ON A STATISTICAL PROBLEM ARISING IN THE CLASSIFICATION OF AN INDIVIDUAL INTO ONE OF TWO GROUPS[1]

By Abraham Wald

Columbia University

1. Introduction. In social, economic and industrial problems we are often confronted with the task of classifying an individual into one of two groups on the basis of a number of test scores. For example, in the case of personnel selection the acceptance or rejection of an applicant is frequently based on a number of test scores obtained by the applicant. A similar situation arises in connection with college entrance examinations. Again, on the basis of a number of test scores, the admission or rejection of a student has to be decided. In all such problems it is assumed that there are two populations, say π_1 and π_2, one representing the population of individuals fit, and the other the population of individuals unfit for the purpose under consideration. The problem is that of classifying an individual into one of the populations π_1 and π_2 on the basis of his test scores. Often, some statistical data from past experience are available which can be utilized in making the classification. Suppose that from past experience we have the test scores of N_1 individuals who *are known* to belong to population π_1, and also the test scores of N_2 individuals who *are known* to belong to population π_2. These data will be utilized in classifying a new individual on the basis of his test scores.

In this paper we shall deal with the statistical problem of classifying an individual into one of the populations π_1 and π_2 on the basis of his test scores and on the basis of past experience, given in the form of two samples, one drawn from π_1 and the other from π_2. In the next section we give a precise formulation of the statistical problem and state the assumptions we make about the populations π_1 and π_2.

2. Statement of the problem. We consider two sets of p variates $(x_1, \cdots, x_p)$ and $(y_1, \cdots, y_p)$. It is assumed that each of the sets $(x_1, \cdots, x_p)$ and $(y_1, \cdots, y_p)$ has a p-variate normal distribution and the two sets are independent of each other. It is furthermore assumed that the covariance matrix of the variates $x_1, \cdots, x_p$ is equal to the covariance matrix of the variates $y_1, \cdots, y_p$, i.e. $\sigma_{x_i x_j} = \sigma_{y_i y_j}$ $(i, j = 1, \cdots, p)$. We will denote this common covariance by σ_{ij}. Let us denote the mean value of x_i by μ_i and the mean value of y_i by ν_i. Furthermore we will denote the normal population with mean values $\mu_1, \cdots, \mu_p$ and covariance matrix $\| \sigma_{ij} \|$ by π_1, and the normal population with mean values $\nu_1, \cdots, \nu_p$ and covariance matrix $\| \sigma_{ij} \|$ by π_2.

A sample of size N_1 is drawn from the population π_1 and a sample of size N_2 is

[1] The author wishes to thank Dr. Irving Lorge, Columbia University, for calling his attention to this problem.

drawn from the population π_2. Denote by $x_{i\alpha}$ the α-th observation on x_i ($i = 1, \cdots, p; \alpha = 1, \cdots, N_1$) and $y_{i\beta}$ the β-th observation on y_i ($i = 1, \cdots, p$; $\beta = 1, \cdots, N_2$). Let z_i ($i = 1, \cdots, p$) be a single observation on the i-th variate drawn from a p-variate population π, where it is known a priori that π is either identical with π_1 or with π_2. The set $(z_1, \cdots, z_p)$ is assumed to be distributed independently of $(x_1, \cdots, x_p)$ and $(y_1, \cdots, y_p)$.

We will deal here with the following statistical problem: On the basis of the observations $x_{i\alpha}, y_{i\beta}, z_i$ ($i = 1, \cdots, p; \alpha = 1, \cdots, N_1; \beta = 1, \cdots, N_2$) we test the hypothesis H_1 that the population π, from which the set $(z_1, \cdots, z_p)$ has been drawn, is equal to π_1. The parameters $\mu_1, \cdots, \mu_p, \nu_1, \cdots, \nu_p$ and $|| \sigma_{ij} ||$ are assumed to be unknown.

3. The statistic to be used for testing the hypothesis H_1. In this problem there exists only a single alternative hypothesis to the O-hypothesis H_1 to be tested, i.e. the hypothesis H_2 that π is equal to π_2. If the parameters $\mu_1, \cdots, \mu_p, \nu_1, \cdots, \nu_p$ and $|| \sigma_{ij} ||$ were known we could easily find (on the basis of a lemma by Neyman and Pearson) the critical region which is most powerful with respect to the alternative H_2. Let us assume for the moment that the parameters $\mu_1, \cdots, \mu_p, \nu_1, \cdots, \nu_p$ and $|| \sigma_{ij} ||$ are known and let us compute the critical region for testing H_1 which is most powerful with respect to the alternative H_2. According to a lemma by Neyman and Pearson[2] this critical region is given by the inequality

$$\frac{p_2(z_1, \cdots, z_p)}{p_1(z_1, \cdots, z_p)} \geq k, \tag{1}$$

where $p_1(z_1, \cdots, z_p)$ denotes the joint probability density function of $z_1, \cdots, z_p$ under the hypothesis H_1, $p_2(z_1, \cdots, z_p)$ denotes the joint probability density function of $(z_1, \cdots, z_p)$ under the hypothesis H_2, and k is a constant determined so that the critical region should have the required size.

Denote the determinant value $| \sigma_{ij} |$ of the matrix $|| \sigma_{ij} ||$ by σ^2. Then

$$p_1(z_1, \cdots, z_p) = \frac{1}{(2\pi)^{p/2}\sigma} e^{-\frac{1}{2}\sum_{j=1}^{p}\sum_{i=1}^{p}\sigma^{ij}(z_i-\mu_i)(z_j-\mu_j)}, \tag{2}$$

and

$$p_2(z_1, \cdots, z_p) = \frac{1}{(2\pi)^{p/2}\sigma} e^{-\frac{1}{2}\sum_{j=1}^{p}\sum_{i=1}^{p}\sigma^{ij}(z_i-\nu_i)(z_j-\nu_j)}, \tag{3}$$

where the matrix $|| \sigma^{ij} ||$ denotes the inverse matrix of the matrix $|| \sigma_{ij} ||$. Taking logarithms of both sizes of the inequality (1), we obtain the inequality

$$-\tfrac{1}{2}\Big\{\sum_j \sum_i \sigma^{ij}[(z_i - \nu_i)(z_j - \nu_j) - (z_i - \mu_i)(z_j - \mu_j)]\Big\} \geq \log k. \tag{4}$$

[2] J. NEYMAN and E. S. PEARSON, "Contributions to the theory of testing statistical hypotheses," *Stat. Res. Mem.*, Vol. 1, London, 1936.

Multiplying both sides of (4) by 2, we have

$$\sum_j \sum_i \sigma^{ij}[(z_i - \mu_i)(z_j - \mu_j) - (z_i - \nu_i)(z_j - \nu_j)] \geq 2 \log k. \tag{5}$$

The critical region (5) is most powerful with respect to the alternative H_2, but it cannot be used for our purposes since the parameters $\mu_1, \cdots, \mu_p, \nu_1, \cdots, \nu_p$ and $\| \sigma_{ij} \|$ are unknown. The optimum estimate of σ_{ij} on the basis of the observations $x_{i\alpha}$ and $y_{i\beta}$ is given by the sample covariance

$$s_{ij} = \frac{\sum_{\alpha=1}^{N_1} (x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j) + \sum_{\beta=1}^{N_2} (y_{i\beta} - \bar{y}_i)(y_{j\beta} - \bar{y}_j)}{N_1 + N_2 - 2} \tag{6}$$

where $\bar{x}_i = \frac{\sum_\alpha x_{i\alpha}}{N_1}$ and $\bar{y}_i = \frac{\sum_\beta y_{i\beta}}{N_2}$. The optimum estimates of μ_i and ν_i are given by $\bar{x}_i$ and $\bar{y}_i$ respectively ($i = 1, \cdots, p$). Hence for testing H_1 it seems reasonable to use the statistic R which we obtain from the left hand side of (5) by substituting the optimum estimates for the unknown parameters. Thus R is given by

$$R = \sum_j \sum_i s^{ij}[(z_i - \bar{x}_i)(z_j - \bar{x}_j) - (z_i - \bar{y}_i)(z_j - \bar{y}_j)], \tag{7}$$

where $\| s^{ij} \| = \| s_{ij} \|^{-1}$. The critical region for testing H_1 is given by the inequality

$$R \geq C, \tag{8}$$

where C is a constant determined in such a way that the critical region should have the required size. It is interesting to notice that R is proportional to the difference $T_1^2 - T_2^2$ where T_i ($i = 1, 2$) denotes the generalized Student's ratio[3] for testing the hypothesis that the set $(z_1, \cdots, z_p)$ is drawn from the population π_i. In our case the statistic T_1 cannot be used for testing H_1, since T_1 is appropriate for this purpose if the class of alternative hypotheses contains *all* p-variate normal populations having the same covariance matrix as π_1. In our case the class of alternatives consists merely of a single alternative, namely, the alternative π_2.

For the sake of certain simplifications we shall propose the use of a statistic U which differs slightly from the statistic R. In order to obtain U, we consider the inequality (5). Since $\sigma^{ij} = \sigma^{ji}$ this inequality can be reduced to

$$\sum_j \sum_i \sigma^{ij} z_i(\nu_j - \mu_j) \geq k', \tag{9}$$

where k' denotes a certain constant. The statistic U is obtained from the left hand side of (9) by substituting the optimum estimates for the unknown para-

[3] See. in this connection H. HOTELLING, "The generalization of Student's ratio," *Annals of Math. Stat.*, Vol. 2, and R. C. BOSE and S. N. ROY, "The exact distribution of the Studentized D^2 statistic," *Sankhya*, Vol. 3.

meters. Thus

$$U = \Sigma\Sigma s^{ij} z_i(\bar{y}_j - \bar{x}_j), \tag{10}$$

and the critical region is given by the inequality

$$U \geq d, \tag{11}$$

where the constant d is chosen so that the critical region should have the required size. The statistic U differs from R merely by a term which does not depend on the quantities $z_1, \cdots, z_p$. If N_1 and N_2 are large the difference $U - R$ is practically constant and therefore the critical regions (8) and (11) are identical. The use of U seems to be as justifiable as that of R and because of certain simplifications we propose the use of the critical region (11).

The statistic U is closely connected with the so called discriminant function[4] introduced by R. A. Fisher for discriminating between the two populations π_1 and π_2. The discriminant function D is given by

$$D = b_1 d_1 + b_2 d_2 + \cdots + b_p d_p \tag{12}$$

where $d_i = \bar{y}_i - \bar{x}_i$ and the coefficient b_i is proportional to $\sum_{j=1}^{p} s^{ij} d_j$. The coefficients $b_1, \cdots, b_p$ are called the coefficients of the discriminant function. We see that U is proportional to the statistic $\sum_{i=1}^{p} b_i z_i$ which is obtained from the right hand side of (12) by substituting z_i for d_i.

4. Solution of the problem when N_1 and N_2 are large. Denote by $F(U, N_1, N_2 \mid \pi_i)$ the cumulative probability distribution of U under the hypothesis that the set $(z_1, \cdots, z_p)$ has been drawn from the population π_i $(i = 1, 2)$. If N_1 and N_2 approach infinity the distribution $F(U, N_1, N_2 \mid \pi_i)$ converges to a normal distribution, since the variates s_{ij}, $\bar{x}_i$ and $\bar{y}_i$ converge stochastically to the constants σ_{ij}, μ_i and ν_i respectively $(i, j = 1, \cdots, p)$. Let us denote $\lim_{N_1 = N_2 = \infty} F(U, N_1, N_2 \mid \pi_i)$ by $\Phi(U \mid \pi_i)$ $(i = 1, 2)$. Furthermore denote by α_i the mean value, and by σ_i the standard deviation of the distribution $\Phi(U \mid \pi_i)$ $(i = 1, 2)$. It is obvious that $\sigma_1 = \sigma_2 = \sigma$ (say). It is easy to verify that the variates

$$\bar{\alpha}_1 = \Sigma\Sigma s^{ij} \bar{x}_i(\bar{y}_j - \bar{x}_j), \tag{13}$$

$$\bar{\alpha}_2 = \Sigma\Sigma s^{ij} \bar{y}_i(\bar{y}_j - \bar{x}_j), \tag{14}$$

$$\begin{aligned} \bar{\sigma}^2 &= \sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p} s^{ik} s^{jl} (\bar{y}_k - \bar{x}_k)(\bar{y}_l - \bar{x}_l) s_{ij} \\ &= \sum_{k=1}^{p}\sum_{l=1}^{p} s^{kl} (\bar{y}_k - \bar{x}_k)(\bar{y}_l - \bar{x}_l), \end{aligned} \tag{15}$$

converge stochastically to the constants α_1, α_2 and σ^2 respectively.

[4] R. A. Fisher, "The statistical utilization of multiple measurements," *Annals of Eugenics*, 1938.

Hence for large values of N_1 and N_2 we can assume that U is normally distributed with mean value $\bar{\alpha}_i$ and standard deviation $\bar{\sigma}$ if the hypothesis H_i ($i = 1, 2$) is true. Thus the critical region for testing H_1 is given by the inequality

$$U \geq \bar{\alpha}_1 + \lambda\bar{\sigma}, \tag{16}$$

where the constant λ is chosen in such a way that $\frac{1}{\sqrt{2\pi}} \int_{\lambda}^{\infty} e^{-t^2/2}\, dt$ is equal to the required size of the critical region.

Finally, some remarks about the proper choice of the size of the critical region may be of interest. Two kinds of error may be committed. H_1 may be rejected when it is true, and H_1 may be accepted when H_2 is true. Suppose that W_1 and W_2 are two positive numbers expressing the importance of an error of the first kind and an error of the second kind respectively. If the purpose of the statistical investigation is given it will usually be possible to determine the values of W_1 and W_2. We shall deal here with the question of determining the size of the critical region as a function of the weights W_1 and W_2. Denote by P_i the probability that (16) holds under the assumption that H_i is true ($i = 1, 2$). Then P_1 is the size of the critical region (also the probability of an error of the first kind), and $1 - P_2$ is the probability of an error of the second kind. Both probabilities P_1 and P_2 are functions of λ and are given by the following expressions:

$$P_1 = \frac{1}{\sqrt{2\pi}} \int_{\lambda}^{\infty} e^{-t^2/2}\, dt, \tag{17}$$

and

$$P_2 = \frac{1}{\sqrt{2\pi}} \int_{((\bar{\alpha}_1 - \bar{\alpha}_2)/\bar{\sigma})+\lambda}^{\infty} e^{-t^2/2}\, dt. \tag{18}$$

From (13) and (14) we obtain

$$\bar{\alpha}_2 - \bar{\alpha}_1 = \sum_j \sum_i s^{ij}(\bar{y}_i - \bar{x}_i)(\bar{y}_j - \bar{x}_j). \tag{19}$$

Since the right hand side of (19) is positive definite, we have $\bar{\alpha}_2 > \bar{\alpha}_1$. Hence because of (17) and (18) we also have $P_2 > P_1$. By the risk of committing a certain error we understand the probability of that error multiplied by its weight. Hence the risk of committing an error of the first kind is given by W_1P_1, and the risk of committing an error of the second kind is given by $W_2(1 - P_2)$. It seems reasonable to choose the value of λ so that the two risks become equal to each other, i.e. such that

$$W_1P_1 = W_2(1 - P_2). \tag{20}$$

Hence using (17) and (18) we obtain the following equation in λ

$$W_1 \frac{1}{\sqrt{2\pi}} \int_{\lambda}^{\infty} e^{-t^2/2}\, dt - W_2 \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{((\bar{\alpha}_1 - \bar{\alpha}_2)/\bar{\sigma})+\lambda} e^{-t^2/2}\, dt = 0. \tag{21}$$

Using a table of the normal distribution, the value of λ which satisfies the equation (21) can easily be found. For $W_1 = W_2$ the solution of (21) is given by

$$\lambda = \frac{\bar{\alpha}_2 - \bar{\alpha}_1}{2\bar{\sigma}},$$

and the critical region is given by the inequality

$$U \geq \bar{\alpha}_1 + \lambda\bar{\sigma} = \bar{\alpha}_1 + \frac{\bar{\alpha}_2 - \bar{\alpha}_1}{2} = \frac{\bar{\alpha}_1 + \bar{\alpha}_2}{2}.$$

5. Some results concerning the exact sampling distribution of the statistic U. If N_1 and N_2 are not large the solution given in section 4 cannot be used and it is necessary to derive the exact sampling distribution of U. Let

$$(\bar{y}_i - \bar{x}_i)\sqrt{\frac{N_1N_2}{N_1 + N_2}} = z_i' \qquad (i = 1, \cdots, p). \tag{22}$$

Then

$$U = \sqrt{\frac{N_1 + N_2}{N_1N_2}} \sum_i \sum_j s^{ij} z_i z_j' \tag{23}$$

where the variates $z_1', \cdots, z_p'$ are distributed independently of the set $(z_1, \cdots, z_p)$, the mean value of z_i' is equal to $(\nu_i - \mu_i)\sqrt{\frac{N_1N_2}{N_1 + N_2}}$ and the covariance between z_i' and z_j' is equal to σ_{ij}. It is known that the set of covariances s_{ij} is distributed independently of the set $(z_1, \cdots, z_p, z_1', \cdots, z_p')$ and therefore the distribution of U remains unchanged if instead of (6) we have

$$s_{ij} = \frac{\sum_{\alpha=1}^{n} t_{i\alpha}^2}{n} \qquad (n = N_1 + N_2 - 2), \tag{24}$$

where the variates $t_{i\alpha}$ are distributed independently of the set $(z_1, \cdots, z_p, z_1', \cdots, z_p')$, have a joint normal distribution with mean values zero, $\sigma_{t_{i\alpha}t_{j\alpha}} = \sigma_{ij}$ and $\sigma_{t_{i\alpha}t_{j\beta}} = 0$ if $\alpha \neq \beta$. It is necessary to derive the distribution of U under both hypotheses H_1 and H_2. In both cases the mean values of $z_1, \cdots, z_p, z_1', \cdots, z_p'$ are not zero. Instead of U we will consider the statistic

$$U' = \sum_{i=1}^{p} \sum_{j=1}^{p} s^{ij} z_i z_j'$$

which differs from U only in the proportionality factor $\sqrt{\frac{N_1 + N_2}{N_1N_2}}$. The distributions of U' under the hypotheses H_1 and H_2 are contained as special cases in the distribution of the statistic

$$V = \sum_j \sum_i s^{ij} t_{i,n+1} t_{j,n+2}, \tag{25}$$

where s_{ij} is given by (24) and the joint distribution of the variates $t_{i\beta}$ $(i = 1, \cdots, p; \beta = 1, \cdots, n + 2)$ is given by

$$(26) \quad \frac{1}{(2\pi)^{p(n+2)/2}\sigma^{n+2}} e^{-\frac{1}{2}\sum_{j=1}^{p}\sum_{i=1}^{p}\sigma^{ij}\left[\sum_{\alpha=1}^{n} t_{i\alpha}t_{j\alpha}+(t_{i,n+1}-\xi_i)(t_{j,n+1}-\xi_j)+(t_{i,n+2}-\eta_i)(t_{j,n+2}-\eta_j)\right]} \times \prod_{\beta=1}^{n+2}\sum_{i=1}^{p} dt_{i\beta}.$$

The quantities $\xi_1, \cdots, \xi_p, \eta_1, \cdots, \eta_p$ are constants and σ^2 denotes the determinant value of the matrix $\| \sigma_{ij} \|$.

We will deal here with the distribution of the statistic V given in (25) under the assumption that the joint distribution of the variates $t_{i\beta}$ $(i = 1, \cdots, p; \beta = 1, \cdots, n + 2)$ is given by (26).

In order to derive the distribution of V we shall have to prove several lemmas.

LEMMA 1. *Let $\| \lambda_{ij} \|$ $(i, j = 1, \cdots, p)$ be an arbitrary non-singular matrix, and let*

$$t'_{i\beta} = \sum_{j=1}^{p} \lambda_{ij} t_{j\beta} \qquad (i = 1, \cdots, p; \beta = 1, \cdots, n + 2).$$

Let furthermore s'_{ij} be given by

$$s'_{ij} = \frac{\sum_{\alpha=1}^{n} t'_{i\alpha} t'_{j\alpha}}{n}.$$

Then $\sum_j \sum_i s^{ij} t_{i,n+1} t_{j,n+2} = \sum_j \sum_i s'^{ij} t'_{i,n+1} t'_{j,n+2}$, i.e. the statistic V is invariant under non-singular linear transformations.

PROOF. We obviously have

$$(27) \quad t'_{i,n+1} t'_{j,n+2} = \sum_{k=1}^{p} \sum_{l=1}^{p} \lambda_{ik} \lambda_{jl} t_{k,n+1} t_{l,n+2}.$$

Furthermore we have

$$(28) \quad s'_{ij} = \sum_{k=1}^{p} \sum_{l=1}^{p} \lambda_{ik} \lambda_{jl} s_{kl}.$$

Hence

$$(29) \quad \| s'_{ij} \| = \| \lambda_{ij} \| \, \| s_{ij} \| \, \| \bar{\lambda}_{ij} \|$$

where $\bar{\lambda}_{ij} = \lambda_{ji}$.

From (29) we obtain

$$(30) \quad \| s'^{ij} \| = \| \bar{\lambda}^{ij} \| \, \| s^{ij} \| \, \| \lambda^{ij} \|,$$

and therefore

$$(31) \quad s'^{ij} = \sum_{k=1}^{p} \sum_{l=1}^{p} \lambda^{ki} \lambda^{lj} s^{kl}.$$

Hence from (27) and (31) we obtain

$$\sum_j \sum_i s'^{ij} t'_{i,n+1} t'_{j,n+2} = \sum_j \sum_i \sum_k \sum_l \sum_u \sum_v \lambda^{ki} \lambda^{lj} s^{kl} \lambda_{iu} \lambda_{jv} t_{u,n+1} t_{v,n+2} . \tag{32}$$

The coefficient of $t_{u,n+1}t_{v,n+2}$ on the right hand side of (32) is given by

$$\sum_j \sum_i \sum_k \sum_l \lambda^{ki} \lambda^{lj} s^{kl} \lambda_{iu} \lambda_{jv} = \sum_k \sum_l \{(\sum_i \lambda^{ki} \lambda_{iu})(\sum_j \lambda^{lj} \lambda_{jv}) s^{kl}\} = s^{uv}. \tag{33}$$

Lemma 1 follows from (32) and (33).

LEMMA 2. *The distribution of V remains unchanged if we assume that the covariance matrix $\| \sigma_{ij} \|$ is equal to the unit matrix, i.e. the joint distribution of the variates $t_{i\beta}$ ($i = 1, \cdots, p$; $\beta = 1, \cdots, n + 2$) is given by*

$$\frac{1}{(2\pi)^{p(n+2)/2}} e^{-\frac{1}{2}\left[\sum_{i=1}^{p} \sum_{\alpha=1}^{n} t_{i\alpha}^2 + \sum_i (t_{i,n+1} - \rho_i)^2 + \sum_i (t_{i,n+2} - \zeta_i)^2\right]}, \tag{34}$$

where the constants ρ_i and ζ_i are functions of the constants $\xi_1, \cdots, \xi_p, \eta_1, \cdots, \eta_p$ and of the σ_{ij}.

Lemma 2 is an immediate consequence of Lemma 1. Hence we have to derive the distribution of V under the assumption that the variates $t_{i\beta}$ have the joint distribution given in (34).

Let R_i ($i = 1, \cdots, p$) be the point of the $n + 2$ dimensional Cartesian space with the coordinates $t_{i1}, \cdots, t_{i,n+2}$. Let $P = (u_1, \cdots, u_{n+2})$ and $Q = (v_1, \cdots, v_{n+2})$ be two arbitrary points such that $\sum_{\beta=1}^{n+2} u_\beta v_\beta = 0$ and $\sum u_\beta^2 = \sum v_\beta^2 = 1$. Denote by 0 the origin of the coordinate system and let $\bar{t}_{i,n+1}$ be the projection of the vector $0R_i$ on the vector $0P$. We have

$$\bar{t}_{i,n+1} = \sum_{\beta=1}^{n+2} t_{i\beta} u_\beta \qquad (i = 1, \cdots, p). \tag{35}$$

Similarly, the projection $\bar{t}_{i,n+2}$ of the vector $0R_i$ on $0Q$ is given by

$$\bar{t}_{i,n+2} = \sum_{\beta=1}^{n+2} t_{i\beta} v_\beta . \tag{36}$$

Let $\bar{R}_i$ ($i = 1, \cdots, p$) be the projection of the point R_i on the n-dimensional hyperplane through 0 and perpendicular to the vectors $0P$ and $0Q$. Denote the coordinates of $\bar{R}_i$ by $r_{i1}, \cdots, r_{i,n+2}$ respectively and let $\bar{s}_{ij}$ be defined by

$$\bar{s}_{ij} = \frac{\sum_{\beta=1}^{n+2} r_{i\beta} r_{j\beta}}{n}. \tag{37}$$

If we rotate the coordinate system so that the $(n + 1)$-axis coincides with $0P$ and the $(n + 2)$-axis coincides with $0Q$, and if $\bar{t}_{i1}, \cdots, \bar{t}_{i,n+2}$ denote the coordinates of R_i ($i = 1, \cdots, p$) referred to the new system, then we have

$$\bar{s}_{ij} = \frac{1}{n} \sum_{\beta=1}^{n+2} r_{i\beta} r_{j\beta} = \frac{1}{n} \sum_{\alpha=1}^{n} \bar{t}_{i\alpha} \bar{t}_{j\alpha}, \text{ and} \tag{38}$$

$$\sum_{\beta=1}^{n+2} t_{i\beta} t_{j\beta} = \sum_{\beta=1}^{n+2} \bar{t}_{i\beta} \bar{t}_{j\beta} . \tag{39}$$

From (38) and (39) we obtain

$$\bar{s}_{ij} = \frac{\sum_{\beta=1}^{n+2} t_{i\beta} t_{j\beta} - \bar{t}_{i,n+1}\bar{t}_{j,n+1} - \bar{t}_{i,n+2}\bar{t}_{j,n+2}}{n}. \tag{40}$$

We will now prove

LEMMA 3. *Let $\bar{V}$ be defined by*

$$\bar{V} = \sum_j \sum_i \bar{s}^{ij}\bar{t}_{i,n+1}\bar{t}_{j,n+2}, \tag{41}$$

where $\bar{t}_{i,n+1}$, $\bar{t}_{i,n+2}$ and $\bar{s}_{ij}$ are given by the formulas (35), (36) *and* (40) *respectively. Let furthermore the joint probability distribution of the variates $t_{i\beta}$ $(i = 1, \cdots, p;\ \beta = 1, \cdots, n + 2)$ be given by*

$$\frac{1}{(2\pi)^{p(n+2)/2}} e^{-\frac{1}{2}\left[\sum_{i=1}^{p}\sum_{\beta=1}^{n+2}(t_{i\beta}-\rho_i u_\beta - \zeta_i v_\beta)^2\right]} \prod_i \prod_\beta dt_{i\beta}. \tag{42}$$

Then the distribution of $\bar{V}$ calculated under the assumption that the quantities $u_1, \cdots, u_{n+2}, v_1, \cdots, v_{n+2}$ are constants and the joint probability distribution of the variates $t_{i\beta}$ is given by (42), *is the same as the distribution of V calculated under the assumption that the joint probability distribution of the variates $t_{i\beta}$ is given by* (34).

PROOF. If we rotate the coordinate system so that the $(n + 1)$-axis coincides with $0P$ and the $(n + 2)$-axis coincides with $0Q$, and if $\bar{t}_{i1}, \cdots, \bar{t}_{i,n+2}$ denote the coordinates of R_i $(i = 1, \cdots, p)$ in the new system, then $\bar{t}_{i,n+1}$ and $\bar{t}_{i,n+2}$ are given by the right hand sides of (35) and (36) respectively. Furthermore

$$\bar{s}_{ij} = \frac{\sum_{\alpha=1}^{n} \bar{t}_{i\alpha}\bar{t}_{j\alpha}}{n}.$$

Hence the distribution of $\bar{V}$ is certainly the same as that of V if the joint probability distribution of the variates $\bar{t}_{i\beta}$ $(i = 1, \cdots, p;\ \beta = 1, \cdots, n + 2)$ is given by the expression which we obtain from (34) by substituting $\bar{t}_{i\beta}$ for $t_{i\beta}$. Thus, in order to prove Lemma 3 we have merely to show that if the variates $\bar{t}_{i\beta}$ have the joint probability distribution (34), the variates $t_{i\beta}$ have the joint probability distribution (42). Since the variates $t_{i1}, \cdots, t_{i,n+2}$ are obtained by an orthogonal transformation of the variates $\bar{t}_{i1}, \cdots, \bar{t}_{i,n+2}$, it follows that the variates $t_{i\beta}$ $(i = 1, \cdots, p;\ \beta = 1, \cdots, n + 2)$ are independently and normally distributed with unit variances. We have

$$t_{i\beta} = \sum_{\gamma=1}^{n+2} \lambda_{\beta\gamma}\bar{t}_{i\gamma} \tag{43}$$

where $\lambda_{\beta\gamma}$ is equal to the cosine of the angle between the β-th axis of the original system and γ-th axis of the new system. Since

$$\lambda_{\beta,n+1} = u_\beta \quad \text{and} \quad \lambda_{\beta,n+2} = v_\beta,$$

and since $E(\bar{t}_{i\gamma}) = 0$ for $\gamma = 1, \cdots, n$, $E(\bar{t}_{i,n+1}) = \rho_i$ and $E(\bar{t}_{i,n+2}) = \zeta_i$, it follows from (43) that

$$E(t_{i\beta}) = \rho_i u_\beta + \zeta_i v_\beta . \tag{44}$$

Hence Lemma 3 is proved.

We will now prove

LEMMA 4. *Let P be a point with the coordinates $u_1, \cdots, u_{n+2}$ and Q a point with the coordinates $v_1, \cdots, v_{n+2}$ such that $\Sigma u_\beta v_\beta = 0$ and $\Sigma u_\beta^2 = \Sigma v_\beta^2 = 1$. Denote by L_p the flat space determined by the vectors $0R_1, \cdots, 0R_p$ ($R_i = (t_{i1}, \cdots, t_{i,n+2})$) and let $\bar{P}$ be the projection of P on L_p and $\bar{Q}$ the projection of Q on L_p. Denote furthermore by θ_1 the angle between the vectors $0P$ and $0\bar{P}$, by θ_1' the angle between $0P$ and $0\bar{Q}$, by θ_2 the angle between $0Q$ and $0\bar{Q}$, by θ_2' the angle between $0Q$ and $0\bar{P}$, and finally by θ_3 the angle between $0\bar{P}$ and $0\bar{Q}$. Then the statistic $\bar{V}$ defined in (41) is equal to*

$$\bar{V} = -\frac{\begin{vmatrix} 0 & a_1 & a_2 \\ b_1 & a_{11} & a_{12} \\ b_2 & a_{12} & a_{22} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{vmatrix}}, \tag{45}$$

where

$$a_1 = \cos^2 \theta_1 ; \quad a_2 = \cos \theta_1' \cos \theta_2 ; \quad b_1 = \cos \theta_1 \cos \theta_2' ; \quad b_2 = \cos^2 \theta_2 ; \tag{46}$$

$$\begin{aligned} a_{11} &= \frac{\cos^2 \theta_1 - a_1^2 - b_1^2}{n}, \qquad a_{22} = \frac{\cos^2 \theta_2 - a_2^2 - b_2^2}{n} \\ \textit{and} \quad a_{12} &= \frac{\cos \theta_1 \cos \theta_2 \cos \theta_3 - a_1 a_2 - b_1 b_2}{n}. \end{aligned} \tag{47}$$

PROOF. If we rotate the coordinate system in such a way that the $(n + 1)$-axis coincides with $0P$ and the $(n + 2)$-axis coincides with $0Q$, and if $\bar{t}_{i1}, \cdots, \bar{t}_{i,n+2}$ are the coordinates of R_i in the new system, then

$$\bar{s}_{ij} = \frac{\sum_{\alpha=1}^{n} \bar{t}_{i\alpha} \bar{t}_{j\alpha}}{n}.$$

According to Lemma 1 the statistic V is invariant under linear transformations of the variables $t_{i\beta}$. Hence $\bar{V}$ is also invariant under linear transformations of the variables $\bar{t}_{i\beta}$. Thus the value of $\bar{V}$ remains unchanged if the points $R_1, \cdots, R_p$ are replaced by arbitrary points $R_1', \cdots, R_p'$ of L_p subject to the condition that the vectors $0R_1', \cdots, 0R_p'$ be linearly independent. Hence we may assume that the vectors $0R_3, \cdots, 0R_p$ are perpendicular to each other and lie in the intersection of L_p with the n-dimensional flat space which goes through 0 and is perpendicular to $0P$ and $0Q$. Furthermore we may assume that $R_1 = \bar{P}$ and $R_2 = \bar{Q}$. Then $0R_i$ is perpendicular to $0P$, $0Q$, $0R_1$ and $0R_2$ ($i = 3, \cdots, p$).

The statistic $\bar{V}$ can obviously be written in the form:

$$(48)\qquad \bar{V} = -\frac{\begin{vmatrix} 0 & \bar{t}_{1,n+1} & \cdots & \bar{t}_{p,n+1} \\ \bar{t}_{1,n+2} & \bar{s}_{11} & \cdots & \bar{s}_{1p} \\ \vdots & \vdots & & \vdots \\ \bar{t}_{p,n+2} & \bar{s}_{p1} & \cdots & \bar{s}_{pp} \end{vmatrix}}{\begin{vmatrix} \bar{s}_{11} & \cdots & \bar{s}_{1p} \\ \vdots & & \vdots \\ \bar{s}_{p1} & \cdots & \bar{s}_{pp} \end{vmatrix}}.$$

Because of our choice of the points $R_1, \cdots, R_p$, we have

$$(49)\qquad \bar{t}_{i,n+1} = \bar{t}_{i,n+2} = 0 \qquad (i = 3, \cdots, p)$$

and

$$(50)\qquad \sum_{\beta=1}^{n+2} \bar{t}_{i\beta}\bar{t}_{j\beta} = 0 \quad \text{if} \quad i \neq j \qquad (i = 3, \cdots, p; j = 1, \cdots, p).$$

From (49) and (50) it follows that $\bar{s}_{ij} = 0$ for $i \neq j$ except $\bar{s}_{12}$ which is not necessarily zero. Hence $\bar{V}$ reduces to the expression

$$(51)\qquad \bar{V} = -\frac{\begin{vmatrix} 0 & \bar{t}_{1,n+1} & \bar{t}_{2,n+1} \\ \bar{t}_{1,n+2} & \bar{s}_{11} & \bar{s}_{12} \\ \bar{t}_{2,n+2} & \bar{s}_{12} & \bar{s}_{22} \end{vmatrix}}{\begin{vmatrix} \bar{s}_{11} & \bar{s}_{12} \\ \bar{s}_{12} & \bar{s}_{22} \end{vmatrix}}.$$

We obviously have $\bar{t}_{1,n+1} = a_1$, $\bar{t}_{2,n+1} = a_2$, $\bar{t}_{1,n+2} = b_1$ and $\bar{t}_{2,n+2} = b_2$.

For any two points A and B denote the length of the vector AB by $\overline{AB}$. Since $n\bar{s}_{11} + (\bar{t}_{1,n+1})^2 + (\bar{t}_{1,n+2})^2 = \overline{0P}^2$, $n\bar{s}_{22} + (\bar{t}_{2,n+1})^2 + (\bar{t}_{2,n+2})^2 = \overline{0Q}^2$ and $n\bar{s}_{12} + \bar{t}_{1,n+1}\bar{t}_{2,n+1} + \bar{t}_{1,n+2}\bar{t}_{2,n+2} = \overline{0P}\cdot\overline{0Q}\cdot\cos\theta_3$, we can easily verify that $\bar{s}_{11} = a_{11}$, $\bar{s}_{12} = a_{12}$ and $\bar{s}_{22} = a_{22}$. Hence Lemma 4 is proved.

The angles θ_1' and θ_2' can be expressed in terms of the angles θ_1, θ_2 and θ_3. In order to show this, let us rotate the coordinate system so that the first p coordinates lie in the flat space L_p defined in Lemma 4. Let $u_1', \cdots, u_{n+2}'$ be the coordinates of P and $v_1', \cdots, v_{n+2}'$ the coordinates of Q referred to the new axes. Then, since $\overline{0P} = \overline{0Q} = 1$, we have

$$\cos\theta_1 = \sqrt{u_1'^2 + \cdots + u_p'^2}; \qquad \cos\theta_1' = \frac{u_1'v_1' + \cdots + u_p'v_p'}{\sqrt{v_1'^2 + \cdots + v_p'^2}};$$

$$\cos\theta_2 = \sqrt{v_1'^2 + \cdots + v_p'^2}; \qquad \cos\theta_2' = \frac{u_1'v_1' + \cdots + u_p'v_p'}{\sqrt{u_1'^2 + \cdots + u_p'^2}};$$

and

$$\cos\theta_3 = \frac{u_1'v_1' + \cdots + u_p'v_p'}{\sqrt{u_1'^2 + \cdots + u_p'^2}\sqrt{v_1'^2 + \cdots + v_p'^2}}.$$

Hence

$$\cos\theta_1' = \cos\theta_1\cos\theta_3 \quad \text{and} \quad \cos\theta_2' = \cos\theta_2\cos\theta_3.$$

Introducing the notations

$$m_1 = \cos^2 \theta_1 , \quad m_2 = \cos^2 \theta_2 \quad \text{and} \quad m_3 = \cos \theta_1 \cos \theta_2 \cos \theta_3 ,$$

we have

$$a_1 = m_1 , \qquad a_2 = m_3 , \qquad b_1 = m_3 , \qquad b_2 = m_2 ;$$

$$\begin{cases} a_{11} = \dfrac{m_1 - m_1^2 - m_3^2}{n} , & a_{12} = \dfrac{m_3(1 - m_1 - m_2)}{n} \\ \text{and} \qquad a_{22} = \dfrac{m_2 - m_2^2 - m_3^2}{n} & \end{cases}$$

Substituting the above values in (45) we obtain

$$\begin{aligned} \bar{V} &= -n \frac{m_3}{m_3^2 - 1 + m_1 + m_2 - m_1 m_2} \\ &= -n \frac{\cos \theta_1 \cos \theta_2 \cos \theta_3}{\cos^2 \theta_1 \cos^2 \theta_2 \cos^2 \theta_3 - \sin^2 \theta_1 \sin^2 \theta_2} . \end{aligned}$$

Hence, Lemma 4 can be written as

LEMMA 4′. *Let P be a point with the coordinates $u_1 , \cdots , u_{n+2}$ and Q a point with the coordinates $v_1 , \cdots , v_{n+2}$. Denote by L_p the flat space determined by the vectors $0R_1 , \cdots , 0R_p$ and let $\bar{P}$ be the projection of P on L_p and $\bar{Q}$ the projection of Q on L_p. Denote furthermore by θ_1 the angle between $0P$ and $0\bar{P}$, by θ_2 the angle between $0Q$ and $0\bar{Q}$ and by θ_3 the angle between $0\bar{P}$ and $0\bar{Q}$. Then the statistic $\bar{V}$ defined in* (41) *is equal to*

$$\bar{V} = -n \frac{\cos \theta_1 \cos \theta_2 \cos \theta_3}{\cos^2 \theta_1 \cos^2 \theta_2 \cos^2 \theta_3 - \sin^2 \theta_1 \sin^2 \theta_2} . \tag{45′}$$

If P is a point of the $(n + 1)$-axis and Q a point of the $(n + 2)$-axis, then $\bar{V}$ is identical with the statistic V given in (25). Hence we obtain the following

Geometric interpretation of the statistic V defined in (25). If θ_1 denotes the angle between the $(n + 1)$-axis and the flat space L_p determined by the vectors $0R_1 , \cdots , 0R_p$, θ_2 the angle between the $(n + 2)$-axis and the flat space L_p, and if θ_3 denotes the angle between the projections of the last two coordinate axes on L_p, then the statistic V is equal to the right hand side of (45′).

Denote by S the $2n + 1$-dimensional surface in the $2n + 4$-dimensional space of the variables $u_1 , \cdots , u_{n+2} , v_1 , \cdots , v_{n+2}$ defined by the following equations

$$\sum_{\beta=1}^{n+2} u_\beta^2 = \sum_{\beta=1}^{n+2} v_\beta^2 = 1; \qquad \sum_{\beta=1}^{n+2} u_\beta v_\beta = 0. \tag{52}$$

denote by C the $2n + 1$-dimensional volume of the surface S, i.e.

$$C = \int_S dS. \tag{53}$$

Now we will assume that $u_1, \cdots, u_{n+2}, v_1, \cdots, v_{n+2}$ are random variables and the joint probability distribution function is defined as follows: the point $(u_1, \cdots, u_{n+2}, v_1, \cdots, v_{n+2})$ is restricted to points of S and the probability density function of S is defined by

$$\frac{dS}{C}. \tag{54}$$

Hence for any subset A of S the probability of A is equal to the $2n + 1$-dimensional volume of A divided by the $2n + 1$-dimensional volume of S. It should be remarked that the probability density function (54) is identical with the probability density function we would obtain if we were to assume that $u_1, \cdots, u_{n+2}, v_1, \cdots, v_{n+2}$ are independently, normally distributed with zero means and unit variances and calculate the conditional density function under the restriction that $(u_1, \cdots, u_{n+2}, v_1, \cdots, v_{n+2})$ is a point of S.

LEMMA 5. *The probability distribution of $\bar{V}$ defined in* (41), *calculated under the assumption that the joint probability density of the variables $u_1, \cdots, u_{n+2}$, $v_1, \cdots, v_{n+2}$ $t_{i\beta}$ $(i = 1, \cdots, p; \beta = 1, \cdots, n + 2)$ is given by the product of* (54) *and* (42), *is the same as the distribution of the statistic V calculated under the assumption that the variables $t_{i\beta}$ have the joint probability density function given in* (34).

Lemma 5 is an immediate consequence of lemma 3.

LEMMA 6. *Let L_p be an arbitrary p-dimensional flat space in the $n + 2$ dimensional Cartesian space, and let M_p be the flat space determined by the first p coordinate axes. Assuming that the joint probability density function of u_β, v_β, $t_{i\beta}$ $(i = 1, \cdots, p; \beta = 1, \cdots, n + 2)$ is given by the product of* (54) *and* (42), *the conditional distribution of $\bar{V}$ calculated under the restriction that the points $R_1, \cdots, R_p$ lie in L_p, is the same as the conditional distribution of $\bar{V}$ calculated under the restriction that the points $R_1, \cdots, R_p$ lie in M_p. The point R_i denotes the point with the coordinates $t_{i1}, \cdots, t_{i,n+2}$.*

PROOF. Let P be the point with the coordinates $u_1, \cdots, u_{n+2}$ and let Q be the point with the coordinates $v_1, \cdots, v_{n+2}$. Let us rotate the coordinate system so that the first p axes lie in the flat space L_p. Denote the coordinates of P in the new system by $u'_1, \cdots, u'_{n+2}$, those of Q by $v'_1, \cdots, v'_{n+2}$, and those of R_i by $t'_{i1}, \cdots, t'_{i,n+2}$ $(i = 1, \cdots, p)$. Let S' be the surface defined by

$$\Sigma u'^2_\beta = \Sigma v'^2_\beta = 1 \quad \text{and} \quad \Sigma u'_\beta v'_\beta = 0. \tag{55}$$

It is clear that the surface S' is identical with the surface S defined in (52). It is furthermore clear that if the joint density function of $u_1, \cdots, u_{n+2}, v_1, \cdots,$ v_{n+2} is given by $\frac{dS}{C}$, the joint density function of $u'_1, \cdots, u'_{n+2}, v'_1, \cdots, v'_{n+2}$ is the same, i.e. it is given by $\frac{dS'}{C}$. It can readily be seen that for any given set of values $u'_1, \cdots, u'_{n+2}, v'_1, \cdots, v'_{n+2}$ the conditional joint probability density of the variates $t'_{i\beta}$ is given by the function obtained from (42) by substituting

$t'_{i\beta}$ for $t_{i\beta}$, u'_β for u_β and v'_β for v_β, provided that for any given set of values $u_1, \cdots, u_{n+2}, v_1, \cdots, v_{n+2}$ the joint conditional distribution of the variates $t_{i\beta}$ is given by (42). Hence, if the joint distribution of $u_1, \cdots, u_{n+2}$, $v_1, \cdots, v_{n+2}$ and $t_{i\beta}$ $(i = 1, \cdots, p; \beta = 1, \cdots, n + 2)$ is given by the product of (54) and (42), the joint probability density function of the variates u'_β, v'_β, $t'_{i\beta}$ $(i = 1, \cdots, p; \beta = 1, \cdots, n + 2)$ is obtained from that of u_β, v_β, $t_{i\beta}$ by substituting S' for S and $t'_{i\beta}$ for $t_{i\beta}$.

According to Lemma 4′, $\bar{V}$ can be expressed as a function of the angles θ_1, θ_2 and θ_3 defined in Lemma 4′. Each angle θ_k $(k = 1, 2, 3)$ can be expressed as a function of the variables $t_{i\beta}$, u_β, v_β. It is obvious that the value of θ_k remains unchanged if we substitute $t'_{i\beta}$ for $t_{i\beta}$, u'_β for u_β and v'_β for v_β. Hence also the value of $\bar{V}$ remains unchanged if we substitute $t'_{i\beta}$ for $t_{i\beta}$, u'_β for u_β and v'_β for v_β. Lemma 6 is a consequence of this fact and of the fact that the joint probability density of the variates $t'_{i\beta}$, u'_β and v'_β is identical with that of the variates $t_{i\beta}$, u_β and v_β.

LEMMA 7. *Assuming that the joint probability distribution of the variates* $u_\beta, v_\beta, t_{i\beta}$ $(i = 1, \cdots, p; \beta = 1, \cdots, n + 2)$ *is given by the product of* (54) *and* (42), *the conditional joint probability distribution of* $u_1, \cdots, u_{n+2}, v_1, \cdots, v_{n+2}$, *calculated under the restriction that the points* $R_i = (t_{i1}, \cdots, t_{i,n+2})$ $(i = 1, \cdots, p)$ *lie in the flat space determined by the first* p *coordinate axes, is given by*

$$\frac{e^{-\frac{1}{2}\sum_{\gamma=p+1}^{n+2}\sum_{i=1}^{p}(\rho_i u_\gamma+\zeta_i v_\gamma)^2} f(u_1, \cdots, u_{n+2}, v_1, \cdots, v_{n+2})\, dS}{\int_S e^{-\frac{1}{2}\sum_{\gamma=p+1}^{n+2}\sum_{i=1}^{p}(\rho_i u_\gamma+\zeta_i v_\gamma)^2} f(u_1, \cdots, u_{n+2}, v_1, \cdots, v_{n+2})\, dS}, \tag{56}$$

where S *denotes the surface defined in* (52), *and* $f(u_1, \cdots, u_{n+2}, v_1, \cdots, v_{n+2})$ *denotes the expected value of*

$$\left(\begin{vmatrix} r_{11} & \cdots & r_{1p} \\ r_{21} & \cdots & r_{2p} \\ \vdots & & \vdots \\ r_{p1} & \cdots & r_{pp} \end{vmatrix}\right)^{\frac{n+2-p}{2}} \qquad \left(r_{ij} = \sum_{\alpha=1}^{p} t_{i\alpha} t_{j\alpha}\right) \tag{57}$$

calculated under the assumption that the joint distribution of the variates $t_{i\beta}$ *is given by* (42).

PROOF. Denote by $\bar{R}_i$ the projection of R_i on the flat space determined by the first p coordinate axes, i.e. $\bar{R}_i = (t_{i1}, \cdots, t_{ip}, 0, \cdots, 0)$. Let $\bar{l}_1$ be the length of $\bar{R}_1$, and let $\bar{l}_i$ be the distance of $\bar{R}_i$ from the flat space determined by the vectors $0\bar{R}_1, \cdots, 0\bar{R}_{i-1}$ $(i = 2, \cdots, p)$. Then, as is known,

$$\bar{l}_1 \bar{l}_2 \cdots \bar{l}_i = \sqrt{\begin{vmatrix} r_{11} & \cdots & r_{1i} \\ r_{21} & \cdots & r_{2i} \\ \cdot & \cdots & \cdot \\ r_{i1} & \cdots & r_{ii} \end{vmatrix}} \qquad (i = 1, \cdots, p), \tag{58}$$

where $r_{kl} = \sum_{\alpha=1}^{p} t_{k\alpha} t_{l\alpha}$.

We introduce the new variables

$$t^*_{i\gamma} = \frac{t_{i\gamma}}{\bar{l}_i} \qquad (i = 1, \cdots, p; \gamma = p + 1, \cdots, n + 2). \tag{59}$$

Then the joint probability density function of the variates u_β, v_β, $t_{i\alpha}$, $t^*_{i\gamma}$ $(i = 1, \cdots, p; \beta = 1, \cdots, n + 2, \alpha = 1, \cdots, p, \gamma = p + 1, \cdots, n + 2)$ is given by

$$\frac{(\bar{l}_1 \cdots \bar{l}_p)^{n+2-p}}{C(2\pi)^{p(n+2)/2}} e^{-\frac{1}{2}\left[\sum_{i=1}^{p} \sum_{\alpha=1}^{p} (t_{i\alpha}-\rho_i u_\alpha-\zeta_i v_\alpha)^2+\sum_{i=1}^{p} \sum_{\gamma=p+1}^{n+2} (\bar{l}_i t^*_{i\gamma}-\rho_i u_\gamma-\zeta_i v_\gamma)^2\right]} \times \left(\prod_i \prod_\alpha dt_{i\alpha}\right)\left(\prod_i \prod_\gamma dt^*_{i\gamma}\right) dS. \tag{60}$$

Substituting zero for $t^*_{i\gamma}$ $(i = 1, \cdots, p, \gamma = p + 1, \cdots, n + 2)$ in (60), we obtain an expression which is proportional to the conditional joint probability density of the variates u_β, v_β, $t_{i\alpha}$ $(\beta = 1, \cdots, n + 2; i = 1, \cdots, p, \alpha = 1, \cdots, p)$, calculated under the restriction that the points R_i $(i = 1, \cdots, p)$ fall in the flat space determined by the first p coordinate axes. Hence this conditional density function is given by

$$Ae^{-\frac{1}{2}\sum_{\gamma=p+1}^{n+2} \sum_{i=1}^{p} (\rho_i u_\gamma+\zeta_i v_\gamma)^2} (\bar{l}_1 \bar{l}_2 \cdots \bar{l}_p)^{n+2-p} \times e^{-\frac{1}{2}\left[\sum_{i=1}^{p} \sum_{\alpha=1}^{p} (t_{i\alpha}-\rho_i u_\alpha-\zeta_i v_\alpha)^2\right]} dS \prod_i \prod_\alpha dt_{i\alpha} \tag{61}$$

where A denotes a constant. The conditional distribution of the variates u_β, v_β $(\beta = 1, \cdots, n + 2)$ is obtained from (61) by integrating it with respect to the variables $t_{i\alpha}$ $(i = 1, \cdots, p; \alpha = 1, \cdots, p)$. Because of (58), we see that the resulting formula is identical with (56). Hence Lemma 7 is proved.

LEMMA 8. *Let* $m_1 = u_1^2 + \cdots + u_p^2$; $m_2 = v_1^2 + \cdots + v_p^2$, *and* $m_3 = u_1 v_1 + \cdots + u_p v_p$. *If the joint distribution of the variates* $u_1, \cdots, u_{n+2}$, $v_1, \cdots, v_{n+2}$ *is given by* (54), *then the joint distribution of* m_1, m_2, m_3 *is given by*

$$\frac{B}{\sqrt{m_1 m_2(1 - m_1)(1 - m_2)}} F_p(m_1)F_p(m_2)\Phi_p\left(\frac{m_3}{\sqrt{m_1 m_2}}\right)F_{n+2+p}(1 - m_1) \times F_{n+2-p}(1 - m_2)\Phi_{n+2-p}\left(\frac{-m_3}{\sqrt{(1 - m_1)(1 - m_2)}}\right) dm_1\, dm_2\, dm_3 \tag{62}$$

where B denotes a constant,

$$F_k(t) = \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} (t)^{(k-2)/2} e^{-\frac{1}{2}t} \quad \textit{and} \quad \Phi_k(t) = \frac{\Gamma\left(\frac{k}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{k-1}{2}\right)} (1 - t^2)^{(k-3)/2}. \tag{63}$$

PROOF. Let $m_1' = u_{p+1}^2 + \cdots + u_{n+2}^2$, $m_2' = v_{p+1}^2 + \cdots + v_{n+2}^2$,

$m_3' = u_{p+1}v_{p+1} + \cdots + u_{n+2}v_{n+2}$, $\bar{m}_3 = \dfrac{m_3}{\sqrt{m_1 m_2}}$ and $\bar{m}_3' = \dfrac{m_3'}{\sqrt{m_1' m_2'}}$. First we calculate the joint distribution of m_1, m_2, $\bar{m}_3$ m_1', m_2', $\bar{m}_3'$ under the assumption that $u_1, \cdots, u_{n+2}, v_1, \cdots, v_{n+2}$ are normally independently distributed with zero means and unit variances. This joint distribution is given by

$$\begin{aligned}(64)\qquad & F_p(m_1)F_p(m_2)\Phi_p(\bar{m}_3)F_{n+2-p}(m_1')F_{n+2-p}(m_2') \\ & \qquad \times \Phi_{n+2-p}(\bar{m}_3')\, dm_1\, dm_2\, d\bar{m}_3\, dm_1'\, dm_2'\, d\bar{m}_3'.\end{aligned}$$

Hence the joint distribution of m_1, m_2, m_3, m_1', m_2', m_3' is given by

$$\begin{aligned}(65)\qquad & \frac{1}{\sqrt{m_1 m_2 m_1' m_2'}} F_p(m_1)F_p(m_2)\Phi_p\left(\frac{m_3}{\sqrt{m_1 m_2}}\right) F_{n+2-p}(m_1')F_{n+2-p}(m_2') \\ & \qquad \times \Phi_{n+2-p}\left(\frac{m_3'}{\sqrt{m_1' m_2'}}\right) dm_1\, dm_2\, dm_3\, dm_1'\, dm_2'\, dm_3'.\end{aligned}$$

The required conditional distribution of m_1, m_2, m_3 is equal to the conditional distribution of m_1, m_2, m_3 obtained from the joint distribution (65) under the restrictions $m_1 + m_1' = 1$, $m_2 + m_2' = 1$ and $m_3 + m_3' = 0$. Hence if in (65) we substitute $1 - m_1$ for m_1', $1 - m_2$ for m_2' and $-m_3$ for m_3' we obtain an expression proportional to the conditional distribution of m_1, m_2, m_3. This proves Lemma 8.

LEMMA 9. *For any point* $(u_1, \cdots, u_{n+2}, v_1, \cdots, v_{n+2})$ *of the surface S defined in* (52) *the expected value of* (57) (*calculated under the assumption that* (42) *is the joint distribution of* $t_{i\beta}$) *is a function of* m_1 m_2, *and* m_3 *only, where* m_1, m_2 *and* m_3 *are defined in Lemma 8.*

PROOF. Let $\| \lambda_{\alpha\beta} \|$ $(\alpha, \beta = 1, \cdots, p)$ be an orthogonal matrix such that

$$(66)\qquad \lambda_{1\beta} = \frac{u_\beta}{\sqrt{u_1^2 + \cdots + u_p^2}} \qquad (\beta = 1, \cdots, p)$$

and

$$(67)\qquad \lambda_{2\beta} = \frac{u_\beta + \lambda v_\beta}{\sqrt{\sum_{\beta=1}^{p} (u_\beta + \lambda v_\beta)^2}} \qquad (\beta = 1, \cdots, p)$$

where

$$\lambda = \frac{-\sum_1^p u_\beta^2}{\sum_1^p u_\beta v_\beta}.$$

Let

$$(68)\qquad t_{i\alpha}' = \sum_{\beta=1}^{p} \lambda_{\alpha\beta} t_{i\beta} \qquad (\alpha = 1, \cdots, p).$$

Then the variates $t'_{i\alpha}$ are independently and normally distributed with unit variances. Since for any point of S, $E(t_{i\alpha}) = \rho_i u_\alpha + \zeta_i v_\alpha$, we have because of (66), (67) and (68)

$$E(t_{i\gamma}) = 0 \qquad (i = 1, \cdots, p, \gamma = 3, 4, \cdots, p),$$

$$E(t_{i1}) = \varphi_{i1}(m_1, m_2, m_3),$$

and $$E(t_{i2}) = \varphi_{i2}(m_1, m_2, m_3).$$

Hence the joint distribution of the variates $t'_{i\alpha}$ $(i = 1, \cdots, p; \alpha = 1, \cdots, p)$ depends merely on m_1, m_2 and m_3. Since $r_{ij} = \sum_{\alpha=1}^{p} t_{i\alpha} t_{j\alpha} = \sum_{\alpha=1}^{p} t'_{i\alpha} t'_{j\alpha}$, the expression (57) can be expressed as a function of the variables $t'_{i\alpha}$. Hence the distribution of the expression (57) depends merely on the parameters m_1, m_2, and m_3. This proves Lemma 9.

The main result of this section is the following

THEOREM. *Let V be the statistic given in* (25) *and let the joint distribution of the variates $t_{i\beta}$ $(i = 1, \cdots, p; \beta = 1, \cdots, n + 2)$ be given by* (34). *Then the probability distribution of V is the same as the distribution of*

$$-n \frac{m_3}{m_3^2 - (1 - m_1)(1 - m_2)} \tag{69}$$

where the joint distribution of m_1, m_2 and m_3 is equal to a constant multiple of the product of the following three factors: the expression (62), *the exponential* $e^{\frac{1}{2}(m_1 \sum \rho_i^2 + 2m_3 \sum \rho_i \zeta_i + m_2 \sum \zeta_i^2)}$ *and the expected value of*

$$\left(\begin{vmatrix} r_{11} & \cdots & r_{1p} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ r_{p1} & \cdots & r_{pp} \end{vmatrix}\right)^{(n+2-p)/2} \qquad \left(r_{ij} = \sum_{\alpha=1}^{p} t_{i\alpha} t_{j\alpha}\right). \tag{70}$$

The expected value of (70) *is calculated under the assumption that the variates $t_{i\alpha}$ are normally and independently distributed with unit variances and $E(t_{i\alpha}) = \rho_i u_\alpha + \zeta_i v_\alpha$ $(i = 1, \cdots, p; \alpha = 1, \cdots, p)$ where $\sum_{\alpha=1}^{p} u_\alpha^2 = m_1$, $\sum_{\alpha=1}^{p} v_\alpha^2 = m_2$ and $\sum_{\alpha=1}^{p} v_\alpha u_\alpha = m_3$. The domain of the variables m_1, m_2 and m_3 is given by the inequalities: $0 \leq m_1 \leq 1; 0 \leq m_2 \leq 1; -\sqrt{m_1 m_2} \leq m_3 \leq \sqrt{m_1 m_2}$.*

PROOF. First we note that the expected value of (70) is a function of m_1, m_2 and m_3 only. Let P be the point with the coordinates $u_1, \cdots, u_{n+2}$, and Q the point with the coordinates $v_1, \cdots, v_{n+2}$. Assume that the points $R_i = (t_{i1}, \cdots, t_{i,n+2})$ $(i = 1, \cdots, p)$ lie in the flat space determined by the first p coordinate axes. Assume furthermore that $u_1 v_1 + \cdots + u_{n+2} v_{n+2} = 0$ and that the lengths of the vectors $0P$ and $0Q$ are equal to 1. Then

$$\cos \theta_1 = \sqrt{u_1^2 + \cdots + u_p^2}; \qquad \cos \theta_2 = \sqrt{v_1^2 + \cdots + v_p^2}$$

and

$$\cos \theta_3 = \frac{u_1 v_1 + \cdots + u_p v_p}{\sqrt{u_1^2 + \cdots + u_p^2}\sqrt{v_1^2 + \cdots + v_p^2}},$$

where θ_1 denotes the angle between $0P$ and the flat space L_p determined by the vectors $0R_1, \cdots, 0R_p$; θ_2 denotes the angle between $0Q$ and L_p, and θ_3 denotes the angle between the projections of $0P$ and $0Q$ on L_p. According to Lemma 4′ the statistic $\bar{V}$ defined in (41) is equal to

$$\begin{aligned} \bar{V} &= -n \frac{\cos\theta_1 \cos\theta_2 \cos\theta_3}{\cos^2\theta_1 \cos^2\theta_2 \cos^2\theta_3 - \sin^2\theta_1 \sin^2\theta_2} \\ &= -n \frac{m_3}{m_3^2 - (1 - m_1)(1 - m_2)} \end{aligned} \tag{71}$$

where

$$m_1 = \cos^2\theta_1 = u_1^2 + \cdots + u_p^2, \qquad m_2 = \cos^2\theta_2 = v_1^2 + \cdots + v_p^2 \tag{72}$$

$$\text{and} \quad m_3 = \cos\theta_1 \cos\theta_2 \cos\theta_3 = u_1v_1 + \cdots + u_pv_p.$$

It follows from Lemmas 5 and 6 that the distribution of V is the same as the conditional distribution of $\bar{V}$ calculated under the assumption that the unconditional joint probability density of the variates u_β, v_β and $t_{i\beta}$ is given by the product of (54) and (42) and under the restriction that the points R_i $(i = 1, \cdots, p)$ fall in the flat space determined by the first p coordinate axes. Since $e^{-\frac{1}{2}\sum_{\gamma=p+1}^{n+2}\sum_{i=1}^{p}(\rho_i u_\gamma + \zeta_i v_\gamma)^2}$ is a constant multiple of

$$e^{\frac{1}{2}(m_1\sum\rho_i^2 + 2m_3\sum\rho_i\zeta_i + m_2\sum\zeta_i^2)} \tag{73}$$

from Lemmas 7, 8 and 9 it follows readily that the joint conditional distribution of $m_1 = u_1^2 + \cdots + u_p^2$, $m_2 = v_1^2 + \cdots + v_p^2$ and $m_3 = u_rv_1 + \cdots + u_pv_p$ is equal to a constant multiple of the product of (62), (73) and the expected value of 70. This proves our theorem.

It can be shown that the variates m_1, m_2 and m_3 are of the order $\frac{1}{n}$ in the probability sense. Hence

$$-n \frac{m_3}{m_3^2 - (1 - m_1)(1 - m_2)} = nm_3(1 + \epsilon) \tag{74}$$

where ϵ is of the order $\frac{1}{n}$. Hence we can say: *even for moderately large n the distribution of the statistic* $\bar{V}$ *is well approximated by the distribution of* nm_3, *where the joint distribution of* m_1, m_2 *and* m_3 *is equal to a constant multiple of the product of* (62), (73) *and the expected value of* (70).

If $n + 2 - p$ is an even integer, the expected value of (70) is obviously an elementary function of m_1, m_2 and m_3. Hence, if $n + 2 - p$ is even, the joint distribution of m_1, m_2 and m_3 is also an elementary function of m_1, m_2 and m_3.

If the constants ρ_i and ζ_i $(i = 1, \cdots, p)$ in formula (34) are equal to zero, the expected value of (70) is a constant and the joint distribution of m_1, m_2 and m_3 is given by (62).