ESTIMATION OF A PARAMETER WHEN THE NUMBER OF UNKNOWN PARAMETERS INCREASES INDEFINITELY WITH THE NUMBER OF OBSERVATIONS

By Abraham Wald

Columbia University

Summary. Necessary and sufficient conditions are given for the existence of a uniformly consistent estimate of an unknown parameter θ when the successive observations are not necessarily independent and the number of unknown parameters involved in the joint distribution of the observations increases indefinitely with the number of observations. In analogy with R. A. Fisher's information function, the amount of information contained in the first n observations regarding θ is defined. A sufficient condition for the non-existence of a uniformly consistent estimate of θ is given in section 3 in terms of the information function. Section 4 gives a simplified expression for the amount of information when the successive observations are independent.

2. Introduction. J. Neyman has recently treated the following estimation problem[1]: Let $X_1, X_2, \cdots$, etc. be a sequence of independent chance variables the distribution of each of which depends on some unknown parameters. Two kinds of parameters are distinguished, structural and incidental parameters. A parameter θ is called structural if there exists an infinite subsequence of the sequence $\{X_i\}$ such that the distribution of each of the chance variables in the subsequence depends on θ. Any parameter which is not structural is called incidental. Neyman has considered the case when there are a finite number of structural parameters, say $\theta_1, \cdots, \theta_s$ and an infinite sequence $\{\xi_i\}$, $(i = 1, 2, \cdots$, ad inf.), of incidental parameters. He has studied the problem of consistent and efficient estimation of the structural parameters and has obtained several interesting results. He has shown, among others, that the maximum likelihood estimate of a structural parameter θ need not be consistent, even when consistent estimates of θ exist. Neyman has also given a method for obtaining consistent estimates of the structural parameters. This method, however, is applicable only under certain restrictive conditions.

In this paper we shall consider a more general case than that treated by Neyman, but we shall concentrate on one aspect of the problem, namely that of the existence of consistent estimates.

Let $\{X_i\}$, $(i = 1, 2, \cdots$, ad inf.), be a sequence of chance variables, not necessarily independent of each other. It is assumed that for each n the chance variables $X_1, \cdots, X_n$ admit a joint probability density function $p_n(x_1, \cdots, x_n \mid \theta, \xi_1, \cdots, \xi_n)$ where $\theta, \xi_1, \xi_2, \cdots$, etc. are unknown parameters.[2]

---

[1] Address given by J. Neyman at the meeting of the Institute of Mathematical Statistics in Atlantic City, January, 1947.

[2] While θ is assumed to be a real variable, we admit ξ_i to be a finite dimensional vector, i.e., $\xi_i = (\xi_{i1}, \ldots, \xi_{ik_i})$ where k_i may be any finite positive integer.

We shall require that the consistency relations among the density functions $p_1, p_2, \cdots$, etc. be fulfilled, i.e.,

$$\int_{-\infty}^{+\infty} p_{n+1}\, dx_{n+1} = p_n, \qquad (n = 1, 2, \cdots, \text{ad inf.}). \tag{1.1}$$

It should be remarked that it is not postulated that p_n actually depends on all the parameters that appear as arguments in p_n. It is merely assumed that p_n does not depend on any parameter that does not appear as an argument in p_n, i.e., p_n does not depend on ξ_i for any $i > n$. It follows, however, from (1.1) that if p_n depends on a parameter ξ, then also p_m depends on ξ for any $m > n$.

Neyman's definition of structural and incidental parameters can be extended to the case of dependent observations considered here by saying that the distribution of X_i does not depend on a parameter ξ if and only if the conditional distribution of X_i for any given values of $X_1, \cdots, X_{i-1}$ does not depend on ξ. It is not postulated that each of the parameters $\xi_1, \xi_2, \cdots$, etc. is incidental; some of them may be structural. We shall not make an explicit distinction between structural and incidental parameters, since for the purposes of the present paper this does not seem to be necessary.

In this paper we shall deal with the problem of formulating conditions under which a uniformly consistent estimate of θ exists. A statistic $t_n(x_1, \cdots, x_n)$ is said to be a uniformly consistent estimate of θ if for any positive δ

$$\lim_{n=\infty} \text{prob. } \{|t_n - \theta| < \delta\} = 1 \tag{1.2}$$

uniformly in θ and the ξ's.

In section 2 a necessary and sufficient condition is given for the existence of a uniformly consistent estimate of θ. In section 3 the amount of information supplied by the first n observations concerning θ is defined. It is then shown that if the amount of information is a bounded function of n over a non-degenerate θ-interval, no uniformly consistent estimate of θ exists. Section 4 gives a simplified formula for the amount of information in the case when the X's are independently distributed.

2. A necessary and sufficient condition for the existence of a uniformly consistent estimate of θ. In deriving a necessary and sufficient condition for the existence of a uniformly consistent estimate of θ, use will be made of some results contained in a publication of the author [1] dealing with statistical decision functions which minimize the maximum risk. In [1] it is assumed that the domain of each of the unknown parameters is a closed and bounded set and that p_n is continuous jointly in all of its arguments. Thus, in order to be able to use the results obtained in [1], we shall have to make the same assumptions here. In what follows we shall, therefore, assume that each of the parameters $\theta, \xi_1, \xi_2, \cdots$, etc. is restricted to a finite closed interval and that p_n is a continuous function of $x_1, \cdots, x_n, \theta, \xi_1, \cdots, \xi_n$.

Let $[a, b]$ $(a < b)$ be the θ-interval to which the values of θ are restricted. Clearly, if $t_n(x_1, \cdots, x_n)$, $(n = 1, 2, \cdots, \text{ad inf.})$, is a uniformly consistent estimate of θ, then also t_n^* is a uniformly consistent estimate of θ when $t_n^* = t_n$ when $a \leqq t_n \leqq b$, $t_n^* =$ a when $t_n < a$ and $t_n^* = b$ when $t_n > b$. Thus, without loss of generality, we can restrict ourselves to estimates t_n which can take values only in the interval $[a, b]$. Uniform consistency of t_n is then equivalent with the condition

$$\lim_{n=\infty} E[(t_n - \theta)^2 \mid \theta, \xi_1, \cdots, \xi_n] = 0 \tag{2.1}$$

uniformly in θ and the ξ's. For any chance variable u the symbol $E(u \mid \theta, \xi_1, \xi_2, \cdots)$ denotes the expected value of u when $\theta, \xi_1, \xi_2, \cdots$ are the true parameter values.

In [1] a non-negative function $W(t_n, \theta)$, called weight function, is introduced which expresses the loss suffered when t_n is the value of the estimate and θ is the true value of the parameter. The risk is defined in [1] as the expected value of the loss, i.e., the risk is given by

$$r_n(\theta, \xi_1, \cdots, \xi_n) = E[W(t_n, \theta) \mid \theta, \xi_1, \cdots, \xi_n]. \tag{2.2}$$

If we put $W(t_n, \theta) = (t_n - \theta)^2$, we have

$$r_n(\theta, \xi_1, \cdots, \xi_n) = E[(t_n - \theta)^2 \mid \theta, \xi_1, \cdots, \xi_n]. \tag{2.3}$$

It can easily be verified that Assumptions 1–4 in section 3 of [1] are fulfilled for the weight function $W(t_n, \theta) = (t_n - \theta)^2$.[3] Thus, all results obtained in [1] can be applied to the risk function given in (2.3). According to Theorem 4.1 in [1] the risk function given in (2.3) is a continuous function of $\theta, \xi_1, \cdots, \xi_n$ for any arbitrary estimate t_n. We shall denote the maximum of (2.3) with respect to $\theta, \xi_1, \cdots, \xi_n$ by $r_n[t_n]$. Thus $r_n[t_n]$ is a functional which associates a non-negative value with any estimate function t_n.

It follows from (2.1) that t_n is a uniformly consistent estimate of θ if and only if

$$\lim_{n=\infty} r_n[t_n] = 0. \tag{2.4}$$

For any θ and for any n let $F_n(\xi_1, \cdots, \xi_n \mid \theta)$ be a cumulative distribution function of $\xi_1, \cdots, \xi_n$. Let, furthermore,

$$\begin{aligned} &q_n(x_1, \cdots, x_n \mid \theta, F_n) \\ &\quad = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} p_n(x_1, \cdots, x_n \mid \theta, \xi_1 \cdots, \xi_n)\, dF_n(\xi_1, \cdots, \xi_n \mid \theta). \end{aligned} \tag{2.5}$$

We do not require that $F_1, F_2, \cdots$, etc. satisfy the consistency relations, i.e., $\lim_{\xi_{n+1}=\infty} F_{n+1}(\xi_1, \cdots, \xi_{n+1} \mid \theta)$ is not necessarily equal to $F_n(\xi_1, \cdots, \xi_n \mid \theta)$.

[3] In verifying Assumption 4, we may assume that p_n is always > 0, since for any given values $\theta, \xi_1, \ldots, \xi_n$ we may restrict the domain of $(x_1, \ldots, x_n)$ to the subset of the sample space where $p_n > 0$.

Hence, also the distributions q_n do not necessarily satisfy the consistency relations. Clearly

$$r_n[t_n] \geqq \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} (t_n - \theta)^2 q_n(x_1, \cdots, x_n \mid \theta, F_n)\, dx_1, \cdots, dx_n \tag{2.6}$$

for any θ and any F_n. Hence, (2.4) and (2.6) imply that if t_n is a uniformly consistent estimate of θ, then t_n remains a uniformly consistent estimate of θ also when q_n is the distribution of $X_1, \cdots, X_n$ for any arbitrary choice of F_n.

For each n let $C_n(\theta, \xi_1, \cdots, \xi_n)$ be a joint cumulative distribution function of $\theta, \xi_1, \cdots, \xi_n$. If this is regarded as an a priori distribution of $\theta, \xi_1, \cdots, \xi_n$, and if our aim is to choose t_n so that

$$\begin{aligned} &E(t_n - \theta)^2 \\ &= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} (t_n - \theta)^2 p_n(x_1, \cdots, x_n \mid \theta, \xi_1, \cdots, \xi_n)\, dC_n\, dx_1 \cdots dx_n \end{aligned} \tag{2.7}$$

is a minimum, then the best choice of t_n is to put it equal to the a posteriori mean value of θ. Let $\overset{*}{t}_n(x_1, \cdots, x_n; C_n)$ denote the a posteriori mean value of θ when C_n is the a priori distribution, i.e.,

$$\overset{*}{t}_n(x_1, \cdots, x_n; C_n) = \frac{\int \theta p_n(x_1, \cdots, x_n \mid \theta, \xi_1, \cdots, \xi_n)\, dC_n}{\int p_n(x_1, \cdots, x_n \mid \theta, \xi_1 \cdots, \xi_n)\, dC_n} \tag{2.8}$$

where the integration is to be taken over the whole domain of the parameters $\theta, \xi_1, \cdots, \xi_n$. Let, furthermore, $\bar{r}_n[C_n]$ denote the value of (2.7) when $t_n = \overset{*}{t}_n(x_1, \cdots, x_n; C_n)$. According to Theorem 4.4 in [1] there exists a particular distribution C_n^0, called a least favorable distribution, such that

$$\bar{r}_n[C_n] \leqq \bar{r}_n[C_n^0] \tag{2.9}$$

for all C_n. Let

$$t_n^0(x_1, \cdots, x_n) = \overset{*}{t}_n(x_1, \cdots x, _n; C_n^0). \tag{2.10}$$

It follows from Theorems (4.5) and (5.1) in [1] that for any estimate t_n we have

$$r_n[t_n] \geqq r_n[t_n^0] = \bar{r}_n(C_n^0). \tag{2.11}$$

Hence, a necessary and sufficient condition for the existence of a uniformly consistent estimate of θ is that

$$\lim_{n=\infty} \bar{r}_n[C_n^0] = 0. \tag{2.12}$$

Let $F_n(\xi_1, \cdots, \xi_n \mid \theta)$ denote the conditional cumulative distribution of $\xi_1, \cdots, \xi_n$ for given θ that results from the joint distribution $C_n(\theta, \xi_1, \cdots \xi_n)$ and let $F_n^0(\xi_1, \cdots, \xi_n \mid \theta)$ correspond to $C_n^0(\theta, \xi_1, \cdots, \xi_n)$. Clearly, any uniformly consistent estimate of θ with respect to $p_n(x_1, \cdots, x_n \mid \theta, \xi_1, \cdots, \xi_n)$

is a uniformly consistent estimate also with respect to $q_n(x_1, \cdots, x_n \mid \theta, F_n)$ for any F_n. On the other hand, if $q_n(x_1, \cdots, x_n \mid \theta, F_n^0)$ admits a uniformly consistent estimate of θ, equation (2.12) must hold and, therefore, $p_n(x_1, \cdots, x_n \mid \theta, \xi_1, \cdots, \xi_n)$ admits a uniformly consistent estimate of θ. Hence we arrive at the following theorem:

THEOREM 2.1. *A necessary and sufficient condition that*

$$p_n(x_1, \cdots, x_n \mid \theta, \xi_1, \cdots, \xi_n)$$

admit a uniformly consistent estimate of θ is that $q_n(x_1, \cdots, x_n \mid \theta, F_n)$ admit a uniformly consistent estimate of θ for any arbitrary choice of F_n.

3. Amount of information contained in the first n observations concerning the parameter θ. We shall make the following assumptions:

Assumption 1. The first two derivatives of $p_n(x_1, \cdots, x_n \mid \theta, \xi_1, \cdots, \xi_n)$ with respect to θ exist.

Assumption 2. We have

$$\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \operatorname*{Max}_{\theta} \left| \frac{\partial p_n}{\partial \theta} \right| dx_1 \cdots dx_n < \infty \tag{3.1}$$

and

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \operatorname*{Max}_{\theta} \left| \frac{\partial^2 p_n}{\partial \theta^2} \right| dx_1 \cdots dx_n < \infty \tag{3.2}$$

for any n.

Assumption 3. The integral

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial^2 \log q_n(x_1, \cdots, x_n \mid \theta, F_n)}{\partial \theta^2} q_n(x_1, \cdots, x_n \mid \theta, F_n)\, dx_1 \cdots dx_n$$

exists for any θ, F_n and n where q_n is defined by (2.5).

Since

$$\frac{\partial^2 \log q_n}{\partial \theta^2} = \frac{1}{q_n} \frac{\partial^2 q_n}{\partial \theta^2} - \left(\frac{\partial \log q_n}{\partial \theta} \right)^2$$

and since, because of Assumptions 1 and 2,

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial^2 q_n}{\partial \theta^2} dx_1 \cdots dx_n = 0,$$

we have

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial^2 \log q_n}{\partial \theta^2} q_n\, dx_1 \cdots dx_n$$

$$= - \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{+\infty} \left(\frac{\partial \log q_n}{\partial \theta} \right)^2 q_n\, dx_1 \cdots dx_n.$$

Let

$$(3.4)\quad c_n(\theta) = \underset{F_n}{\text{g.l.b.}} \left\{ - \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\frac{\partial^2 \log q_n}{\partial \theta^2} \right) q_n \, dx_1 \cdots dx_n \right\}.$$

Clearly $c_n(\theta) \geqq 0$. We shall now show that

$$(3.5)\qquad c_{n+1}(\theta) \geqq c_n(\theta) \qquad \text{for } n = 1, 2, \cdots, \text{ad inf.}$$

In fact, we can write

$$(3.6)\quad \frac{-\partial^2 \log q_{n+1}(x_1, \cdots, x_{n+1} \mid \theta, F_{n+1})}{\partial \theta^2} = - \frac{\partial^2 \log q_n(x_1, \cdots, x_n \mid \theta, F_n^*)}{\partial \theta^2} - \frac{\partial^2 \log f_{n+1}(x_{n+1} \mid x_1, \cdots, x_n, \theta, F_{n+1})}{\partial \theta^2}$$

where $F_n^* = \lim_{\xi_{n+1}=\infty} F_{n+1}(\xi_1, \cdots, \xi_{n+1} \mid \theta)$ and $f_{n+1}(x_{n+1} \mid x_1, \cdots, x_n, \theta, F_{n+1})$ is the conditional probability density function of X_{n+1} given the values of $x_1, \cdots, x_n$ and assuming that the joint density function of $X_1, \cdots, X_{n+1}$ is given by $q_{n+1}(x_1, \cdots, x_{n+1} \mid \theta, F_{n+1})$. Since $c_n(\theta) \leqq$ expected value of

$$- \frac{\partial^2 \log q_n(x_1, \cdots, x_n \mid \theta, F_n^*)}{\partial \theta^2}$$

and since the expected value of $- \frac{\partial^2 \log f_{n+1}}{\partial \theta^2}$ is $\geqq 0$, inequality (3.5) must hold.

In analogy with R. A. Fisher's information function, we shall call $c_n(\theta)$ the amount of information contained in the first n observations regarding θ. We shall now prove the following theorem:

THEOREM 3.1. *If* $\lim_{n=\infty} c_n(\theta) \leqq c < \infty$ *over a finite non-degenerate* θ*-interval* I, *then there is no uniformly consistent estimate of* θ.

PROOF. If for any n, $c_n(\theta) \leqq c < \infty$ over the interval I, for each n there exists a distribution $F_n(\xi_1, \cdots, \xi_n \mid \theta)$ such that

$$(3.7)\quad 0 \leqq - \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial^2 \log q_n(x_1, \cdots, x_n \mid \theta, F_n)}{\partial \theta^2} \cdot q_n(x_1, \cdots, x_n \mid \theta, F_n) \, dx_1 \cdots dx_n \leqq c + 1$$

for all n and for all θ in I. Let t_n be any estimate and let

$$(3.8)\quad \begin{aligned} b_n(\theta) = E(t_n - \theta) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (t_n - \theta) q_n(x_1, \cdots, x_n \mid \theta, F_n) \, dx_1 \cdots dx_n \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} t_n q_n(x_1, \cdots, x_n \mid \theta, F_n) \, dx_1 \cdots dx_n - \theta. \end{aligned}$$

Since t_n is bounded, it follows from Assumptions 1 and 2 that $\frac{db_n(\theta)}{d\theta}$ exists and is a continuous function of θ. According to a theorem by Cramér [2] we have

$$(3.9)\qquad E(t_n - \theta)^2 = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (t - \theta)^2 q_n \, dx_1 \cdots dx_n \geqq \frac{\left(1 + \frac{db_n}{\partial\theta}\right)^2}{c + 1}$$

for all θ in I. Thus, in order that $\lim_{n=\infty} E(t_n - \theta)^2 = 0$ uniformly in θ, we must have

$$(3.10)\qquad \lim_{n=\infty} \frac{db_n(\theta)}{d\theta} = -1$$

uniformly in θ over I. Let I be the interval ranging from g to h ($g < h$). From (3.10) it follows that

$$(3.11)\qquad \lim_{n=\infty} [b_n(h) - b_n(g)] = g - h.$$

Hence

$$\liminf_{n=\infty} \max_{\theta \text{ in } I} [b_n(\theta)]^2 \geqq \frac{(g - h)^2}{4}.$$

Since $E(t_n - \theta)^2 \geqq [b_n(\theta)]^2$, $E(t_n - \theta)^2$ cannot converge to zero uniformly in θ and Theorem 3.1 is proved.

4. Formula for $c_n(\theta)$ when $p_n(x_1, \cdots, x_n \mid \theta, \xi_1, \cdots, \xi_n)$ is equal to $\varphi_1(x_1 \mid \theta, \xi_1)\, \varphi_2(x_2 \mid \theta, \xi_2) \cdots \varphi_n(x_n \mid \theta, \xi_n)$. Let $g_i(x_i \mid x_1, \cdots, x_{i-1}, \theta, F_n)$ be the conditional probability density of X_i given $x_1, \cdots, x_{i-1}$ when the joint density function of $x_1, \cdots, x_n$ is given by $q_n(x_1, \cdots, x_n \mid \theta, F_n)$, ($i \leqq n$). Clearly,

$$(4.1)\qquad -E\left(\frac{\partial^2 \log q_n}{\partial\theta^2}\right) = -\sum_{i=1}^{n} E\left(\frac{\partial^2 \log g_i}{\partial\theta^2}\right).$$

Now

$$(4.2)\qquad g_i(x_i \mid x_1, \cdots, x_{i-1}, \theta, F_n) = \int_{-\infty}^{\infty} \varphi_i(x_i \mid \theta, \xi_i)\, dH_i(\xi_i \mid x_1, \cdots, x_{i-1}, \theta, F_n)$$

where $H_i(\xi_i \mid x_1, \cdots, x_{i-1}, \theta, F_n)$ denotes the conditional cumulative distribution of ξ_i given $x_1, \cdots, x_{i-1}$, assuming that $F_n(\xi_1, \cdots, \xi_n \mid \theta)$ is the joint cumulative distribution of $\xi_1, \cdots, \xi_n$ and $p_n(x_1, \cdots, x_n \mid \theta, \xi_1, \cdots, \xi_n)$ is the joint density of $X_1, \cdots, X_n$ for any given values of $\theta, \xi_1, \cdots, \xi_n$.

It follows from (4.2) that

$$-\int_{-\infty}^{+\infty} \frac{\partial^2 \log g_i}{\partial\theta^2} g_i \, dx_i \geqq c_{ni}(\theta)$$

$$= \underset{C_i(\xi_i)}{\text{g.l.b.}} \left\{ -\int_{-\infty}^{+\infty} \left[\frac{\partial^2 \log \int_{-\infty}^{+\infty} \varphi_i(x_i \mid \theta, \xi_i)\, dC_i(\xi_i)}{\partial\theta^2} \int_{-\infty}^{\infty} \varphi_i \, dC_i \right] dx_i \right\}$$

where $C_i(\xi_i)$ may be any cumulative distribution of ξ_i. Hence

$$\underset{F_n}{\text{g.l.b.}}\left[-E\left(\frac{\partial^2 \log g_i}{\partial\theta^2}\right)\right] = c_{ni}(\theta) \tag{4.3}$$

and, therefore,

$$c_n(\theta) = \sum_{i=1}^{n} c_{ni}(\theta). \tag{4.4}$$

The quantity $c_{ni}(\theta)$ is simply the amount of information contained in the ith observation alone. Thus, formula (4.4) says that if $X_1, \cdots, X_n$ are independent, the total information contained in the first n observations is equal to the sum of the amounts of information contained in each of these observations singly.

REFERENCES

[1] A. Wald, "Statistical decision functions which minimize the maximum risk", *Annals of Math.*, Vol. 46 (1945).

[2] H. Cramér, *Mathematical Methods of Statistics*, Princeton Univ. Press, 1946.