THE CONSISTENCY AND ULTIMATE DISTRIBUTION OF OPTIMUM STATISTICS*

BY
HAROLD HOTELLING

1. Definitions. It has been customary to assume somewhat loosely that when a quantity is calculated from a random sample to estimate a parameter of a hypothetical frequency distribution, the accuracy of the determination will increase without limit as the number in the sample increases. A *consistent statistic* is a function of the observations actually possessing this property, which we define more precisely as follows. If p is the true value of the parameter (e.g. the true declination of a star) and p' an estimate of p calculated from data (e.g. the geometric mean of the declinations arrived at by a class in astronomy), then p' is a consistent statistic if for every two positive numbers δ and ϵ a number N exists such that if the sample contains more than N individuals the probability is less than ϵ that

$$|p - p'| > \delta.$$

However, some statistics, such as averages of biased measurements, or rank correlation coefficients as direct estimates of product-moment correlations, are not consistent. The utility of such a statistic, if any, lies in ease of computation, together with sufficient accuracy for ordinary purposes, though the accuracy cannot be increased beyond a certain point by multiplying observations.

It has also been common to assume loosely that the probability distribution of a statistic calculated from n observations has a shape which approaches that of the normal ("Gaussian") curve as n increases. But in some cases this is not true. For example the range r of a sample of n from a rectangular population of breadth p has the probability

$$\frac{n(n-1)r^{n-2}(p-r)dr}{p^n} \qquad (r < p)$$

of falling in the range dr. The form of this distribution does not approach normality.

If the probability of obtaining an observation in the range $x \pm \frac{1}{2}dx$ is $f(x, p)dx$, and if observations $x_1, \cdots, x_n$ have been obtained, then

* Presented to the Society, December 30, 1929; received by the editors in December, 1929.

$$L = \sum_{i=1}^{n} \log f(x_i, p),$$

or any quantity whose difference from this does not involve p, has been defined by R. A. Fisher as the logarithm of the *likelihood.** The same definition holds if the probability that an observation shall be exactly x is finite and equal to $f(x, p)$. Thus we deal simultaneously with continuous and with discontinuous distributions of x. For a multivariate distribution of any kind and for a multiplicity of parameters this and the following definitions may be extended in obvious fashion.

The value $\hat{p}$ of p which makes L a maximum is a function of $x_1, \cdots, x_n$; it will be convenient to use Fisher's designation of this function as the *optimum statistic* or *optimum estimate of* p. Examples of optimum statistics are the mean of a sample from a normal distribution for estimating the mean of the population, and the range r of a sample from a rectangular distribution for estimating the range of this distribution. The *variance* of any quantity, the square of its standard error, is defined as the mean value, or mathematical expectation, of the square of its deviation from its mean value. The *covariance* of two quantities is the mean value of the product of their deviations from their respective mean values. In the paper cited Fisher shows† that, if an optimum statistic has a distribution approaching the normal form, its variance approximates to the reciprocal of the mean value of $-\partial^2 L/\partial p^2$ as n increases. As an extension to the case of simultaneous estimation of parameters $p_1, \cdots, p_k$, he uses for the variances and covariances of the optimum statistics the elements of the inverse of the matrix

$$\left|\left| -E\left(\frac{\partial^2 L}{\partial p_i \partial p_j}\right) \right|\right|$$

where E signifies the mathematical expectation. In this way Fisher arrives at results of wide practical importance, including a condemnation of the general use of the method of moments for fitting frequency curves.

Apart from the question whether the hypothesis of an ultimately normal distribution is satisfied, it is not clear what conditions, particularly of continuity, are necessary in order that the proofs which have been given shall

---

* *On the mathematical foundations of theoretical statistics*, Philosophical Transactions, vol. 222A (1922), pp. 309–368.

† By a method similar to one used for the same purpose by Pearson and Filon, Philosophical Transactions, vol. 191A (1898), p. 229.

be valid. Thus Fisher's application (op. cit. p. 333) to the location of the Type III curve breaks down if $p \leqq 1$, though all values greater than -1 are possible.

We shall put $\phi = \log f$. Then since the total frequency of the population is unity, $\int e^{\phi} dx = 1$. Differentiating twice we have

$$n \int_{-\infty}^{\infty} (\partial\phi/\partial p)^2 e^{\phi} dx = - n \int_{-\infty}^{\infty} (\partial^2\phi/\partial p^2) e^{\phi} dx,$$

provided the second derivative exists and the integrals converge. The right-hand side of the equation is the mean value used by Fisher and others. Likewise for problems of simultaneous estimation of two or more parameters we find that the mean value of

$$\frac{\partial\phi}{\partial p_i} \frac{\partial\phi}{\partial p_j}$$

and that of

$$- \frac{\partial^2\phi}{\partial p_i \partial p_j}$$

are identical.

2. **Theorems.** We shall prove explicitly the theorems below for frequency functions of one continuous variable. However, the extensions to any number of variables are perfectly obvious; and the corresponding theorems for discrete variables follow immediately by replacing each value of the variable by an interval within which the probability may be supposed uniformly distributed.

The true value of p will be denoted by p_0, its optimum estimate by $\hat{p}$, and the probability that a random value of x lies between α and β by $\int_{\alpha}^{\beta} f(x, p) dx$. The following properties of the distribution function f. will be used:

(a) f is a continuous function of x except on a set of values of x of measure zero.

(b) f is a continuous monotonic function of p in a p-interval including p_0 for all values of x in some interval.

(c) In a p-interval including p_0, $\partial f/\partial p$ is a continuous function of p for every value of x; $x^2 \partial f/\partial p$ approaches a continuous function of p as $x \to \pm \infty$; in some x-interval $\partial f/\partial p$ does not vanish.

The theorems are the following:

I. *If* (a) *and* (b) *are satisfied, $\hat{p}$ is a consistent statistic.*

II. *The distribution of $\hat{p}$ approaches the normal form if* (a) *and* (c) *are satisfied.*

III. *Under hypotheses* (a) *and* (c) *the variance of* $\hat{p}$ *bears to*

$$E\left(\frac{\partial L}{\partial p}\right)^2_{p=p_0} = n\int_{-\infty}^{\infty}[\partial \log f(x, p_0)/\partial p_0]^2 f(x, p_0)dx$$

$$= n\int_{-\infty}^{\infty}[\partial\phi(x, p_0)/\partial p_0]^2 e^{\phi(x,p_0)}dx$$

$$= -n\int_{-\infty}^{\infty}[\partial^2\phi(x, p_0)/\partial p_0^2]e^{\phi(x,p_0)}dx,$$

a ratio which approaches unity as n increases.

IV. *If* $f(x, p_1, \cdots, p_k)$ *is a frequency function satisfying the conditions* (a) *and* (c) *for each* p_i, *the optimum joint estimates of* $p_1, \cdots, p_k$ *have a distribution approaching the normal form as n increases. The variances and covariances, each multiplied by n, approach the elements of a matrix whose inverse consists of the elements*

$$\int_{-\infty}^{\infty}\frac{\partial\phi}{\partial p_i}\frac{\partial\phi}{\partial p_j}e^{\phi}dx,$$

where $\phi = \log f$.

In §6 we shall correct the usual proof of the formula for the covariance of two frequencies.

The hypothesis of Theorem II which is violated in locating a Type III curve for which the parameter of shape, p, is not greater than unity is that of continuity of the first derivative with respect to the parameter of location, which Fisher denotes by m. Variation of m determines a translation of the curve along the axis. When $p=1$ the curve makes an angle at one end with the axis; when $p<1$ the angle is $\pi/2$. Translation therefore causes an ordinate corresponding to a fixed value of x to change at a discontinuous rate when p has such values. By a transformation of x the curve can be thrown into a form tangent to the axis and so satisfying the hypotheses (c), but the transformation must depend upon the location as well as the shape of the curve.

3. **Reduction of infinite to finite range.** Three kinds of infinity are involved in the problem. The range of x may be infinite, or may increase indefinitely as p approaches some value. Secondly, we shall divide the range into subintervals whose lengths will eventually approach zero. Finally, the number in the sample is to increase indefinitely.

The infinite range we dispose of at once by the simple expedient of transforming the variable into one of finite range and using the invariantive property of the optimum statistic. Thus we may put $x=\tan\theta$. The distri-

bution $f(x, p)dx$ then becomes

$$F(\theta, p)d\theta = f(\tan\theta, p)\sec^2\theta d\theta,$$

and the likelihood,

$$L = \sum \log [f(\tan\theta_i, p)\sec^2\theta_i] = Q + \sum \log f(x_i, p),$$

where $x_i = \tan\theta_i$. Since Q does not involve p, the value of p which makes L a maximum is exactly the same as before. However the range of the variable θ is finite.

The new frequency function satisfies the same conditions as the old concerning the continuity and non-vanishing of its first derivative with respect to p. This is true even at the ends of the range, since there

$$\partial F/\partial p = \lim (1 + x^2)\partial f/\partial p,$$

which is continuous by hypothesis (c). The integrals in Theorems III and IV have the same values whether calculated from the distribution of x or from that of θ.

4. **Approximation by grouped distribution.** Let the range of the variable x, which we shall from now on assume finite, be divided into m intervals $J_1, J_2, \cdots, J_m$ of respective lengths $l_1, \cdots, l_m$. For convenience we take these intervals such that the frequencies

$$f_t(p) = \int_{J_t} f(x, p)dx \qquad (t = 1, 2, \cdots, m)$$

are all equal when p has the value p_0 which it takes in the population in question. The maximum length l can be made arbitrarily small by using a sufficiently great number m of intervals, provided that in measuring the length of an interval we exclude any subintervals in which $\int f dx$ is zero.

Hypothesis (a) shows that if x is in J_t, the ratio $f_t(p)/l_t$ can by increase of m be made to differ as little as we please from $f(x, p)$, unless x is one of a set of values of measure zero and so can be neglected in considering probabilities. Consequently, if in a sample of n the number falling in the class J_t be denoted by n_t, the logarithm of the likelihood derived from the grouped data, which may be taken as

$$L' = \sum_{t=1}^{m} n_t \log f_t(p) - \sum n_t \log l_t,$$

the second sum not involving p and so not affecting the maximizing value, differs arbitrarily little from

$$L = \sum_{i=1}^{n} \log f(x_i, p).$$

This is true for every value of p; and by hypothesis $f(x, p)$, and therefore L and L', are continuous functions of p. Therefore if $\hat{p}'$ and $\hat{p}$ denote respectively the values maximizing L' and L, the difference $|\hat{p}'-\hat{p}|$ can be made arbitrarily small by increasing m. Consequently if our theorems are true for data grouped in this way, and for every sufficiently large value of m, they are true also for ungrouped data. This fact enables us to concentrate our attention upon the variables $n_1, \cdots, n_m$, which are finite in number.

5. **Proof of Theorem I.** Consider the expression

$$L = \sum_{t=1}^{m} n_t \log z_t,$$

where n_t is the number of observations in a particular sample of n falling in the interval J_t, so that $\sum n_t = n$. Let

$$L = L_R \text{ when } z_t = n_t/n;$$
$$L = L_Z \text{ when } z_t = f_t(\hat{p}');$$
$$L = L_M \text{ when } z_t = f_t(p_0) = 1/m \qquad (t = 1, 2, \cdots, m).$$

Then L_R is the maximum value of L when the z_t are subject only to the restriction

$$\sum_{t=1}^{m} z_t = 1; \tag{5.1}$$

L_Z is the maximum value when they are subject to the severer condition that a value of p exists such that

$$z_t = f_t(p) \qquad (t = 1, 2, \cdots, m) \tag{5.2}$$

and L_M satisfies the last condition but is not the maximum. Hence

$$L_R \geqq L_Z \geqq L_M.$$

Now Bernoulli's law of great numbers shows that, when n is large enough, we can assert with a probability greater than $1-\epsilon$ that

$$| n_t/n - f_t(p_0) | < \delta, \tag{5.3}$$

ϵ and δ being arbitrarily small. When this condition is satisfied, since L is a continuous function of the z_t, L_R-L_M has an upper limit which approaches zero with δ. The same will therefore be true of L_R-L_Z. We shall next prove it true of each of the quantities

$$| f_t(\hat{p}') - f_t(p_0) | .$$

Since the hypotheses of Theorem I show that some at least of the functions

$f_t(p)$ have single-valued continuous inverses when m is large enough, it will then follow that $|\hat{p}'-p_0|$, and so, by §4, $|\hat{p}-p_0|$, are arbitrarily small, thus establishing the theorem.

To fill in the gap in the argument, let L_1 be the maximum value of L for a fixed value of z_1, and subject otherwise only to (5.1). Then $L_R \geqq L_1$. It is easy to see that this maximum is determined by

$$z_i = n_i(1 - z_1)/(n - n_1) \qquad (i = 2, 3, \cdots, m), \tag{5.4}$$

and equals

$$L_1 = n_1 \log z_1 + (n - n_1)[\log (1 - z_1) - \log (n - n_1)] + \sum_{i=2}^{m} n_i \log n_i.$$

Obviously L_1 is a continuous function of z_1 from 0 to 1, not inclusive, and

$$dL_1/dz_1 = n_1/z_1 - (n - n_1)/(1 - z_1)$$

changes sign only for $z_1 = n_1/n$. By the theorem on inverse functions, z_1 is therefore a double-valued *continuous* function of L_1, the two branches becoming equal for $z_1 = n_1/n$, where $L_1 = L_R$. Hence, for any $\delta' > 0$,

$$|z_1 - n_1/n| < \delta',$$

if $L_R - L_1$ is sufficiently small. This will be true a fortiori if $z_2, z_3, \cdots, z_m$ have values other than (5.4) and if $L_R - L$ is sufficiently small, for $L_R - L_1$ is still smaller.

In the same way each of the quantities

$$|z_t - n_t/n| < \delta'$$

when $L_R - L$ is small enough. Putting $L = L_Z$, which as we have seen will be arbitrarily close to L_R, we have

$$|f_t(\hat{p}') - n_t/n| < \delta'$$

and therefore from (5.3),

$$|f_t(\hat{p}') - f_t(p_0)| < \delta + \delta',$$

the required inequality.

A geometrical interpretation may render the foregoing argument more intelligible. The z_t being coördinates in m-space, (5.1) is a hyperplane containing the curve (5.2), the point R representing the observations, the point Z on the curve at which the parameter p takes the optimum value p', and the point M where p takes its true value p_0. The likelihood L is constant over

a system of approximately spherical hypersurfaces about R. The point Z is the point of the curve which lies on the smallest of the approximate spheres meeting the curve, and is therefore approximately the nearest point on the curve to R. The argument shows with probability greater than $1-\epsilon$ that Z lies arbitrarily close to R and therefore to M, and that the values of the parameter at Z and at M differ arbitrarily little.

The curve (5.2) may be quite irregular. For example the arc length may fail to exist, owing to infinitely rapid oscillations of some of the ordinates of the frequency function when p varies. The conditions of Theorem I require only one coördinate—that is, the integral of the frequency function in some x-interval—to be continuous and monotonic with respect to p, and then only in a short p-interval. The subsequent theorems require functions of a smoother sort.

6. **Distribution of class frequencies.** The familiar demonstration by means of Stirling's formula shows that the probability that n_t will fall between any limits is given by the integral between these limits of a normal curve of variance proportional to n, apart from terms which vanish as n increases. We now examine the joint normal distribution of all these class frequencies in order to deduce that of $\hat{p}$.

Let the deviations

$$r_t = n_t/n - 1/m$$

of the observed relative frequencies from their values in the population be taken as the cartesian coördinates of a point R in m dimensions. The values of the n_t/n obtained from random samples have the binomial distribution with mean $1/m$, and therefore with variance

$$\sigma^2 = (1 - 1/m)/(mn).$$

If each r_t were determined from a different random sample of n, the points R would form a globular cluster in the m-space. The distribution of probability would have, in fact, a spherical symmetry whenever n is large enough for the binomial to be replaced by the normal distribution; for the probability of R falling in the volume element $dr_1 dr_2 \cdots dr_m$ would be the product of the m separate probabilities, or

$$e^{-U/2\sigma^2} dr_1 dr_2 \cdots dr_m/(2\pi\sigma^2)^{m/2},$$

where

$$U = \sum_{t=1}^{m} r_t^2.$$

However the points R which we are considering are confined to the hyperplane V_{m-1} of equation $\sum r_t = 0$, since all the coördinates are derived from

the same sample. The imposition of this restriction destroys the mutual independence of the class frequencies, giving instead a covariance between any two of them which will now be seen to equal $-1/(nm^2)$.

The derivations which have been given of the covariance or correlation between sampling deviations in class frequencies are objectionable because of the inconsistent assumption that the deviation of an observed class frequency from its mean value is compensated, in the same sample, by deviations of the opposite sign in all the other class frequencies distributed exactly in proportion to the mean values. Since we are attempting an accurate treatment we must therefore turn aside to give an accurate proof of this much-used proposition. The accurate proof is simpler than the customary one.

In any problem of sampling grouped values, let f_i be the mean frequency in the ith class in a sample of n, and let $f_i+\delta f_i$ be the observed frequency. Then from the elementary theory of the Bernoulli distribution,

$$E(\delta f_i)^2 = f_i(1 - f_i/n),$$
$$E(\delta f_j)^2 = f_j(1 - f_j/n).$$

Combining the ith and jth classes we have, by the same principle,

$$E(\delta f_i + \delta f_j)^2 = (f_i + f_j)\left(1 - \frac{f_i + f_j}{n}\right) \qquad (i \neq j).$$

Subtracting the first two of these equations from the last and dividing by 2 we have the familiar expression for the covariance,

$$r_{f_if_j}\sigma_{f_i}\sigma_{f_j} = E(\delta f_i\,\delta f_j) = -f_if_j/n.$$

Substituting $f_i=f_j=1/m$, we have the result needed for present purposes: the covariance of any two of our r_i is $-1/(nm^2)$. The density of the points R in V_{m-1} is therefore proportional to

$$\exp\left(-\frac{1}{2}\sum_{s=1}^{m-1}\sum_{t=1}^{m-1}a_{st}r_sr_t\right),$$

where $||a_{st}||$ is the inverse of the matrix of variances and covariances. In accordance with the results just obtained, the latter matrix is

$$\begin{Vmatrix} (1-1/m)/(nm) & -1/(nm^2) & \cdots & -1/(nm^2) \\ -1/(nm^2) & (1-1/m)/(nm) & \cdots & -1/(nm^2) \\ \cdots & \cdots & \cdots & \cdots \\ -1/(nm^2) & -1/(nm^2) & \cdots & (1-1/m)/(nm) \end{Vmatrix}.$$

It is easy by straightforward algebra to obtain the inverse of this matrix or of the more general one in which the class frequencies are unequal, though Karl Pearson resorted for the latter purpose to a complicated trigonometric method.* With the help of $\sum_{t=1}^{m} r_t = 0$ we obtain simply

$$\sum_{s=1}^{m-1}\sum_{t=1}^{m-1} a_{st} r_s r_t = mn \sum_{t=1}^{m} r_t^2 .$$

Thus the density distribution in V_{m-1} is spherically symmetrical.

Let us take new cartesian coördinates $y_1, y_2, \cdots, y_m$ with the same origin as the r_t and such that $y_m = 0$ on V_{m-1}. On this hyperplane the density distribution, being symmetrical, will be given by

$$(mn/(2\pi))^{(m-1)/2} \exp\left(-\frac{1}{2} mn \sum_{t=1}^{m-1} y_t^2 \right) dy_1 \cdots dy_{m-1}.$$

Now suppose that each point R is projected orthogonally upon a line which, without loss of generality, we assume to be the y_1-axis. The linear density of the resulting points is found by integrating the last expression with respect to $y_2, y_3, \cdots, y_{m-1}$, an easy matter because it is the product of factors involving only one variable each. The result shows that the linear density of the projected points is normally distributed with variance

$$\sigma_s^2 = 1/(mn). \tag{6.1}$$

7. **Proof of normal distribution.** In addition to R, each sample determines a point Z in V_{m-1}, the coördinates $z_1, \cdots, z_m$ of Z being the deviations from $1/m$ of the proportions falling into the several classes of the total frequency in a population having for its parameter p the value $\hat{p}'$ for which the likelihood of the given sample is a maximum. We shall now find an approximate geometrical construction for Z when R is known.

Since the coördinates of Z are the relative frequencies in a population of the form under consideration, minus $1/m$, they must satisfy the equations

$$z_t = f_t(p) - 1/m \qquad (t = 1, 2, \cdots, m) \tag{7.1}$$

which define a curve in V_{m-1} in terms of the parameter p. This curve C passes through the center P of our globular cluster; and since it follows from the hypotheses of Theorem II that the functions $f_t(p)$ have continuous first derivatives of which one at least does not vanish, C is approximated in a neighborhood of P by the tangent straight line, along which the distance measured from P is

* *On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling*, Philosophical Magazine, vol. 50 (1900), p. 157.

$$s = (p - p_0)\left\{ \sum [f_t'(p_0)]^2 \right\}^{1/2}. \tag{7.2}$$

The point Z for a particular sample is that for which

$$L = \sum_{t=1}^{m} n_t \log (z_t + 1/m) \tag{7.3}$$

is a maximum for variations of z_t subject to (7.1). For any constant value of L, (7.3) is the equation of a hypersurface. The problem of maximizing L subject only to the condition

$$\sum z_t = 0 \tag{7.4}$$

would be the problem of determining one of these hypersurfaces tangent to the hyperplane (7.4). The solution of this problem is simply $z_t = n_t/n - 1/m$; the point of tangency is R.

Putting L_R for the value of L at this point, L_Z for its value at Z, and $\delta z_1, \cdots, \delta z_m$ for the differences in coördinates of these points, we have, apart from terms of higher order,

$$L_Z = L_R + \sum_t (\partial L/\partial z_t)_R \, \delta z_t + \tfrac{1}{2} \sum_s \sum_t (\partial^2 L/\partial z_s \partial z_t)_R \, \delta z_s \, \delta z_t.$$

Now from (7.3) we have at R, where $z_t = n_t/n - 1/m$,

$$\partial L/\partial z_t = n, \qquad \partial^2 L/\partial z_s \, \partial z_t = 0 \text{ if } s \neq t, \qquad \cdot = -n^2/n_t \text{ if } s = t.$$

Moreover $\sum \delta z_t = 0$, since both Z and R lie in V_{m-1}. Thus Z is the point on the curve C for which

$$\sum (\delta z_t)^2/n_t$$

is a minimum. For large values of n the denominators tend to equality and the point Z therefore to the nearest point on the curve to R.

This shows that the points R are in the limit projected orthogonally upon C, or approximately upon the tangent line. The distance s along this line will then have a normal distribution with variance (6.1). The distribution of p will therefore approach the normal form with a variance, derived from that of s by means of (7.2), equal to

$$1/\left\{ mn \sum [f_t'(p_0)]^2 \right\}. \tag{7.5}$$

Since this p is identical with $\hat{\hat{p}}'$, which approximates $\hat{\hat{p}}$ for the fine grouping implied by large values of m, Theorem II is established.

It is important that the last of the conditions (c) prevents the vanishing of the denominator of (7.5). If this condition were not insisted upon it would be easy to find parameters whose optimum estimates did not tend to a normal distribution and whose sampling deviations might therefore easily

be misinterpreted. Such for example would be $q=(p-p_0)^2$; the distribution of an optimum estimate of q would approach the form $e^{-kq}\,dq/[2(k\pi q)^{1/2}]$.

8. **Evaluation of variance.** To prove Theorem III we multiply (7.5) by n and let m increase without limit. Since by definition

$$f_t(p_0) = 1/m,$$

and since, apart from terms of higher order in m^{-1},

$$f_t(p) = f(x, p)l_t,$$

so that

$$f_t'(p)^2/f_t(p) = [\partial \log f(x, p)/\partial p]^2 f(x, p)l_t,$$

it follows that the expression

$$m\sum[f_t'(p_0)]^2 = \sum[f_t'(p_0)]^2/f_t(p_0)$$

will, as m increases, approach the value for $p=p_0$ of

$$\int_{-\infty}^{\infty}(\partial \log f/\partial p)^2 f\,dx, \quad \text{or} \quad \int_{-\infty}^{\infty}(\partial\phi/\partial p)^2 e^{\phi}dx,$$

where $f=e^{\phi}$ is taken as zero outside the range of x. This proves Theorem III.

9. **Extension to more than one parameter.** For each of k parameters $p_1, \cdots, p_k$ separately, Theorems I, II, and III hold. Since the optimum estimate of each has in the limit a normal distribution, their joint distribution approaches that for which the frequency element is proportional to $e^{-nT/2}$, where

$$T = \sum_{i=1}^{k}\sum_{j=1}^{k} g_{ij}(p_i - p_{i0})(p_j - p_{j0})$$

is a positive definite quadratic form. Here p_{i0} represents the true value of p_i, and the carats have been dropped from the p_i in denoting the optimum estimates. The matrix of the g_{ij} is the inverse of that of the variances and covariances.

Let us apply a linear transformation which reduces T to

$$T' = q_1^2 + q_2^2 + \cdots + q_n^2,$$

and let the inverse transformation be

$$q_h = \sum_i A_{hi}(p_i - p_{i0}). \tag{9.1}$$

Equating coefficients,

$$g_{ij} = \sum_h A_{hi}A_{hj}. \tag{9.2}$$

Taking the q's instead of the p's as parameters we observe from the form of T' that their optimum estimates are in the limit distributed normally and independently with variance $1/n$. Hence from Theorem III,

$$\int (\partial\phi/\partial q_h)^2 e^\phi \, dx = 1. \tag{9.3}$$

To show further that

$$\int (\partial\phi/\partial q_h)(\partial\phi/\partial q_l) e^\phi \, dx = 0 \text{ if } h \neq l, \tag{9.4}$$

we put

$$q_h = (u_h - u_l)/2^{1/2},$$
$$q_l = (u_h + u_l)/2^{1/2},$$

that is,

$$u_h = (q_h + q_l)/2^{1/2},$$
$$u_l = (-q_h + q_l)/2^{1/2}.$$

The left member of (9.4) equals one-half of

$$\int (\partial\phi/\partial u_h)^2 e^\phi \, dx - \int (\partial\phi/\partial u_l)^2 e^\phi dx.$$

But these integrals are equal because they are respectively n times the variances of u_h and u_l, both of which have, by their definition, the same variance as the q's.

Now since, by (9.1),

$$\partial\phi/\partial p_i = \sum(\partial\phi/\partial q_h)(\partial q_h/\partial p_i) = \sum A_{hi}\partial\phi/\partial q_h,$$

we have

$$(\partial\phi/\partial p_i)(\partial\phi/\partial p_j) = \sum_h \sum_l A_{hi} A_{lj} (\partial\phi/\partial q_h)(\partial\phi/\partial q_l).$$

Taking the mean values of both sides and using (9.3), (9.4), and then (9.2),

$$\int [(\partial\phi/\partial p_i)(\partial\phi/\partial p_j)] e^\phi \, dx = \sum A_{hi} A_{hj} = g_{ij}.$$

Thus Theorem IV is proved.

STANFORD UNIVERSITY,
STANFORD UNIVERSITY, CALIF.