

# A NOTE ON THE DISTRIBUTION OF THE STUDENTISED D<sup>2</sup>-STATISTIC

BY

S. N. ROY

STATISTICAL LABORATORY

## INTRODUCTION

In an earlier paper<sup>1</sup> published by the author jointly with Mr. R. C. Bose the Studentised D<sup>2</sup>-Statistic was defined, its statistical use and object were fully set forth and its sampling distribution was worked out. It is the object of the present note to considerably simplify the derivation of the distribution obtained there, by a twist in the geometrical arguments employed in that paper and a consequent change of procedure. This will incidentally throw fresh light on the use of hyperspace geometry in tackling this class of problems. To make this note intelligible even without reference to the earlier paper<sup>1</sup>, or in other words, to make this note practically independent of that, we have to take over three small sections from that paper, where definitions and notations were laid down and certain preliminaries were also introduced which will suffice for our present purpose.

## §1. PRELIMINARIES.

Consider two samples  $\Sigma$  and  $\Sigma'$  of sizes  $n$  and  $n'$  from two multivariate normal populations  $\Pi$  and  $\Pi'$  with the same set of variances and covariances  $a_{ij}$  ( $i, j = 1, 2, \dots, p$ ) where  $a_{ij} = \rho_{ij} \sigma_i \sigma_j$ ,  $\sigma_i$ ,  $\sigma_j$  and  $\sigma_k$  being the standard deviations for the  $i$ -th and  $j$ -th characters respectively, and  $\rho_{ij}$  the correlation co-efficient between the  $i$ -th and  $j$ -th characters.<sup>1</sup> The matrix  $\|a_{ij}\|$  will be said to be the common dispersion matrix for the two populations. Let  $a_{ij}$ ,  $a'_{ij}$  ( $i, j = 1, 2, \dots, p$ ) denote the respective variances and covariances of the samples  $\Sigma$  and  $\Sigma'$ , so that  $\|a_{ij}\|$  and  $\|a'_{ij}\|$  are their respective dispersion matrices. Let  $\sigma_i$ ,  $\sigma'_i$  ( $i = 1, 2, \dots, p$ ) be the means for the  $i$ -th character for the populations  $\Pi$  and  $\Pi'$  and let  $a_i$ ,  $a'_i$  denote the corresponding quantities for the samples  $\Sigma$  and  $\Sigma'$ .

Let us set

$$c_{ij} = \frac{na_{ij} + n'a'_{ij}}{n + n'} \quad \dots \quad (1.1)$$

Let  $c^u$  as usual denote the minor of  $c_{ij}$  in the determinant  $|c_{ij}|$  divided by the determinant itself. A like definition holds for  $a^u$ . Then the Studentised D<sup>2</sup>-statistic is defined by

$$D^2 = c^{11}(a_1 - a'_1)^2 + c^{22}(a_2 - a'_2)^2 + \dots + c^{pp}(a_p - a'_p)^2 + 2c^{12}(a_1 - a'_1)(a_2 - a'_2) + \dots + 2c^{p-1,p}(a_{p-1} - a'_{p-1})(a_p - a'_p) \quad \dots \quad (1.15)$$

Likewise if  $\Delta^3$  is the population value of  $D^3$ , then

$$\rho \Delta^3 = a^{11}(a_1 - a'_1)^2 + a^{22}(a_2 - a'_2)^2 + \dots + a^{pp}(a_p - a'_p)^2 + 2a^{12}(a_1 - a'_1)(a_2 - a'_2) + \dots + 2a^{p-1,p}(a_{p-1} - a'_{p-1})(a_p - a'_p) \dots \quad (1.2)$$

We know that the joint distribution of the sample readings  $x_{1k}, x'_{1k} (i=1, 2, \dots, \dots, p; k=1, 2, \dots, \dots, n; k'=1, 2, \dots, \dots, n')$  is

$$\text{const. } e^{-\frac{1}{2} \sum_{i=1}^p \sum_{k=1}^n a^{ii} [n(a_i - a_i) (a_i - a_i) + n'(a'_i - a'_i) (a'_i - a'_i) + (na_i + n'a'_i)]} \times \Pi dx_{1k} \Pi dx'_{1k} \dots \quad (1.25)$$

where  $\Pi dx_{1k}$  stands for  $dx_{11} dx_{12} dx_{13} \dots dx_{1n}$  and a similar meaning attaches to  $\Pi dx'_{1k}$ .

Sample  $\Sigma$  can be represented in the usual Fisherian space  $S_n$  of  $n$  dimensions, by the points with co-ordinates

$$(x_{11}, x_{12}, \dots, x_{1n}), i=1, 2, \dots, p. \dots \quad (1.3)$$

or what is the same thing, by  $p$  vectors  $x_i$  joining the points to the origin. We take another space  $S_{n'}$  of  $n'$  dimensions absolutely orthogonal to the former space, and represent in it the sample  $\Sigma'$  by  $p$  other similar vectors. Let

$$y_{1k} = x_{1k} - a_i, y'_{1k} = x'_{1k} - a'_i, \dots \quad (1.35)$$

where  $i=1, 2, \dots, p; k=1, 2, \dots, n; k'=1, 2, \dots, n'$

Let  $y_i, y'_i$  denote the vectors, with components  $(y_{1i}, y_{2i}, \dots, y_{ni})$  and  $(y'_{1i}, y'_{2i}, \dots, y'_{n'i})$  lying in the space  $S_n$  and  $S_{n'}$  respectively. Then the vectors  $y_i$  lie in a flat  $S_{n-1}$  of  $n-1$  dimensions, perpendicular to the equiangular line in  $S_n$ . Similar considerations apply to the vector  $y'_i$ .

Let  $O$  be the origin of co-ordinates and let  $M_i$  be the point on the equiangular line in  $S_n$  such that  $OM_i = \frac{1}{\sqrt{n}}$  times the projection of  $x_i$  on the equiangular line. Then  $OM_i = a_i$ . In the same way we can find  $M'_i$  on the other equiangular line such that  $OM'_i = a'_i$ . Also if  $y_i \cdot y_j$  is the scalar product of the vectors  $y_i$  and  $y_j$ , then clearly

$$a_{ij} = (y_i \cdot y_j)/n, a'_{ij} = (y'_i \cdot y'_j)/n' \dots \quad (1.4)$$

Let us now take a new set of  $(n+n')$  variable  $a_i, a'_i, z_{ik}, z'_{ik} (i=1, 2, \dots, p; k=1, 2, \dots, n-1; k'=1, 2, \dots, n'-1)$

such that

$$(i) a_i = \frac{1}{n} \sum_{k=1}^n x_{ik} \quad (i=1, 2, \dots, p)$$

$$(ii) a'_i = \frac{1}{n'} \sum_{k'=1}^{n'} x'_{ik'} \quad (i=1, 2, \dots, p)$$

STUDENTISED D<sup>3</sup>-STATISTIC

(iii)  $z_{ik}$  ( $k = 1, 2, \dots, n-1$ ) are the components of  $y_i$  along any  $n-1$  mutually orthogonal lines in  $S_{n-1}$ . Then  $z_{ik}$  is naturally a linear function of

$$x_{1i}, x_{2i}, \dots, x_{ni}$$

(iv) Similar considerations apply to  $z'_{ik}$ .

The distribution (1.25) can now be written in the form

$$\text{const} \times e^{-\frac{1}{2} \sum_{i=1}^n \sum_{k=1}^{n-1} a^i |n(a_i - a_i') + n'(a_i' - a_i')|} \times \prod_{i=1}^n \prod_{k=1}^{n-1} da_i' \prod_{i=1}^n \prod_{k=1}^{n-1} dz_{ik} \prod dz'_{ik} \dots \quad (1.45)$$

It should be noted that  $a_0$ 's are expressible purely in terms of  $z_{ik}$ 's and  $a_0$ 's are expressible purely in terms of  $z'_{ik}$ 's

Next we introduce the new variables  $a_i - a_i'$  and  $a_i + a_i'$  in place of  $a_i$  and  $a_i'$  and integrating for  $a_i + a_i'$  we get the distribution in the form

$$\text{const} \times e^{-\frac{1}{2} \sum_{i=1}^n \sum_{k=1}^{n-1} a^i \left[ \frac{1}{2} \{ (a_i - a_i') - (a_i - a_i') \} \{ (a_i - a_i') - (a_i - a_i') \} + N c_{ij} \right]} \times \prod_{i=1}^n d(a_i - a_i') \prod dz_{ik} \prod dz'_{ik} \dots \quad (1.5)$$

where  $c_{ij}$  is given by (1.1) and

$$\frac{2}{N} = \frac{1}{n} + \frac{1}{n'}, \quad N = n + n' \quad \dots \quad (1.55)$$

In the plane of the equiangular lines of  $S_n$  and  $S'_n$  let  $R_i$  be the point, whose projections on the equiangular lines coincide with  $M_i$  and  $M'_i$ . Then if  $Q_i$  is the projection of  $R_i$  on  $OY$ , the external bisector of the equiangular lines

$$OQ_i = \frac{1}{\sqrt{2}} (a_i - a_i') \quad (i = 1, 2, \dots, p) \quad \dots \quad (1.6)$$

Let the vector  $v_i$  be the resultant of the vectors  $y_i$  and  $y'_i$ . Then it is easily seen that

$$N c_{ij} = v_i \cdot v_j \quad \dots \quad (1.65)$$

where the dot denotes the scalar product.

We may note that the spaces  $S_{n-1}$ ,  $S'_{n-1}$  containing the vectors  $y_i$ ,  $y'_i$  respectively ( $i = 1, 2, \dots, p$ ) are orthogonal to one another, as also to  $OY$ . Hence the new vectors  $v_i$  are also orthogonal to  $OY$ , and lie in the space  $S_{n-2}$  which comprises both  $S_{n-1}$  and  $S'_{n-1}$ .

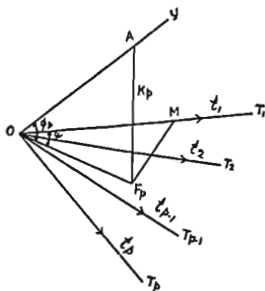
Let  $\ell_i$  be the resultant of the vectors  $a_i$  and the vector  $\bar{n} \perp Q \Omega$ , where  $\bar{n}$  is given by (1.55).

Then

$$\begin{aligned} \ell_i \cdot \ell_j &= \frac{\bar{n}}{2} (a_i - a'_i) (a_j - a'_j) + (na_{ij} + n'a_{ij}) \\ &= \frac{\bar{n}}{2} (a_i - a'_i) (a_j - a'_j) + N c_{ij} = g_{ij} \text{ (say)} \end{aligned} \quad (1.7)$$

The distribution (1.5) now takes the form

$$\begin{aligned} \text{const } x e^{-\frac{1}{2} p \Delta^2 - \frac{1}{2} \sum_{i=1}^p \sum_{j=1}^p a^{ij} \{g_{ij} - \bar{n}(a_i - a'_i) (a_j - a'_j)\}} \\ \times \prod d(a_i - a'_i) \prod dz_{1k} \prod dz'_{1k} \end{aligned} \quad (1.75)$$



Let  $T_i$  denote the extremity of the vector  $\ell_i$ .

Let  $OA$  be the unit vector along  $OY$ . We shall denote it by  $i$ . Let  $F_1$  denote the foot of the perpendicular from  $A$  to the space  $(T_1 T_2 \dots T_p)$ . Let the length  $AF_1$  be denoted by  $K_1$ .

Let us now consider the length  $K_p$ . We shall show that it is very closely connected with  $D^2$

$$\text{Now } K_p^2 = \frac{\text{Vol } (i, \ell_1, \ell_2, \dots, \ell_p)}{\text{Vol } (\ell_1, \ell_2, \dots, \ell_p)} \quad \dots (1.8)$$

But  $\text{Vol } (i, \ell_1, \ell_2, \dots, \ell_p)$  is the same as the volume formed by the unit vector  $QA$ , and the projections of the vectors  $\ell_1, \ell_2, \dots, \ell_p$  on the space perpendicular to  $OA$ , which we have earlier called  $S_{n,n-1}$ . But these projections, from the way in which

STUDENTISED D<sup>2</sup>-STATISTIC

$t_1, t_2, \dots, t_p$  have been derived, are at once seen to be  $v_1, v_2, \dots, v_p$ . Hence the numerator in (1.95) is numerically equal to  $\text{vol}(v_1, v_2, \dots, v_p)$

$$\begin{aligned} \text{Therefore, } K_p^* &= \frac{|v_1, v_2|}{|t_1, t_2|} \\ &= \frac{|N c_U|}{|N c_U + \frac{\bar{n}}{2} (a_1 - a'_1) (a_1 - a'_1)|} \quad \text{from (1.65) and (1.7)} \\ &= \frac{|N c_U|}{|N c_U| + \frac{\bar{n}}{2} N^{p-1} \leq C_U (a_1 - a'_1) (a_1 - a'_1)} \quad \dots \quad (1.81) \end{aligned}$$

$C_U$  denoting the minor of  $c_U$  in  $|c_U|$

$$\text{Therefore from (1.15) } K_p^* = \frac{1}{1 + \frac{\bar{n}}{2N} D^2} \quad \dots \quad (1.82)$$

Denoting by  $\phi_p$  the angle  $\angle OF_p$ , i.e. the angle between the unit vector  $i$ , and the flat  $(t_1, t_2, \dots, t_p)$  we have, since  $k_p = \sin \phi_p$

$$D^2 = \frac{2N}{\bar{n}} \cot^2 \phi_p \quad \dots \quad (1.83)$$

If a statistic similar to  $D^2$  were constructed for the first  $i$  variates only, then we could denote it by  $D_i^2$ ; we should then have had

$$K_i^* = \frac{1}{1 + \frac{\bar{n}}{2N} D_i^2} \quad \dots \quad (1.84)$$

and

$$D_i^2 = \frac{2N}{\bar{n}} \cot^2 \phi_i \quad \dots \quad (1.85)$$

It should be noted that consistently with this notation  $D^2$  should be preferably replaced by  $D_p^2$ . Let  $M$  be the foot of the perpendicular from  $F_p$  to  $t_i$  and let  $v$  be the angle  $F_p OM$ .

§2. DERIVATION OF THE SAMPLING DISTRIBUTION OF  $D^2$ 

It was shown in section 2 of the earlier paper<sup>1</sup> that  $D^2$  and  $\Delta^2$  defined in the present note by equations (1.15), (1.2), and  $\sum_{i=1}^p \sum_{j=1}^p a^{ij} a_{ij}$ ,  $\sum_{i=1}^p \sum_{j=1}^p a^{ij} (a_i - a'_i) (a_i - a'_i)$  occurring in the present note in the relation (1.75) are all invariant under any linear transformation. It was further shown in section 8 of the earlier paper<sup>1</sup> that a special linear transformation could be constructed such that

$$\left. \begin{aligned} a_{ij} &= 0, \text{ if } i \neq j, a_{ii} = 1 \\ (a_i - a'_i) &= \bar{n} \Delta^2 \\ (a_i - a'_i) &= 0 \text{ for } i = 2, 3, \dots, p \end{aligned} \right\} \quad \dots \quad (2.1)$$

$$\text{Therefore, } a^{ij} = 0 \text{ if } i \neq j, a^{ii} = 1 \quad \dots \quad (2.11)$$

We may consider  $\Delta$  to be given by the positive root of  $\Delta^2$  and  $\sqrt{p}$  the positive root of  $p$ . Then by suitably naming the populations as first and second we can write

$$a_1 - a'_1 = \Delta \sqrt{p} \quad \dots (2.12)$$

Therefore without any loss of generality we can take the distribution in (1.75) as

$$\text{const} \times \epsilon \quad -\frac{1}{2} \bar{n} p \Delta^2 - \frac{1}{2} \sum_{i=1}^p g_{ii} - \frac{\bar{n}}{2} \sqrt{p} \Delta (a_1 - a'_1) \quad \Pi d(a_1 - a'_1) \Pi dz_{1k} \Pi dz'_{1k} \quad \dots (2.13)$$

The density factor in (2.13) does not contain any mean difference except the first. Therefore integrating out for all mean differences  $(a_i - a'_i)$ , ( $i = 2, 3, \dots, p$ ) we have the distribution in the form

$$\text{const} \times \epsilon \quad -\frac{1}{2} \bar{n} p \Delta^2 - \frac{1}{2} \sum_{i=1}^p g_{ii} - \frac{\bar{n}}{2} \sqrt{p} \Delta (a_1 - a'_1) \quad d(a_1 - a'_1) \Pi dz_{1k} \Pi dz'_{1k} \quad \dots (2.14)$$

Now it is clear that the volume element  $d(a_1 - a'_1) \Pi dz_{1k} \Pi dz'_{1k}$  is really the volume described by the points  $(A, T_1, T_2, \dots, T_p)$  in a space of  $(n + n' - 1)$  dimensions, when the points are given whatever freedom we please but subject to the restriction that the variates will have to lie within infinitesimal ranges round about the values occurring in the density factor in (2.14). Let us denote the lengths  $OT_1, OT_2, \dots, OT_p$  by  $l_1, l_2, \dots, l_p$ . It is now clear from the definitions of  $(a_1 - a'_1)$  and  $g_{ii}$  in (1.6) and (1.7) and from the figure in section 1 that

$$g_{ii} = l_i^2, \text{ and } (\sqrt{\bar{n}/2} (a_1 - a'_1)) = \text{projection of } OT_1 \text{ on } OY = l_1 \cos \phi_p \cos \psi \quad \dots (2.2)$$

The density factor in (2.14) now reduces to

$$\text{const} \times \epsilon \quad -\frac{1}{2} \bar{n} p \Delta^2 - \frac{1}{2} \sum_{i=1}^p l_i^2 - (\bar{n} p / 2)^{1/2} \Delta l_1 \cos \phi_p \cos \psi \quad \dots (2.21)$$

Give now the points  $(A, T_1, T_2, \dots, T_p)$  freedom to move subject to the restrictions that  $l_1$  lies between  $l_1$  and  $l_1 + dl_1$  ( $i = 1, 2, \dots, p$ ) and  $\phi_p$  and  $\psi$  lie respectively between  $\phi_p$  and  $\phi_p + d\phi_p$  and  $\psi$  and  $\psi + d\psi$ . Let this be done in two stages. It is clear that keeping the vectors  $(l_1, l_2, \dots, l_p)$  fixed the volume element described by  $A$  is

$$\text{const. } (\sin \phi_p)^{N-p-2} (\cos \phi_p)^{p-1} (\sin \psi)^{p-2} d\phi_p d\psi \quad \dots (2.22)$$

where  $N$  of course is  $n + n'$

Release now the points  $T_1, T_2, \dots, T_p$  subject to the restrictions mentioned before; the volume element described by these points is

$$\text{const. } (l_1, l_2, \dots, l_p)^{N-3} dl_1, dl_2, \dots, dl_p \quad \dots (2.23)$$

STUDENTISED D<sup>3</sup>-STATISTIC

Therefore, the joint volume element described by  $(\lambda, T_1, T_2, \dots, T_p)$  is

$$\text{const. } (\sin \varphi_p)^{N-p-2} (\cos \varphi_p)^{p-1} (\sin v)^{p-2} d\varphi_p d\psi (t_1, t_2, \dots, t_p)^{N-1} \Pi dt_i \quad \dots (2.24)$$

If we restrict now each of  $\varphi_p$  and  $v$  to lie between 0 and  $\pi/2$  we easily see from (2.21) and (2.24) that the distribution (2.14) now reduces to

$$\begin{aligned} & -\frac{1}{2} \bar{n} \beta \Delta^2 - \frac{1}{2} \sum_{i=1}^p t_i^2 \\ \text{const} \times c & \cosh (c \Delta t_i \cos \varphi_p \cos v) (\sin \varphi_p)^{N-p-2} (\cos \varphi_p)^{p-1} (\sin v)^{p-2} \\ & \times d\varphi_p d\psi (t_1, t_2, \dots, t_p)^{N-1} \prod_{i=1}^p dt_i \quad \dots (2.25) \end{aligned}$$

where  $c^2 = \bar{n} \beta / 2$  ... (2.255)

Integrating out for  $dt_1, dt_2, \dots, dt_p$ , we have

$$\begin{aligned} & -\frac{1}{2} \bar{n} \beta \Delta^2 - \frac{1}{2} t_1^2 \\ \text{const} \times c & \cosh (c \Delta t_1 \cos \varphi_p \cos v) (\sin \varphi_p)^{N-p-2} (\cos \varphi_p)^{p-1} (\sin v)^{p-2} \\ & \times d\varphi_p d\psi t_1^{N-1} dt_1 \quad \dots (2.26) \end{aligned}$$

It should be noted that  $t_1$  varies from 0 to  $\alpha$  and both  $\varphi_p$  and  $v$  vary from 0 to  $\pi/2$

Using now the well known relation<sup>(1)</sup>

$$I_\nu(z) = \frac{2(z/2)^\nu}{\Gamma(\nu + \frac{1}{2}) \Gamma(\frac{1}{2})} \int_0^{\pi/2} \cosh(z \cos \theta) \sin^{2\nu} \theta d\theta \quad \dots (2.8)$$

which is valid for  $R(\nu + \frac{1}{2}) > 0$ , for the special case  $v=0$ , and integrating (2.26) over  $v$  from 0 to  $\pi/2$  the distribution reduces to

$$\text{const} \times c \int_0^\infty t_1^{N-2-(p-2)/2} I_{\frac{p-2}{2}}(c \Delta t_1 \cos \varphi_p) (\sin \varphi_p)^{N-p-2} (\cos \varphi_p)^{p/2} dt_1 d\varphi_p \quad \dots (2.35)$$

Using now another well known relation<sup>(2)</sup>

$$\int_0^\infty I_\nu(at) e^{-bt} t^{\mu-1} dt = \frac{\Gamma(\frac{1}{2} \nu + \frac{1}{2} \mu)}{2^{\nu} \Gamma(\nu+1)} {}_1F_1\left(\frac{1}{2} \nu + \frac{1}{2} \mu, \nu+1, \frac{a^2}{4b^2}\right) \quad \dots (2.4)$$

for the special case  $b^2 = \frac{1}{2}$ ,  $\nu = (p-2)/2$ ,  $\mu = N-1-(p-2)/2$ ,  $a = c \Delta \cos \varphi_p$ , which do not violate the restrictions on the validity of the formula, and integrating (2.35) over  $t_1$  from 0 to  $\infty$ , the distribution of  $\varphi_p$  is obtained in the form

$$\text{const} \times (\sin \varphi_p)^{N-p-2} (\cos \varphi_p)^{p-1} {}_1F_1\left(\frac{N-1}{2}, \frac{\beta}{2}, \frac{c^2 \Delta^2 \cos^2 \varphi_p}{2}\right) d\varphi_p \quad \dots (2.45)$$

But remembering from (1.83) and (2.255) that

$$D^2 = \frac{N}{c^2} \cot^2 \varphi_p \quad \dots (2.46)$$

and therefore,

$$\left. \begin{aligned} d\varphi_p &= \frac{c \sqrt{N} \cdot d D^2}{2 (D^2)^{1/2} (N + c^2 D^2)} \\ \cos \varphi_p &= \frac{c (D^2)^{1/2}}{(N + c^2 D^2)^{1/2}}, \quad \sin \varphi_p = \frac{N}{(N + c^2 D^2)^{1/2}} \end{aligned} \right\} \quad \dots (2.51)$$

we have the distribution of  $D^2$  given in the form

$$\text{const.} \cdot \frac{(D^2)^{(p-3)/2}}{(N + c^2 D^2)^{p/2}} \cdot \frac{N - p - 2}{2} {}_1F_1 \left( \frac{N-1}{2}, \frac{p}{2}, \frac{c^2 D^2}{2} \frac{D^2}{N + c^2 D^2} \right) d(D^2) \quad \dots (2.55)$$

#### BIBLIOGRAPHY.

1. BOSE, R. C. AND ROY, S. N.: The Distribution of the Studentised  $D^2$ -Statistic. The Proceedings of the Indian Statistical Conference, Calcutta, 1938, pp. 19-38.
2. WATSON, G. N.: Theory of Bessel Functions, p. 79.
3. Loc. Cit., p. 100 and p. 393.

[Paper received: 22nd June, 1939].