163

THE SAMPLING DISTRIBUTION OF SOME STATISTICS OBTAINED FROM NON-LINEAR EQUATIONS

Author's Note (CMS 36. 237a)*

For nearly two decades prior to the date of this publication, the arithmetical procedure of the analysis of variance had been found in a rapidly expanding field of applications, to provide the most commodious approach to the problem of summarising thoroughly, and interpreting in critical fashion, the kinds of observational data available. In many cases the properties naturally postulated for the observations in question were such as to render the interpretation, and the standard tests of significance, mathematically exact; but this was not always so. New extensions were constantly being made, such, for example, as that implied by the discriminant functions, and it seemed to the author worth while, as in the opening section of this paper, to specify explicitly the conditions for the exactitude of the z-test; and in the succeeding sections to illustrate cases in which it must be inexact, though often presumably a good approximation, the limitations of which could not be specified without the solution of problems so far seemingly intractable.

The paper incorporates the solution of the simultaneous distribution of the latent roots which arise in discriminant analysis, without the formidable notation of matrix algebra. The method of resolution of this rather difficult problem may therefore be of interest in view of other possible applications.

Annals of Eugenics, 9: 238-249, (1939).

THE SAMPLING DISTRIBUTION OF SOME STATISTICS OBTAINED FROM NON-LINEAR EQUATIONS

By R. A. FISHER

1. The field of the z-test

It has long been recognized (Fisher, 1936) that that aspect of the analysis of variance which consists in comparing the mean square ascribed to some possible causes, or discrepancies, with an appropriate residual mean square, or error, is absolutely valid for normally distributed errors, subject to a certain limitation of the form in which the adjustable parameters are involved.

For example, to take a case of wide generality, in testing the goodness of fit of a regression line, or surface, having a given hypothetical form, no difficulty is introduced merely by reason of the line being curved, or the form non-linear. The idea that special difficulty is involved in non-linear regression is an illusion widely disseminated by the Pearsonian school, who were indeed completely at fault, even in testing the goodness of fit of linear regression.

If the form of the regression line be

$$Y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p,$$

where $X_1, \ldots, X_p$ are any functions of the independent variate x, we may minimize

$$S(y - Y)^2$$

for variations of $\beta_1, \ldots, \beta_p$, and obtain *linear* equations for these adjustable parameters,

$$SX_1(y - Y) = 0$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$SX_p(y - Y) = 0$$

having solutions $\beta_1 = b_1, \ldots, \beta_p = b_p$.

Multiplying by $b_1, \ldots, b_p$, and adding, it follows that, for the solution,

$$S(yY) = S(Y^2),$$

and therefore that

$$S(y - Y)^2 = S(y^2) - S(Y^2).$$

This is the residual sum of squares with $n - p$ degrees of freedom, where n is the number of degrees of freedom possessed by the set of values y, which may themselves be deviations from some simpler formula; while $S(Y^2)$ has p degrees of freedom. Then in general the analysis of variance

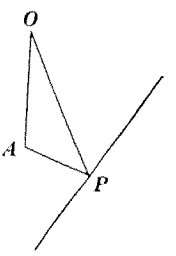
	Degrees of freedom	Sum of squares
Regression	p	$S(Y^2)$
Residual	$n-p$	$S(y-Y)^2$
Total	n	$S(y^2)$

supplies a direct and exact test of significance for the set of adjustments represented by b_1 to b_p.

For goodness of fit, where we have several observations y in each array, the most complicated form of curve is one which has as many variable constants as there are arrays. All such curves will pass through the array means, and all are indistinguishable, so far as such data are concerned. The goodness of fit of any curve involving $a-p$ restrictions on the array means may therefore be tested in a similar analysis:

	Degrees of freedom	Sum of squares
Deviation from fitted line	$a-p$	$\Sigma n_x(\bar{y}_x-Y)^2$
Within arrays	$n-a$	$S(y-\bar{y}_x)^2$
Total	$n-p$	$S(y-Y)^2$

It will be noticed that the expected values from which deviations are calculated in the second line are the array means, so that the vanishing of this sum of squares implies for all observations, $y=\bar{y}_x$. If y is used as a co-ordinate in generalized space, these conditions are satisfied in a plane continuum of a dimensions, within which lies a more restricted plane continuum of p dimensions representing all possible fitted curves of the form chosen. The observation point O will not in general lie within either space. The foot A of the perpendicular dropped from it on the a-space determines the observed means of the arrays, its length, OA, determines the sum of squares within arrays. The foot P of the perpendicular dropped on the p-space determines the fitted curve; its projection AP in the a-space gives the sum of squares of deviations. The test of goodness of fit consists in recognizing that, if this is unduly large compared with OA, the hypothetical form is incompatible with the data.



What is essential for the generality of the analysis is clearly not the linearity of the regression line, but the linearity of the relations among observation points for different sets of deviations to vanish simultaneously. Thus, if expectations are expressed in terms of parameters in any way, we may eliminate the parameters and obtain the relations which hold among the expectations of direct observations, equally liable to error. It is these which must be linear for the simple procedure of analysis of variance itself to supply exact tests of significance.

It is, of course, to be anticipated that the ordinary tests will still be sufficiently exact when the radii of curvature of the spaces concerned are all large compared with the distances. This condition has been recognized in the use of least square formulae for non-linear equations in astronomy and geodesy; the real importance of non-linearity does not lie, however, as has been supposed, in the fact that with equations non-linear in the parameters employed, the process of fitting may have to be iterated. This in itself presents no difficulty;

what is important is that after a good fit is obtained, and the sums of squares to be compared have been accurately evaluated, curvatures of the restriction spaces, if they are not small relative to the distances indicated by the sums of squares, may appreciably affect the frequency distribution of their ratios.

2 The test of significance of a harmonic component

Examples, in which the analysis is not too difficult, are rare; an early one was provided by the process of testing the reality of harmonic components in a series of equally spaced observations, u_r.

Supposing such a series to be composed of independent and normally distributed values, then any function

$$S(a_r u_r)$$

will have a mean value zero, if $S(a) = 0$, and will be normally distributed with the same variance as a single observation, if

$$S(a_r^2) = 1.$$

If the number in the series is odd, $2s+1$, then

$$a_r = \sqrt{\left(\frac{2}{2s+1}\right)} \cos \frac{2\pi pr}{2s+1},$$

and

$$a_r = \sqrt{\left(\frac{2}{2s+1}\right)} \sin \frac{2\pi pr}{2s+1}$$

for values of p from 1 to s, will constitute $2s$ mutually orthogonal components, occurring in pairs having periods $(2s+1)/p$ units. The like is true of an even number, $2s+2$, only in this case there is an unpaired component

$$a_r = \frac{1}{\sqrt{(2s+2)}} (-)^r u_r$$

of period 2 units. We may omit this component from consideration by deducting

$$\frac{1}{2s+2} (u_1 - u_2 + u_3 - \ldots - u_{2s+2})^2$$

from

$$S(u - \bar{u})^2,$$

so that the remainder consists of two components for each of s different periods.

The sum of two squares for each period constitutes a certain fraction of the whole. If we choose that period out of s available which makes the largest contribution, it has been shown (Fisher, 1929) that the fraction, g, taken by this period is distributed so that the probability of exceeding any value g is

$$P = s(1-g)^{s-1} - \frac{s(s-1)}{2}(1-2g)^{s-1} + \ldots + (-)^{k-1} \frac{s!}{(s-k)!\,k!}(1-kg)^{s-1}, \quad \ldots\ldots(1)$$

in which k is the greatest integer less than $1/g$.

Now, the choice of one period out of the s Fourier submultiples available, together with the evaluation of the two corresponding coefficients, gives that function

$$U = A\cos\frac{2\pi pr}{2s+1} + B\sin\frac{2\pi pr}{2s+1},$$

involving the adjustable constants A, B and p, which minimizes

$$S(u-U)^2$$

for integral values of p. If p were adjusted to some intermediate value it would presumably give a lower value still.

To find the mean of g as distributed in random samples, it is convenient to use the expression for $1-P$,

$$1-P = \sum_k (-)^{s-k}\frac{s!}{k!\,(s-k)!}(kg-1)^{s-1}, \qquad \ldots\ldots(2)$$

for values of k up to s for which $kg>1$.

Then the mean of g may be evaluated as

$$\int\frac{d}{dg}(1-P)\,g\,dg = [g(1-P)] - \int(1-P)\,dg,$$

in which the limits for term k are $1/k$ and 1.

The contribution of each term is therefore

$$(-)^{s-k}\frac{s!}{k!\,(s-k)!}\left\{(k-1)^{s-1} - \frac{1}{ks}(k-1)^s\right\},$$

which is to be summed for values of k from 2 to s.

Now, if u_k is a polynomial in k of degree less than s,

$$\sum_2^s (-)^{s-k}\frac{s!}{k!\,(s-k)!}u_k = (-)^s(su_1-u_0);$$

putting

$$u_k = (k-1)^{s-1} - \frac{1}{sk}\{(k-1)^s - (-1)^s\},$$

the mean is found to be

$$1 + \sum_2^s (-)^{k-1}\frac{s!}{k!\,(s-k)!}\frac{1}{sk}$$

$$= \frac{1}{s}\sum_1^s (-)^{k-1}\frac{s!}{k!\,(s-k)!}\frac{1}{k}.$$

This may be expressed as the definite integral

$$\frac{1}{s}\int_0^1 \{1-(1-x)^s\}\frac{dx}{x},$$

or

$$\frac{1}{s}\int_0^1 \{1+(1-x)+(1-x)^2+\ldots+(1-x)^{s-1}\}\,dx.$$

Hence
$$\bar{g} = \frac{1}{s}\left(1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{s}\right), \qquad \ldots\ldots(3)$$

or
$$\frac{1}{s}\{F(s)-F(0)\}.$$

The values of $\bar{g}$, of $3/2s$, and of $2s\bar{g}/3$ are given below for the first few values of s.

Table I. *Values of $\bar{g}$ for small samples*

Length of series	s	$\bar{g}$	$3/2s$	$2s\bar{g}/3$
5	2	3/4	3/4	1
7	3	11/18	1/2	1·2222
9	4	25/48	3/8	1·3889
11	5	137/300	3/10	1·5222
13	6	49/120	1/4	1·6333
15	7	363/980	3/14	1·7286

Instead of accounting for only the fraction $3/2s$ of the total sum of squares, as would be the case for a formula linear in three adjustable parameters, the fraction accounted for in large samples is about
$$\frac{\gamma+\log s}{s},$$
the ratio of which to $3/2s$ increases indefinitely as the size of the sample is increased; if $s = 12$ it is double, and at $s = 227$ about four times expectation.

Note that if p were fixed, and only A and B adjustable, the eliminant would be
$$u_{r-1}+u_{r+1} = 2u_r \cos\frac{2\pi p}{2s+1},$$
a linear equation between any three adjacent observations; whereas, when p also is adjustable, we have
$$\frac{u_{r-1}+u_{r+1}}{u_r} = \frac{u_r+u_{r+2}}{u_{r+1}},$$
for any four adjacent observations; and this eliminant, being non-linear, shows that the formula cannot be replaced by one linear in the three constants required.

3. The simultaneous distribution of latent roots

The most important case, for practical statistics, of an analysis derived from the solution of non-linear equations, is one which has been approached from different standpoints by several different groups of writers, notably Hotelling (1933, 1935, 1936), Mahalanobis (1930, 1936), Bose (1936) and Bose & Roy (1938), and which I have recently discussed under the heading of "The statistical utilization of multiple measurements" (Fisher, 1938).

The problem that emerges, when the effect of these researches is viewed in its mathematical generality may be stated as follows:

If $x_1, \ldots, x_i, \ldots, x_p$ stand for p different variates and a_{ij} for the sum of products of any two of them, an analysis of covariance will give values a_{ij} separately for several different distinguishable causes.

In connexion with any series a_{ij} corresponding with n_1 degrees of freedom we shall wish to consider the totals A_{ij} (of n_1+n_2 degrees of freedom) including the contributions of error, or of other causes, from which our chosen series may or may not be significantly distinguishable.

If, using arbitrary multipliers b^i, positive or negative, we make a compound variate

$$X = \sum_i b^i x_i,$$

then the sums of squares for X are

	D.F.	S.S.
Treatment	n_1	$\Sigma\Sigma\, b^i b^j a_{ij}$
Remainder	n_2	$\Sigma\Sigma\, b^i b^j (A_{ij} - a_{ij})$
Total	n_1+n_2	$\Sigma\Sigma\, b^i b^j A_{ij}$

and it is to be noted that the ratio of the sums of squares is stationary for all variations of b, if

$$\frac{\Sigma b^i a_{1i}}{\Sigma b^i A_{1i}} = \frac{\Sigma b^i a_{2i}}{\Sigma b^i A_{2i}} = \ldots = \frac{\Sigma b^i a_{ip}}{\Sigma b^i A_{ip}} = \theta.$$

Hence θ is one of the roots of the equation

$$| a_{ij} - \theta A_{ij} | = 0, \qquad \ldots\ldots(4)$$

which is at most of degree p. For the present we may suppose that n_1 and n_2 both exceed $p-1$.

Note that θ is also the ratio of the part to the whole of the sums of squares of X, so that all the roots of the equation lie in the range from 0 to 1. The fundamental problem is the sampling distribution of these roots considered simultaneously.

The simultaneous distribution of the sums of squares and products, of a couple of variates x and y, was given by the author (Fisher, 1915) in a form equivalent to

$$df = \frac{1}{\{\frac{1}{2}(N-3)\}!\{\frac{1}{2}(N-4)\}!\sqrt{\pi}} (\alpha_{11}\alpha_{22} - \alpha_{12}^2)^{\frac{1}{2}(N-1)} e^{-\alpha_{11}a_{11} - 2\alpha_{12}a_{12} - \alpha_{22}a_{22}} \times (a_{11}a_{22} - a_{12}^2)^{\frac{1}{2}(N-4)} da_{11}\, da_{12}\, da_{22}, \qquad \ldots\ldots(5)$$

in which the frequency element of the parent population is

$$\frac{\sqrt{(\alpha_{11}\alpha_{22} - \alpha_{12}^2)}}{4\pi} e^{-\alpha_{11}x_1^2 - 2\alpha_{12}x_1x_2 - \alpha_{22}x_2^2} dx_1\, dx_2.$$

Clearly N, the size of the bivariate sample, may in general be replaced by n_1+1, when n_1 is the number of degrees of freedom on which the estimates a_{11}, a_{12}, a_{22} are based.

If now a second sample yields values a'_{11}, a'_{12}, a'_{22}, we shall have a similar distribution in which these values are substituted for a_{11}, a_{12}, a_{22}, and n_2 for n_1. The frequency with which the two bivariate samples give values in the range

$$da_{11}da_{12}da_{22}da'_{11}da'_{12}da'_{22},$$

is given by the product of two such expressions.

With homogeneous samples we may argue in like manner from the totals, obtaining the total frequency of samples for which

$$a_{11}+a'_{11} = A_{11}, \quad a_{12}+a'_{12} = A_{12}, \quad a_{22}+a'_{22} = A_{22},$$

have given values. Noting also that, when a_{11}, a_{12}, a_{22} are fixed,

$$da'_{11} = dA_{11}, \quad da'_{12} = dA_{12}, \quad da'_{22} = dA_{22},$$

it follows that, given A_{11}, A_{12}, A_{22}, the simultaneous distribution of a_{11}, a_{12} and a_{22} is

$$df = \frac{\{\frac{1}{2}(n_1+n_2-2)\}!\,\{\frac{1}{2}(n_1+n_2-3)\}!}{\{\frac{1}{2}(n_1-2)\}!\,\{\frac{1}{2}(n_1-3)\}!\,\{\frac{1}{2}(n_2-2)\}!\,\{\frac{1}{2}(n_2-3)\}!\,\sqrt{\pi}} \times \frac{(a_{11}a_{22}-a_{12}^2)^{\frac{1}{2}(n_1-3)}\,(a'_{11}a'_{22}-a'^2_{12})^{\frac{1}{2}(n_2-3)}}{(A_{11}A_{22}-A_{12}^2)^{\frac{1}{2}(n_1+n_2-3)}}\,da_{11}da_{12}da_{22}. \quad \ldots\ldots(6)$$

This is the case of two variates of the more general distribution for p variates

$$df = \frac{\{\frac{1}{2}(n_1+n_2-2)\}!\ldots\{\frac{1}{2}(n_1+n_2-p-1)\}!}{\{\frac{1}{2}(n_1-2)\}!\ldots\{\frac{1}{2}(n_1-p-1)\}!\,\{\frac{1}{2}(n_2-2)\}!\ldots\{\frac{1}{2}(n_2-p-1)\}!\,\sqrt{\pi}} \times \frac{|a_{ij}|^{\frac{1}{2}(n_1-p-1)}\,|a'_{ij}|^{\frac{1}{2}(n_2-p-1)}}{|A_{ij}|^{\frac{1}{2}(n_1+n_2-p-1)}}\,da_{11}\ldots da_{pp}. \quad \ldots\ldots(7)$$

This very general distribution is derived by reasoning, exactly parallel with that above, from the distribution of a single set of variances and covariances

$$df = \frac{1}{\{\frac{1}{2}(n-2)\}!\ldots\{\frac{1}{2}(n-p-1)\}!\,\sqrt{\pi}}\,|\alpha|^{\frac{1}{2}n}e^{-\Sigma\Sigma\alpha x}\,|a|^{\frac{1}{2}(n-p-1)}da_{11}\ldots da_{nn},$$

where $|\alpha|$ and $|a|$ stand for the determinants of the α's and a's respectively. This is the generalization for p variates of the bivariate distribution given above. It may be derived either by the geometrical argument used by Wishart (1928), or as Hotelling has shown (1933, p. 51*a*), by an inductive chain, using the fact that the distribution of a partial correlation from which i variates have been eliminated, is exactly the same as that of a total correlation based on a sample smaller by i, and that such partial correlations must be distributed independently of the variates eliminated.

The distribution with which we are concerned is invariant for all linear transformations of the variates, so that we may at this point simplify the algebra by choosing

$$\begin{aligned} A_{11} &= 1, & A_{12} &= 0, & A_{22} &= 1, \\ a_{11} &= a, & a_{12} &= b, & a_{22} &= c, \\ a'_{11} &= 1-a, & a'_{12} &= -b, & a'_{22} &= 1-c. \end{aligned}$$

The roots of the equation θ, θ', then satisfy the symmetric relations

$$\theta+\theta' = a+c = p,$$

$$\theta\theta' = ac-b^2 = q,$$

whence

$$(1-\theta)(1-\theta') = (1-a)(1-c)-b^2 = 1-p+q,$$

$$(a-c)^2 = p^2-4q-4b^2,$$

and

$$\frac{\partial(p,q)}{\partial(a,c)} = a-c.$$

The simultaneous distribution of p, q and b is therefore

$$df = \frac{(n_1+n_2-2)!}{4\pi(n_1-2)!\,(n_2-2)!}q^{\frac{1}{2}(n_1-3)}(1-p+q)^{\frac{1}{2}(n_2-3)}dp\,dq\frac{2db}{\sqrt{(p^2-4q-4b^2)}}, \quad \ldots\ldots(8)$$

where the factor 2 has been inserted on the understanding that the integration is taken over the region in which a exceeds c.

For given p and q, b may take any value between the limits $\pm\sqrt{(\frac{1}{4}p^2-q)}$. Integrating between these limits, the last factor then is replaced by the constant, π.

To obtain the simultaneous distribution of the roots, note that

$$\frac{\partial(p,q)}{\partial(\theta,\theta')} = \theta-\theta',$$

giving

$$df = \frac{(n_1+n_2-2)!}{4(n_1-2)!\,(n_2-2)!}\theta^{\frac{1}{2}(n_1-3)}(1-\theta)^{\frac{1}{2}(n_2-3)}\theta'^{\frac{1}{2}(n_1-3)}(1-\theta')^{\frac{1}{2}(n_2-3)}(\theta-\theta')\,d\theta\,d\theta'. \quad \ldots\ldots(9)$$

For p variates, when p does not exceed n_1 or n_2, the general distribution of the p roots is, as might, apart from a factor involving p only, at this stage be expected,

$$df = \frac{\{\frac{1}{2}(n_1+n_2-2)\}!\ldots\{\frac{1}{2}(n_1+n_2-p-1)\}!}{\{\frac{1}{2}(n_1-2)\}!\ldots\{\frac{1}{2}(n_1-p-1)\}!\{\frac{1}{2}(n_2-2)\}!\ldots\{\frac{1}{2}(n_2-p-1)\}!}$$

$$\times\frac{\pi^{\frac{1}{2}p}}{\{\frac{1}{2}(p-2)\}!\ldots(-\frac{1}{2})!}\{\theta_1\ldots\theta_p\}^{\frac{1}{2}(n_1-p-1)}\{(1-\theta_1)\ldots(1-\theta_p)\}^{\frac{1}{2}(n_2-p-1)}$$

$$\times(\theta_1-\theta_2)\ldots(\theta_1-\theta_p)(\theta_2-\theta_3)\ldots(\theta_{p-1}-\theta_p)\,d\theta_1\ldots d\theta_p. \quad \ldots\ldots(10)$$

The general form, apart from the constant factor, may be demonstrated by using the transformation,

$$a_{ij} = \sum_k e_{ik}e_{jk}\theta_k,$$

which, if

$$\sum_i e_{ij}^2 = 1,\ \sum_i e_{ij}e_{il} = 0, \text{ when } j\neq k$$

satisfies the requirement that the sum of the principal minors of $|a|$ of degree s is equal to the sum of the products of θ taken s at a time. For example,

$$\sum_i a_{ii} = \sum_k \theta_k,$$

$$\sum\sum(a_{ii}a_{jj}-a_{ij}^2) = \sum_k\sum_l \theta_k\theta_l,$$

and so on.

We may now replace the $\frac{1}{2}p(p+1)$ variates a_{ij} by p variates θ and $\frac{1}{2}p(p-1)$ functionally independent values e. The Jacobian of this transformation is of degree $\frac{1}{2}p(p-1)$ in θ, and must consist, apart from a constant involving p, of the product of the $\frac{1}{2}p(p-1)$ differences, for it can be shown to contain such differences as $\theta_1-\theta_2$ as a factor.

If as variates we choose those e_{ij} for which $j>i$, then to satisfy the condition

$$\frac{\partial}{\partial e_{ij}}\sum_k e_{ik}e_{jk}=0,$$

we find

$$\frac{\partial e_{i1}}{\partial e_{12}}=-\frac{e_{12}}{e_{11}},$$

$$\frac{\partial e_{i2}}{\partial e_{12}}=\frac{e_{i1}}{e_{11}},$$

$$\frac{\partial e_{ij}}{\partial e_{12}}=0, \quad j>2,$$

whence

$$\frac{\partial}{\partial e_{12}}a_{ij}=\frac{e_{i2}e_{j1}}{e_{11}}(\theta_2-\theta_1).$$

For all values of i and j, this contains the factor $(\theta_1-\theta_2)$, which, therefore, divides the Jacobian. Similarly, this is true of every difference $(\theta_l-\theta_m)$, thus establishing the form of the distribution. The constant factor in the Jacobian cannot involve n_1 or n_2, so that, for any value of p, putting $n_1=n_2=p+1$, we may evaluate it by direct integration. An elementary evaluation of this function of p is given in the following section.

A more formal demonstration of this important distribution is exhibited in this number by Dr P. L. Hsu.

4. Some special cases

A common source of such an analysis of variance as was considered in § 3 is the expression of p dependent in terms of n_1 independent variates. This situation has been most elegantly elucidated by Hotelling. The linear compound corresponding to the largest root has been termed by him the "most predictable criterion". The compounds corresponding with the whole series of roots he has termed the canonical series of dependent variates, while the multiple regression formulae for these form the corresponding covariant series of dependent variates. The roots θ are then the squares of the correlation coefficients between different pairs. The same roots are thus obtained if the sets of variates are interchanged. Thus, if p exceeds n_1 there are only n_1 roots, and n_1 and p are interchanged in the distribution formula. If either set of variates is normally distributed the distribution of the roots, for independence of the sets of variates, or for homogeneity in respect of the second set of samples selected in respect of the first set, is unaffected by non-normality of the other set. This explains the curious equivalence of the discriminant function for a single contrast with the partial regression formula for an artificial variate introduced to register the contrast to which attention was called in an earlier paper (Fisher, 1938).

This case is reproduced by putting $n_1 = 1$, then the distribution of θ is

$$\frac{(\frac{1}{2}n-\frac{1}{2})!}{(\frac{1}{2}p-1)!\,(\frac{1}{2}n-\frac{1}{2}p-\frac{1}{2})!}\,\theta^{\frac{1}{2}p-1}(1-\theta)^{\frac{1}{2}n-\frac{1}{2}p-\frac{1}{2}}\,d\theta,$$

being the test of significance of Hotelling's generalized "Student's" ratio, as used in simple discriminant analysis.

Again, in the case $p = 2$, the simultaneous distribution of the two roots is

$$\begin{aligned} df &= \frac{(\frac{1}{2}n_1+\frac{1}{2}n_2-1)!\,(\frac{1}{2}n_1-\frac{1}{2}n_2-\frac{3}{2})!\,\sqrt{\pi}}{(\frac{1}{2}n_1-1)!\,(\frac{1}{2}n_1-\frac{3}{2})!\,(\frac{1}{2}n_2-1)!\,(\frac{1}{2}n_2-\frac{3}{2})!} \\ &\qquad \times \theta^{\frac{1}{2}n_1-\frac{3}{2}}\,\theta_2^{\frac{1}{2}n_1-\frac{3}{2}}\,(1-\theta_1)^{\frac{1}{2}n_2-\frac{3}{2}}\,(1-\theta_2)^{\frac{1}{2}n_2-\frac{3}{2}}(\theta_1-\theta_2)\,d\theta_1\,d\theta_2 \\ &= \frac{(n_1+n_2-2)!}{4(n_1-2)!\,(n_2-2)!}\,q^{\frac{1}{2}n_1-\frac{3}{2}}\,(1-p+q)^{\frac{1}{2}n_2-\frac{3}{2}}\,dp\,dq, \end{aligned}$$

where p is the sum, and q the product of the roots. For given q the limits of p are $2\sqrt{q}$ and $1+q$. Integrating with respect to p, we have

$$\begin{aligned} &\frac{(n_1+n_2-2)!}{4(n_1-2)!\,(n_2-2)!}\,q^{\frac{1}{2}n_1-\frac{3}{2}}\,\frac{2}{n_2-1}\,\{1-2\sqrt{q}+q\}^{\frac{1}{2}n_2-\frac{1}{2}}\,dq \\ &\qquad = \frac{(n_1+n_2-2)!}{(n_1-2)!\,(n_2-2)!}\,(\sqrt{q})^{n_1-2}\,(1-\sqrt{q})^{n_2-1}\,d(\sqrt{q}), \end{aligned}$$

the curious result found by Wilkes in studying the distribution of the ratio of the so-called "generalized variance", which is the product of the roots. The significance of such a product may thus be tested by a z-test. Such a test would, of course allow a large root due to some relevant causation to be obscured by the accident that the second root happened to be exceptionally small. It will generally be the largest root, or the largest of those doubtfully significant, which will need to be tested.

An important section of the general distribution is the limiting form when n_2 is large. If

$$n_2 \to \infty, \quad n_1 = n$$

and

$$n_2\theta \to 2\phi,$$

the simultaneous distribution of the p variates ϕ is seen to be

$$\begin{aligned} &\frac{\pi^{\frac{1}{2}p}}{(\frac{1}{2}n-1)!\ldots(\frac{1}{2}n-\frac{1}{2}p-\frac{1}{2})!\,(\frac{1}{2}p-1)!\ldots(-\frac{1}{2})!} \\ &\qquad \times e^{-\phi_1-\ldots-\phi_p}(\phi_1\ldots\phi_p)^{\frac{1}{2}n-\frac{1}{2}p-\frac{1}{2}}(\phi_1-\phi_2)\ldots(\phi_{p-1}-\phi_n)\,d\phi_1\ldots d\phi_p, \end{aligned}$$

where

$$0<\phi_p<\phi_{p-1}<\ldots<\phi_1<\infty. \qquad \ldots\ldots(11)$$

Apart from its own analytic interest, this form supplies a simple demonstration of the form of the function of p referred to in § 3. For let

$$\int \frac{(\phi_1\ldots\phi_p)^{\frac{1}{2}n-\frac{1}{2}p-\frac{1}{2}}e^{-\phi_1-\ldots-\phi_p}}{(\frac{1}{2}n-1)!\ldots(\frac{1}{2}n-\frac{1}{2}p-\frac{1}{2})!}\,(\phi_1-\phi_2)\ldots(\phi_{p-1}-\phi_p)\,d\phi_1\ldots d\phi_p = F(p),$$

when the integration extends over all admissible sets of values; then if $n = p+1$

$$(\tfrac{1}{2}p-\tfrac{1}{2})!\dots 0!\,F(p) = \int e^{-\phi_1-\dots-\phi_p}(\phi_1-\phi_2)\dots(\phi_{p-1}-\phi_p)\,d\phi_1\dots d\phi_p. \quad \dots\dots(12)$$

Now substitute in the right-hand element

$$\phi_p = x,\ \phi_i-\phi_p = \phi_i',\quad i = 1,\dots,p-1;$$

then we have the recurrence equation

$$\begin{aligned}(\tfrac{1}{2}p-\tfrac{1}{2})!\dots(0)!\,F(p)&\\ &= \int_0^\infty e^{-px}dx\int(\phi_1'\phi_2'\dots\phi_{p-1}')\,e^{-\phi_1'-\dots-\phi_{p-1}'}(\phi_1'-\phi_2')\dots(\phi_{p-1}'-\phi_{p-1}')\,d\phi_1'\dots d\phi_{p-1}'\\ &= \frac{1}{p}(\tfrac{1}{2}p)!\dots(1)!\,F(p-1).\end{aligned}$$

Removing the common factors, it appears that

$$\begin{aligned}F(p)/F(p-1) &= \frac{1}{p}(\tfrac{1}{2}p)!/(\tfrac{1}{2})!\\ &= (\tfrac{1}{2}p-1)!/\sqrt{\pi}. \quad \dots\dots(13)\end{aligned}$$

Since also, when $p = 1$, $\quad F(p) = 1 = (-\tfrac{1}{2})!/\sqrt{\pi}$

it follows in general that

$$F(p) = \pi^{-\frac{1}{2}p}(\tfrac{1}{2}p-1)!\dots(-\tfrac{1}{2})!, \quad \dots\dots(14)$$

the form given in § 3.

Returning to equation (12) it is interesting to note that it may be written in the form

$$\begin{aligned}\int e^{-\phi_1-\dots-\phi_p}(\phi_1-\phi_2)\dots(\phi_{p-1}-\phi_p)\,d\phi_1\dots d\phi_p&\\ &= \prod_{i=1}^{p}(\tfrac{1}{2}p-\tfrac{1}{2}i)!\,(\tfrac{1}{2}p-\tfrac{1}{2}i-\tfrac{1}{2})!\,\pi^{-\frac{1}{2}}\\ &= \prod_{i=1}^{p}(p-i)!\,2^{-(p-i)}\\ &= 2^{-\frac{1}{2}p(p-1)}(p-1)!\,(p-2)!\dots(0)!. \quad \dots\dots(15)\end{aligned}$$

Summary

Hitherto little has been known of the distribution of statistics requiring the solution of non-linear equations. The test of significance of a selected harmonic, obtained some years ago, shows, however, that some complexity is to be expected in exact tests of significance for these.

In the present paper the solution is given of the simultaneous distribution of the roots of certain quantic equations which arise in discriminant analysis.

REFERENCES

R. C. Bose (1936). "On the exact distribution of D^2 statistic." *Sankhyā*, **2**, 143–54.

R. C. Bose & S. N. Roy (1938). "The exact distribution of the Studentized D^2 statistic." *Sankhyā*, **4**, 19–38.

R. A. Fisher (1915). "Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population." *Biometrika*, **10**, 507–21.

—— (1929). "Tests of significance in harmonic analysis." *Proc. Roy. Soc.* A, **125**, 54–9.

—— (1936). "The significance of regression coefficients." *Cowles Commission for Research Conference, Colorado College Publication*, no. 208, pp. 63–7.

—— (1938). "The statistical utilization of multiple measurements." *Ann. Eugen., Lond.*, **8**, 376–86.

H. Hotelling (1933). "Analysis of a complex of statistical variables into principal components." *J. Educ. Psychol.* **24**, 417–41 and 498–520.

—— (1935). "The most predictable criterion." *J. Educ. Psychol.* **26**, 139–42.

—— (1936). "Relations between two sets of variates." *Biometrika*, **28**, 321–77.

P. C. Mahalanobis (1930). "On tests and measures of group divergence. Part I. Theoretical formulae." *J. Asiat. Soc. Beng.* **26**, 541–88.

—— (1936). "On the generalized distance in statistics." *Proc. Nat. Inst. Sci. Ind.* **12**, 49–55.

J. Wishart (1928). "The generalized product moment distribution in samples from a normal multivariate population." *Biometrika*, **20** A, 32–52.