\ l_{ii} = k_{ii}\gamma_{ii}\gamma_{ii}' = \frac{k\gamma_{ii}\,\gamma_{ii}'}{\gamma_{ii} + \gamma_{ii}'}$$

Therefore,

$$\beta_{ij} = 0 \text{ when } i \neq j, \quad \beta_{ii} = \frac{4}{\bar{n}} \cdot \frac{k\gamma_{ii}\,\gamma_{ii}'}{\gamma_{ii} + \gamma_{ii}'}$$

or

$$\beta_{ii} = \frac{4k}{\bar{n}} \cdot \frac{1}{(1/\gamma_{ii}) + (1/\gamma_{ii}')} = \frac{n + n'}{n'\alpha_{ii} + n\alpha_{ii}'}$$

Therefore,

$$D^2 = \frac{1}{p\beta} \{\beta_{11}(a_1 - a_1')^2 + \ldots\ldots\}$$

$$= \frac{1}{p} \left\{ \frac{(a_1 - a_1')^2}{\frac{n'\alpha_{11} + n\alpha_{11}'}{n + n'}} + \ldots\ldots \right\}$$

Thus for the uncorrelated case the generalised definition agrees with the one originally given by Mahalanobis.

May, 1930.
Statistical Laboratory, Calcutta.