AN OPTIMUM PROPERTY OF CONFIDENCE REGIONS ASSOCIATED WITH THE LIKELIHOOD FUNCTION[1]

By S. S. Wilks and J. F. Daly

One of the authors [1] has recently established a connection between the method of maximum likelihood and shortest average confidence intervals for the case of one unknown parameter, and has reported a generalization [2] of this result for the case of several parameters. It is the object of this paper to consider the several-parameter problem in greater detail and at the same time to make the previously obtained result slightly stronger, particularly in the one-parameter case.

Let x denote a set of random variables, and θ a set of parameters $\theta_1, \cdots, \theta_h$. Suppose Π_0 is a population with the cumulative distribution function $F(x, \theta_0) \equiv F_0$ say. Then the logarithm of the likelihood associated with the population Π_0^n of random samples $0_n: x_1, x_2, \cdots, x_n$ drawn from Π_0 is

$$L^n(x, \theta_0) = \sum_{\alpha=1}^{n} \log dF(x_\alpha, \theta_0).$$

For a given sample 0_n we shall say that a set of functions $H_i^n(x, \theta)$ is of class K if there exists a domain R of parameter points θ: $(\theta_1, \cdots, \theta_h)$ in a θ-space such that for each θ_0 in R:

(i) $H_i^n(x, \theta_0) = H_{i0}^n$ is of the form $\sum_{\alpha=1}^{n} h_i(x_\alpha, \theta_0)$;

(ii) $h_i(x, \theta_0) = h_{i0}$ exists for all x except possibly for a set of zero probability;

(iii) $E_0[h_{i0}] = 0$, where E_0 means that the expected value is taken for the population Π_0;

(iv) $\| E_0[h_{i0}h_{j0}] \|$ exists and is non-singular;

(v) the moments $E_0[h_{i0}h_{j0}h_{k0}]$ are all finite.

(Here and throughout the remainder of the paper, the indices i, j, k, l have the range $1, \cdots, h$.) If, in addition,

(iii′) $E_0[h_{i0}]$ can be differentiated under the integral sign;

(iv′) the moments $E_0[h_{i0}h_{j0}]$ are differentiable with respect to the θ's;

the H_i will be said to be of class K'.

We shall need the following lemma, which is very closely related to Theorem 1′ and Theorem 2 in [1] and which can be proved by the method of characteristic functions.

---

[1] Incorporated in this paper is a note presented by one of us (c.f. [2]) at a meeting of the Institute of Mathematical Statistics, December 27, 1938.

LEMMA: *Let $H_i^n(x, \theta)$ be of class K for each n, and put*

$$B_{ij0}^n = \frac{1}{n} E_0[H_{i0}^n H_{j0}^n] = E_0[h_{i0} h_{j0}].$$

Let $\| b_{ij0}^n \|$ be the positive definite matrix satisfying the equation

$$\| b_{ij0}^n \|^2 = \| B_{ij0}^n \|$$

and write

$$\| b_{ij0}^n \|^{-1} = \| b_0^{nij} \|$$

Then for any point θ_0 in R the functions

$$\varphi_{i0}^n = \frac{1}{\sqrt{n}} \sum_{j=1}^{h} b_0^{nij} H_{j0}^n \tag{1}$$

computed from Π_0^n have a joint distribution which converges in large samples to normality, with the density function

$$(2\pi)^{-\frac{h}{2}} e^{-\frac{1}{2} \sum_{i=1}^{h} \varphi_{i0}^2}$$

Now whenever we are justified in assuming a definite functional form for $F(x, \theta)$, and have a set of functions $\varphi_i(x, \theta)$ whose distribution under this last assumption is known and is independent of the θ's, as is the case in the limit for the functions (1), we can obtain, from a sample, information about the values of the θ's. For, given any region S in the space of the functions φ_i, we can determine the probability $P_0\{\varphi_{i0} \subset S\}$ that in samples from Π_0 the point $(\varphi_{10}, \cdots, \varphi_{h0})$ will fall in the region S, even though we do not know the population values θ_0. Suppose, then, that we pick a region S such that $P_0\{\varphi_{i0} \subset S\} > .95$, and agree that each time we encounter such a problem we shall substitute the observed x's into the φ's, and call the set of all points $(\theta_1, \cdots, \theta_h)$ for which $\varphi_i(x, \theta) \subset S$ the *confidence region T*. If this procedure is followed consistently, we can assert that the probability is more than .95 that the region T thus determined contains the true parameter point θ_0.

Evidently the size of the confidence region, i.e., the accuracy with which it serves to locate the true parameter point θ_0, depends upon our choice of the auxiliary functions φ_i. Consider now the case in which there is but one parameter θ, and let $\varphi(x, \theta)$ and $\varphi^*(x, \theta)$ be two functions with the same distribution $D(u)$, where $D(u)$ does not depend on θ. For the set S of the above discussion take the interval $\underline{u} < u < \bar{u}$. Then

$$P_0\{\varphi_0 \subset S\} = P_0\{\varphi_0^* \subset S\} = \alpha$$

where $\alpha = .95$, say. Given a set of observed x's, $\varphi(x, \theta)$ will map S into a confidence region T, while $\varphi^*(x, \theta)$ will map it into a confidence region T^*. Both T and T^* may be expected to contain the true value θ_0 in 95% of the cases; hence a reasonable way to compare the size of T with that of T^* is to compare the quantities $\frac{\partial \varphi}{\partial \theta}(x, \theta_0)$ and $\frac{\partial \varphi^*}{\partial \theta}(x, \theta_0)$; for these derivatives give an indication of the amount of change one can make in θ without forcing φ or φ^* out of the interval S.

The result obtained in [1] in this connection may now be stated as follows: Let $H = \frac{\partial L}{\partial \theta}$ be of class K', and let $H^* = \sum_{\alpha=1}^{n} h(x_\alpha, \theta)$ be any other function of class K'. Then in large samples from Π_0 both

$$\varphi = \frac{H}{\left(nE\left[\left\{\frac{\partial}{\partial \theta} \log dF\right\}^2\right]\right)^{\frac{1}{2}}}$$

and

$$\varphi^* = \frac{H^*}{(nE[\{h(x, \theta)\}])^{\frac{1}{2}}}$$

are distributed almost normally with zero mean and unit variance. But the confidence regions obtained from φ will, on the average, be smaller than those from φ^*, in the sense that, for large samples the inequality

$$\left\{E_0\left[\frac{\partial \varphi_0}{\partial \theta}\right]\right\}^2 > \left\{E_0\left[\frac{\partial \varphi_0^*}{\partial \theta}\right]\right\}^2 \tag{2}$$

will hold (unless $h(x, \theta) \equiv c \frac{\partial}{\partial \theta} \log dF$, in which case alone the inequality (2) becomes an equality).

Now let us return to the several-parameter case. One method of attack which suggests itself is to consider the jacobian determinant

$$\left| \frac{\partial \varphi_{i0}}{\partial \theta_j} \right|$$

for this bears the same relation to the area of the region dS which maps into the region

$$dT\colon \theta_0 - \tfrac{1}{2}d\theta < \theta < \theta_0 + \tfrac{1}{2}d\theta$$

as does the derivative $\frac{\partial \varphi_0}{\partial \theta}$ in the one parameter case. To this end, let us put $L_i^n(x, \theta) = \frac{\partial L^n}{\partial \theta_i}$, and for each n and for each θ_0 in R assume that

(a) L_{i0}^n is defined for all x except perhaps on a set of probability 0;
(b) $E_0[L_{i0}^n] = 0$;
(c) $E_0[L_{i0}^n]$ can be differentiated under the integral sign;
(d) $\| E_0[L_{i0}^n L_{j0}^n] \|$ exists and is non-singular;
(e) $E_0[L_{i0}^n L_{j0}^n]$ is differentiable in the θ's.

Let $H_i^n(x, \theta)$ be any other set of functions satisfying the same conditions. Set

$$E_0[L_{i0}^n L_{j0}^n] = nA_{ij0}^n \qquad E_0[H_{i0}^n H_{j0}^n] = nB_{ij0}^n$$

and define the matrices

$$\| a^n_{ij0} \|^2 = \| A^n_{ij0} \| \qquad \| a_0^{nij} \| = \| a^n_{ij0} \|^{-1}$$
$$\| b^n_{ij0} \|^2 = \| B^n_{ij0} \| \qquad \| b_0^{nij} \| = \| b^n_{ij0} \|^{-1}$$

Now consider the normalized functions

$$\tilde{L}^n_{i0} = \sum_{j=1}^{h} a_0^{nij} L^n_{j0}$$
$$\tilde{H}^n_{i0} = \sum_{j=1}^{h} b_0^{nij} H^n_{j0}$$

We then have

$$\frac{1}{n}\frac{\partial \tilde{L}^n_{i0}}{\partial \theta_k} = \sum_{j=1}^{h} \frac{\partial a_0^{nij}}{\partial \theta_k} \cdot \frac{1}{n} \cdot L^n_{j0} + \sum_{j=1}^{h} a_0^{nij} \cdot \frac{1}{n} \frac{\partial L^n_{j0}}{\partial \theta_k} \tag{3}$$

and by virtue of assumptions (b) and (c) it follows that (c.f. [1], pp. 171–2)

$$E_0\left[\frac{1}{n}\frac{\partial \tilde{L}^n_{i0}}{\partial \theta_k}\right] = -\frac{1}{n}\sum_{j=1}^{h} a_0^{nij} E_0[L^n_{j0} L^n_{k0}]$$

In similar fashion

$$E_0\left[\frac{1}{n}\frac{\partial \tilde{H}^n_{i0}}{\partial \theta_k}\right] = -\frac{1}{n}\sum_{j=1}^{h} b_0^{nij} E_0[H^n_{j0} L^n_{k0}]$$

Consequently

$$(-1)^h \left| E_0\left[\frac{1}{n}\frac{\partial \tilde{L}^n_{i0}}{\partial \theta_k}\right]\right| = | A^n_{ij0} |^{\frac{1}{2}} \tag{4}$$

and

$$(-1)^h \left| E_0\left[\frac{1}{n}\frac{\partial \tilde{H}^n_{i0}}{\partial \theta_k}\right]\right| = | B^n_{ij0} |^{-\frac{1}{2}} \cdot \left| \frac{1}{n} E_0[H^n_{i0} L^n_{j0}] \right| \tag{5}$$

We can find a relation between these two determinants by going over to the matrix

$$M_n = \left\| \begin{matrix} \| E_0[L^n_{i0} L^n_{j0}] \| & \| E_0[L^n_{i0} H^n_{j0}] \| \\ \| E_0[H^n_{i0} L^n_{j0}] \| & \| E_0[H^n_{i0} H^n_{j0}] \| \end{matrix} \right\|$$

This matrix is positive definite unless there is a linear relation with constant coefficients, say $\sum (c_i L_i + d_i H_i) = 0$, which holds for all x's except a set of zero probability; and in this event it is positive semidefinite. From the theory of compound matrices [3] we can then conclude that the matrix whose elements are the h-th order minors of M_n arranged in lexicographic order on both row and column indices has the same property, so that

$$| E_0[L^n_{i0} L^n_{j0}] | \cdot | E_0[H^n_{i0} H^n_{j0}] | \geq | E_0[L^n_{i0} H^n_{j0}] |^2$$

The relations (4) and (5) then imply that

$$\left| \det E_0 \left[\frac{1}{n} \frac{\partial \tilde{L}_{i0}^n}{\partial \theta_k} \right] \right| \geq \left| \det E_0 \left[\frac{1}{n} \frac{\partial \tilde{H}_{i0}^n}{\partial \theta_k} \right] \right| \tag{6}$$

It may be observed that no use has been made of the assumption of linearity (i) in deriving (6). And since in the one parameter case the determinants have but one row and column, we see that in this case the result in [1] remains valid for functions of an even more general type than those of class K'. In order to give the inequality a statistical meaning it seems necessary, however, to require not only that H and L satisfy (a), $\cdots$ (e) but also that in large samples $\frac{1}{\sqrt{n}} \tilde{H}_1^n$ and $\frac{1}{\sqrt{n}} \tilde{L}_1^n$ tend to be distributed independently of θ, with the same (though not necessarily normal) distribution.

For the case of several parameters the transition from the above determinants of expected values to the jacobian determinants requires further argument and further assumptions. To begin with, suppose that the L_i^n and H_i^n are of class K', and that

(vi) the moments $E_0 \left[\frac{\partial h_{i0}}{\partial \theta_j} \frac{\partial h_{k0}}{\partial \theta_l} \right]$ are all finite,

with a corresponding condition on the variances and covariances of $\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log dF(x, \theta_0)$. Let us put

$$Y_{ij0}^n = \frac{1}{n} \frac{\partial H_{i0}^n}{\partial \theta_j} - E_0 \left[\frac{1}{n} \frac{\partial H_{i0}^n}{\partial \theta_j} \right]$$

$$y_{ij0} = \frac{\partial h_{i0}}{\partial \theta_j} - E_0 \left[\frac{\partial h_{i0}}{\partial \theta_j} \right]$$

The characteristic function of the Y_{ij}^n is

$$\varphi_n(t_{11}, \cdots, t_{hh}) = \varphi_n(t) = E_0 [\exp (i \sum t_{ij} Y_{ij})]$$
$$= \left\{ E_0 \left[\exp \left(\frac{i}{n} \sum t_{ij} y_{ij} \right) \right] \right\}^n$$

Expanding the exponential in powers of the t's and using (vi), we find that

$$\varphi_n(t) = \left\{ 1 - O\left(\frac{1}{n^2} \right) \right\}^n$$

so that we have

$$\lim_{n \to \infty} \varphi_n(t) = 1$$

uniformly in every finite interval $| t_{ij} | < M$. A basic theorem on sequences of characteristic functions [4] then guarantees that for any $\epsilon > 0$

$$\lim_{n \to \infty} P_0 \left\{ \left| \frac{1}{n} \frac{\partial H_{i0}^n}{\partial \theta_j} - E_0 \left[\frac{1}{n} \frac{\partial H_{i0}^n}{\partial \theta_j} \right] \right| > \epsilon \right\} = 0$$

that is to say, that $\frac{1}{n}\frac{\partial H_{i0}^n}{\partial\theta_j}$ converges stochastically to its expected value. Under the assumptions of this paragraph the same type of reasoning may be used to show that the quantities $\frac{1}{n}H_{i0}^n$, $\frac{1}{n}L_{i0}^n$, and $\frac{1}{n}\frac{\partial L_{i0}^n}{\partial\theta_j}$ all converge stochastically to their respective mean values. It will then follow from equation (3) that the functions $\frac{1}{n}\frac{\partial \tilde{L}_{i0}^n}{\partial\theta_j}$ converge stochastically to the values $E_0\left[\frac{1}{n}\frac{\partial \tilde{L}_{i0}^n}{\partial\theta_j}\right]$. In fact, it can be shown [5] that any polynomial in these functions must converge stochastically to the same polynomial in their expected values. Hence, given any $\epsilon > 0$, the probability that the determinant $\left|\frac{1}{n}\frac{\partial \tilde{L}_{i0}^n}{\partial\theta_j}\right|$ differs in samples from Π_0 from the determinant $\left|E_0\left[\frac{1}{n}\frac{\partial \tilde{L}_{i0}^n}{\partial\theta_j}\right]\right|$ by more than ϵ can be made arbitrarily small by taking n sufficiently large. Similarly, the determinant $\left|\frac{1}{n}\frac{\partial \tilde{H}_{i0}^n}{\partial\theta_j}\right|$ converges stochastically to $\left|E_0\left[\frac{1}{n}\frac{\partial \tilde{H}_{i0}^n}{\partial\theta_j}\right]\right|$. Thus, given any two positive numbers ϵ, ϵ', we have the relation

$$P_0\left\{\left|\frac{1}{n}\frac{\partial \tilde{L}_{i0}^n}{\partial\theta_j}\right|^+ > \left|\frac{1}{n}\frac{\partial \tilde{H}_{i0}^n}{\partial\theta_j}\right|^+ - \epsilon\right\} > 1 - \epsilon'$$

(where + indicates the absolute values of the determinants), provided n is sufficiently large.

As in the one parameter case, the restrictions which have been put on the class of functions L and H are not entirely necessary. But it is difficult to replace them by any other set of conditions which are not obviously *ad hoc*. Let us now summarize the above results.

THEOREM 1. *If the functions L_i^n and H_i^n satisfy the conditions* (a), $\cdots$ (e), *and if*

(f) *the functions $\frac{1}{n}\frac{\partial \tilde{L}_{i0}^n}{\partial\theta_j}$ and $\frac{1}{n}\frac{\partial \tilde{H}_{i0}^n}{\partial\theta_j}$ converge stochastically to their mean values;*

(g) *the large sample distribution of the functions $\frac{1}{\sqrt{n}}\tilde{L}_{i0}^n$ is the same as that of the functions $\frac{1}{\sqrt{n}}\tilde{H}_{i0}^n$ and is independent of the θ_0's;*

then in large samples the confidence regions derived from the $\tilde{L}$'s will almost certainly be smaller than those derived from the $\tilde{H}$'s, in the sense that

$$\lim_{n\to\infty} P_0\left\{\left|\frac{1}{n}\frac{\partial \tilde{L}_{i0}^n}{\partial\theta_j}\right|^+ > \left|\frac{1}{n}\frac{\partial \tilde{H}_{i0}^n}{\partial\theta_j}\right|^+\right\} = 1$$

unless there is linear dependence between the L's and H's.

THEOREM 2. *The assumptions of Theorem 1 will be satisfied if the L_i and H_i are of class K', are linearly independent, and satisfy vi).*

THEOREM 3. *For the case of only one unknown parameter, the relation*

$$\left\{E_0\left[\frac{\partial \bar{L}_{10}^n}{\partial \theta_1}\right]\right\}^2 \geq \left\{E_0\left[\frac{\partial \tilde{H}_{10}^n}{\partial \theta_1}\right]\right\}^2$$

$\left(\text{equality holding only in case } H_1^n \equiv c \frac{\partial L^n}{\partial \theta_1}\right)$ *can be derived under assumptions* (a), $\cdots$, (e) *alone. Its interpretation in terms of smallest average confidence intervals depends, however, on whether or not (g) is satisfied.*

At first sight it may appear that the functions

$$\psi_{ni} = \frac{1}{\sqrt{n}} \sum_{j=1}^{h} b^{nij} H_j^n$$

to which these theorems apply are too complicated to be of any practical use, involving as they do the square root of the inverse of the matrix

$$\| B_{ij}^n \| = \frac{1}{n} \| E[H_i^n H_j^n] \|.$$

But in employing the method of fiducial argument in the several parameter case there is no need to take the region S in the ψ space to be an interval

$$\underline{\psi}_i < \psi_i < \bar{\psi}_i .$$

Instead, we may take S to be the interior of the sphere

$$\sum_{i=1}^{h} \psi_i^2 < R^2 \tag{7}$$

This enables us to avoid the computation of the b^{nij}; for

$$\sum_{i=1}^{h} \psi_{ni}^2 = \frac{1}{n} \sum_{i,j,k=1}^{h} b^{nij} b^{nik} H_j^n H_k^n = \frac{1}{n} \sum_{j,k=1}^{h} B^{njk} H_j^n H_k^n$$

where $\| B^{njk} \|$ is the inverse of $\| B_{jk}^n \|$.

To indicate more precisely how the function $\sum_{i=1}^{h} \psi_{ni}^2$ may be used to determine confidence regions for the parameter point θ, we note that if the distribution law of the ψ_{ni} tends to the form

$$(2\pi)^{-\frac{h}{2}} e^{-\frac{1}{2}\sum \psi_i^2}$$

then $\sum_{i=1}^{h} \psi_{ni}^2$, which is identically equal to $\frac{1}{n} \sum_{i,j} B^{nij} H_i^n H_j^n$, is approximately distributed according to the χ^2 law with h degrees of freedom. We then have

$$P\left(\frac{1}{n} \sum_{i,j} B^{nij} H_i^n H_j^n < \chi_\alpha^2\right) = \alpha \tag{8}$$

approximately, where χ_α is given by the relation

$$\frac{1}{2\Gamma(\frac{1}{2}h)} \int_0^{\chi_\alpha} (\tfrac{1}{2}\chi^2)^{\frac{1}{2}h-1} e^{-\frac{1}{2}\chi^2} d\chi^2 = \alpha.$$

The confidence region T corresponding to a particular sample $0_n : x_1, x_2, \cdots, x_n$ consists of those points in the θ space for which $\frac{1}{n}\sum_{i,j} B^{nij} H_i^n H_j^n < \chi_\alpha^2$ when the x's are substituted in the H's. Since the region T depends on the sample, it is essentially a random variable and the probability is α that T will include the point θ_0, that is, the point in the θ-space corresponding to the values of the θ's in the population.

For example, suppose the population Π is known to have the multinomial distribution law

$$f(x_0, \cdots, x_h; p_0, \cdots, p_h) = p_0^{x_0} \cdots p_h^{x_h}$$

In this case each x has but two possible values, 0 and 1, and

$$x_0 + \cdots + x_h = 1, \qquad p_0 + \cdots + p_h = 1. \tag{9}$$

The likelihood function for random samples 0_n drawn from Π has for its logarithm

$$L^n = \sum_{\nu=0}^{h} n_\nu \log p_\nu$$

where $n_\nu = \sum_{\alpha=1}^{n} x_{\nu\alpha}$, $x_{\nu\alpha}$ being the value of x_ν for the α-th observation. Because of (9) there are only h independent parameters, say p_i $(i = 1, \cdots, h)$. Thus

$$L_i^n = \frac{n_i}{p_i} - \frac{n_0}{p_0}$$

and

$$A_{ij}^n = \frac{\delta_{ij}}{p_i} + \frac{1}{p_0}$$

where δ_{ij} is unity if $i = j$ and 0 if $i \neq j$. It is not necessary to compute the a^{nij}, for, as we have seen,

$$\sum_{i=1}^{h} (\psi_i^n)^2 = \frac{1}{n} \sum_{i,j=1}^{n} A^{nij} L_i^n L_j^n$$

And one can immediately verify that

$$A^{nij} = \delta_{ij} p_i - p_i p_j$$

so that we have

$$\sum_{i=1}^{h} \psi_{ni}^2 = \frac{1}{n} \sum_{i,j=1}^{h} (\delta_{ij} p_i - p_i p_j) \left(\frac{n_i}{p_i} - \frac{n_0}{p_0}\right)\left(\frac{n_j}{p_j} - \frac{n_0}{p_0}\right) \tag{10}$$

Since in this case the L_i^n satisfy the conditions of the lemma, we know that $\sum_{i=1}^{h} \psi_{ni}^2$ is distributed, in large samples, approximately like χ^2 with h degrees of freedom.

As a matter of fact, (10) is precisely the Pearson χ^2 which is ordinarily used, in connection with the problem of deciding whether a sample supports the hypothesis that the population from which it has been drawn has specified values of the p's. For, making use of the fact that

$$\sum_{i=1}^{h} (n_i - np_i) + (n_0 - np_0) = 0$$

we find that

$$\frac{n_i}{p_i} - \frac{n_0}{p_0} = \sum_{j=1}^{h} A_{ij}^n(n_j - np_j)$$

so that $\sum_{i=1}^{h} \psi_{ni}^2$ reduces to

$$\frac{1}{n} \sum_{i,j=1}^{h} A_{ij}^n(n_i - np_i)(n_j - np_j) = \sum_{\nu=0}^{h} (n_\nu - np_\nu)^2/np_\nu$$

which is the familiar form. Thus in particular the Pearson χ^2 is the best fiducial function of its type which can be formed from H's satisfying Theorem 1, in the sense that for sufficiently large samples its constituent functions $\tilde{L}_i^n$ will almost certainly have a greater jacobian with respect to the parameters p_i than will the corresponding $\tilde{H}_i^n$ computed from a set of H_i^n independent of the L_i^n.

The confidence regions determined by (8) when the H_i^n are replaced by the L_i^n have an associated optimum property which may be stated as

THEOREM 4: *Let Δ_0 denote the differential of $\frac{1}{n} \sum_{i,j} B^{nij} H_i^n H_j^n$ with respect to the θ_i, evaluated at the true parameter point θ_0. Let Δ_0^* be the corresponding differential when the H_i^n are replaced by the L_i^n. Let the H_i^n and L_i^n satisfy conditions* (i), (ii), $\cdots$, (vi) *and let the mean value of the product of two, three or four factors taken from the set $\left\{h_{i0}, \frac{\partial h_{j0}}{\partial \theta_k}\right\}$ be finite, no product containing more than two factors of the type $\frac{\partial h_{i0}}{\partial \theta_j}$. Let similar assumptions hold for the set $\left\{l_{i0}, \frac{\partial l_{i0}}{\partial \theta_i}\right\}$ where $l_{i0} = \frac{\partial \log dF_0}{\partial \theta_i}$. Then if n is sufficiently large*

$$E_0\left(\frac{1}{n}\Delta_0^{*2}\right) - E_0\left(\frac{1}{n}\Delta_0^2\right) \geq 0 \tag{11}$$

The equality in (11) *will hold for all differential vectors if and only if each h_{i0} is a linear function of the l_{i0}.*

This theorem can be proved in a straightforward manner by using the following characteristic functions

$$\varphi_H = \exp\left(i\sum_{i=1}^h t_i H_{i0}^n + i\sum_{i,j=1}^h u_{ij}\frac{\partial H_{i0}^n}{\partial\theta_j}\right)$$
$$= \left[\exp\left(i\sum_{i=1}^h t_i h_{i0} + i\sum_{i,j=1}^h u_{ij}\frac{\partial h_{i0}}{\partial\theta_j}\right)\right]^n$$
$$\varphi_L = \exp\left(i\sum_{i=1}^h t_i L_i^n + i\sum_{i,j=1}^h u_{ij}\frac{\partial L_i^n}{\partial\theta_j}\right)$$
$$= \left[\exp\left(i\sum_{i=1}^h t_i l_{i0} + i\sum_{i,j=1}^h u_{ij}\frac{\partial l_{i0}}{\partial\theta_j}\right)\right]^n,$$

where $u_{ij} \equiv u_{ji}$. Now

$$\Delta_0 = \frac{1}{n}\sum_{i,j,k=1}^h \frac{\partial B^{nij}}{\partial\theta_k} H_i^n H_j^n\, d\theta_k + \frac{2}{n}\sum_{i,j,k=1}^h B^{nij}\frac{\partial H_i^n}{\partial\theta_k} H_j^n\, d\theta_k$$

with a similar expression for Δ_0^*. The problem of finding the mean values $E\left(\frac{1}{n}\Delta_0^2\right)$ and $E\left(\frac{1}{n}\Delta_0^{*2}\right)$ is a matter of evaluating a set of fourth order derivatives of φ_H and φ_L at $t_i = 0$, $u_{ij} = 0$.

If the appropriate differentiations are carried out it is found that

$$E_0(\Delta_0^2) = 4n\left[\sum_{i,j,k,l} B_{kl0} C_{ki0} C_{lj0}\, d\theta_i d\theta_j + 0\left(\frac{1}{n}\right)\right]$$
$$E_0(\Delta_0^{*2}) = 4n\left[\sum_{i,j,k,l} A_{ij0}\, d\theta_i d\theta_j + 0\left(\frac{1}{n}\right)\right]$$

where $A_{ij0} = E_0[l_{i0}l_{j0}]$, $B_{kl0} = E_0[h_{k0}h_{l0}]$, $C_{ki0} = E_0[h_{k0},\ l_{i0}]$. Denoting $E_0\left(\frac{1}{n}\Delta_0^{*2}\right) - E_0\left(\frac{1}{n}\Delta_0^2\right)$ by δ, we have

$$\delta = 4\left\{\sum_{i,j} M_{ij0}\, d\theta_i d\theta_j + 0\left(\frac{1}{n}\right)\right\}$$

where $\|M_{ij0}\| = \|A_{ij0} - \sum_{k,l} B_0^{kl} C_{ki0} C_{lj0}\|$. If the h_{k0} and l_{i0} are linearly independent then $\|M_{ij0}\|$ is a positive definite matrix and hence $\sum_{i,j} M_{ij0}\, d\theta_i\, d\theta_j \equiv \delta'$ say, will be non-negative and can vanish only when all $d\theta_i$ are zero. If each h_{k0} is a linear combination of the l_{i0} and if the h_{k0} are linearly independent, then each l_{i0} is a linear combination of the h_{k0}. In this case it can be readily shown that every element in $\|M_{ij0}\|$ will vanish, and hence $\delta' \equiv 0$.

In case some of the h_{i0} are linearly dependent on the l_{j0}, it can be shown that δ' is positive semidefinite, that is, there exists no differential vector for which δ' is negative, although there will exist non-zero differential vectors for which δ' is zero.

It can be shown under the assumptions made in Theorem 4 that $\frac{1}{n}(\Delta_0^{*2} - \Delta_0^2)$ actually converges stochastically to $4\delta'$, and thus if the h_{i0} and l_{i0} are linearly independent, the difference $\frac{1}{n}(\Delta_0^{*2} - \Delta_0^2)$ converges stochastically to a positive number. Stated in another way: for sufficiently large samples, the square of the differential change in $\frac{1}{n}\sum_{i,j} A^{nij}L_i^n L_j^n$, for a given change $d\theta_i$ in the θ_i from the values θ_{i0}, will almost certainly exceed that of $\frac{1}{n}\sum_{i,j} B^{nij}H_i^n H_j^n$. The statistical interpretation of this result amounts to the following: by taking sufficiently large samples, we can make it as certain as we please that the confidence regions for locating θ_0 determined by using $\frac{1}{n}\sum_{i,j} A^{nij}L_i^n L_j^n$ in (8) will be smaller than those determined by using $\frac{1}{n}\sum_{i,j} B^{nij}H_i^n H_j^n$ in (8).

REFERENCES

[1] S. S. Wilks, "Shortest Average Confidence Intervals from Large Samples," *Annals of Mathematical Statistics*, vol. IX, No. 3 (1938), pp. 166–175.

[2] S. S. Wilks, "Optimum Fiducial Regions for Simultaneous Estimation of Several Population Parameters from Large Samples," (Abstract), *Annals of Mathematical Statistics*, vol. X, No. 1 (1939), p. 85.

[3] C. C. MacDuffee, *The Theory of Matrices*, Ergebnisse Series, vol. II, No. 5, Berlin, Julius Springer (1933), p. 87.

[4] H. Cramér, *Random Variables and Probability Distributions*, Cambridge Tracts in Mathematics and Mathematical Physics, No. 36, Cambridge University Press, 1937.

[5] J. L. Doob, "Limiting Distributions of Certain Statistics," *Annals of Mathematical Statistics*, vol. VI, No. 3 (1935), p. 164.

PRINCETON UNIVERSITY, PRINCETON, N. J.