NOTES

This section is devoted to brief research and expository articles and other short items.

NOTE ON THE CONSISTENCY OF THE MAXIMUM LIKELIHOOD ESTIMATE[1]

By Abraham Wald

Columbia University

1. Introduction. The problem of consistency of the maximum likelihood estimate has been treated in the literature by several authors (see, for example, Doob [1][2] and Cramér [2][3]). The purpose of this note is to give another proof of the consistency of the maximum likelihood estimate which may be of interest because of its relative simplicity and because of the easy verifiability of the underlying assumptions. The present proof has some common features with that given by Doob, insofar that both proofs make no differentiability assumptions (thus, not even the existence of the likelihood equation is postulated) and both are based on the strong law of large numbers and an inequality involving the log of a random variable. The assumptions in the present note are stronger in some respects than those made by Doob, but also the results obtained here are stronger. For the sake of simplicity, the author did not attempt to give the most general results or to weaken the underlying assumptions as much as possible. Remarks on possible generalizations are made in Section 4.

Let $X_1, X_2, \cdots$, etc. be independently and identically distributed chance variables. The most frequently considered case in the literature is that where the common distribution is known, except for the values of a finite number of

[1] The author wishes to thank J. L. Doob for several comments and suggestions he made in connection with this note.

[2] According to a communication from Doob, his Theorem 4 is incorrect, but is correct if the class of almost everywhere continuous functions in that theorem is replaced by a suitable class C of functions. The class C can be any one of a variety of classes; for example, the class of bounded almost everywhere continuous functions, or the larger class of almost everywhere continuous functions each of which is less than or equal in modulus to any one of a prescribed sequence of functions with finite expectations. His Theorem 5 on the consistency of the maximum likelihood is then dependent on the class C used in Theorem 4.

[3] The proof given by Cramér [2], pp. 500–504, establishes the consistency of some root of the likelihood equation but not necessarily that of the maximum likelihood estimate when the likelihood equation has several roots. Recently, Huzurbazar [3] showed that under certain regularity conditions the likelihood equation has at most one consistent solution and that the likelihood function has a relative maximum for such a solution. Since there may be several solutions for which the likelihood function has relative maxima, Cramér's and Huzurbazar's results taken together still do not imply that a solution of the likelihood equation which makes the likelihood function an absolute maximum is necessarily consistent.

parameters, $\theta^1, \theta^2, \cdots, \theta^k$. In this note we shall treat the parametric case. For any parameter point $\theta = (\theta^1, \cdots, \theta^k)$, let $F(x, \theta)$ denote the corresponding cumulative distribution function of X_i; i.e., $F(x, \theta) = \text{prob.}\ \{X_i < x\}$. The totality Ω of all possible parameter points is called the parameter space. Thus, the parameter space Ω is a subset of the k-dimensional Cartesian space.

It is assumed in this note that for any θ, the cumulative distribution function $F(x, \theta)$ admits an elementary probability law $f(x, \theta)$. If $F(x, \theta)$ is absolutely continuous, $f(x, \theta)$ denotes the density at x. If $F(x, \theta)$ is discrete, $f(x, \theta)$ is equal to the probability that $X_i = x$.

Throughout this note the following assumptions will be made.

ASSUMPTION 1. *$F(x, \theta)$ is either discrete for all θ or is absolutely continuous for all θ.*

Before formulating the next assumption, we shall introduce the following notations: for any θ and for any positive value ρ let $f(x, \theta, \rho)$ be the supremum of $f(x, \theta')$ with respect to θ' when $|\theta - \theta'| \leqq \rho$. For any positive r, let $\varphi(x, r)$ be the supremum of $f(x, \theta)$ with respect to θ when $|\theta| > r$. Furthermore, let $f^*(x, \theta, \rho) = f(x, \theta, \rho)$ when $f(x, \theta, \rho) > 1$, and $=1$ otherwise. Similarly, let $\varphi^*(x, r) = \varphi(x, r)$ when $\varphi(x, r) > 1$, and $=1$ otherwise.

ASSUMPTION 2. *For sufficiently small ρ and for sufficiently large r the expected values $\int_{-\infty}^{\infty} \log f^*(x, \theta, \rho)\, dF(x, \theta_0)$ and $\int_{-\infty}^{\infty} \log \varphi^*(x, r)\, dF(x, \theta_0)$ are finite where θ_0 denotes the true parameter point.*[4]

ASSUMPTION 3. *If $\lim_{i=\infty} \theta_i = \theta$, then $\lim_{i=\infty} f(x, \theta_i) = f(x, \theta)$ for all x except perhaps on a set which may depend on the limit point θ (but not on the sequence θ_i) and whose probability measure is zero according to the probability distribution corresponding to the true parameter point θ_0.*

ASSUMPTION 4. *If θ_1 is a parameter point different from the true parameter point θ_0, then $F(x, \theta_1) \neq F(x, \theta_0)$ for at least one value of x.*

ASSUMPTION 5. *If $\lim_{i=\infty} |\theta_i| = \infty$, then $\lim_{i=\infty} f(x, \theta_i) = 0$ for any x except perhaps on a fixed set (independent of the sequence θ_i) whose probability is zero according to the true parameter point θ_0.*

ASSUMPTION 6. *For the true parameter point θ_0 we have*

$$\int_{-\infty}^{\infty} |\log f(x, \theta_0)|\, dF(x, \theta_0) < \infty.$$

ASSUMPTION 7. *The parameter space Ω is a closed subset of the k-dimensional Cartesian space.*

ASSUMPTION 8. *$f(x, \theta, \rho)$ is a measurable function of x for any θ and ρ.*

It is of interest to note that if we forbid the dependence of the exceptional set on θ in Assumption 3, Assumption 8 is a consequence of Assumption 3, as can easily be verified.

---

[4] The measurability of the functions $f^*(x, \theta, \rho)$ and $\varphi^*(x, r)$ for any θ, ρ and r follows easily from Assumption 8.

In the discrete case, Assumption 8 is unnecessary. In fact, we may replace $f(x, \theta, \rho)$ everywhere by $\bar{f}(x, \theta, \rho)$ where $\bar{f}(x, \theta, \rho) = f(x, \theta, \rho)$ when $f(x, \theta_0) > 0$, and $\bar{f}(x, \theta, \rho) = 1$ when $f(x, \theta_0) = 0$. Here θ_0 denotes the true parameter point. Since $f(x, \theta_0) > 0$ only for countably many values of x, $\bar{f}(x, \theta, \rho)$ is obviously a measurable function of x.

In the absolutely continuous case, $F(x, \theta)$ does not determine $f(x, \theta)$ uniquely. If Assumptions 3, 5 and 8 hold for one choice of $f(x, \theta)$, they do not necessarily hold for another choice of $f(x, \theta)$. This is in a way undesirable, but assumptions of such nature are unavoidable if we want to insure the consistency of the maximum likelihood estimate. It is, however, possible to formulate assumptions which remain valid for all possible choices of $f(x, \theta)$ and which insure the consistency of the maximum likelihood estimate for a particular choice of $f(x, \theta)$. In this connection the following remark due to Doob is of interest. Let Assumptions 3′ and 5′ be the same as 3 and 5, respectively, except that the exceptional set is permitted to depend on the sequence θ_i. If 3′ and 5′ hold for one choice of $f(x, \theta)$, they also hold for any other choice. Doob has shown that Assumptions 3′ and 5′ insure the existence of a choice of $f(x, \theta)$ for which Assumptions 3, 5 and 8 hold. Thus, one may say that Assumptions 3′ and 5′ are the essential ones and the stronger assumptions 3, 5 and 8 are needed merely to exclude a "bad" choice of $f(x, \theta)$.

2. Some lemmas. In this section we shall prove some lemmas which will be used in the next section to obtain the main theorems. Let θ_0 be the true parameter point. By the expected value Eu of any chance variable u we shall mean the expected value determined under the assumption that θ_0 is the true parameter point. For any chance variable u, u' will denote the chance variable which is equal to u when $u > 0$ and equal to zero otherwise. Similarly, for any chance variable u, the symbol u'' will be used to denote the chance variable which is equal to u when $u < 0$ and equal to zero otherwise. We shall say that the expected value of u exists if $Eu' < \infty$. If the expected value of u' is finite but that of u'' is not, we shall say that the expected value of u is equal to $-\infty$.

LEMMA 1. *For any $\theta \neq \theta_0$ we have*

$$E \log f(X, \theta) < E \log f(X, \theta_0) \tag{1}$$

where X is a chance variable with the distribution $F(x, \theta_0)$.

PROOF. It follows from Assumption 2 that the expected values in (1) exist. Because of Assumption 6, we have

$$E \mid \log f(X, \theta_0) \mid < \infty. \tag{2}$$

If $E \log f(X, \theta) = -\infty$, Lemma 1 obviously holds. Thus, we shall merely consider the case when $E \log f(X, \theta) > -\infty$. Then

$$E \mid \log f(X, \theta) \mid < \infty. \tag{3}$$

Let $u = \log f(X, \theta) - \log f(X, \theta_0)$.[5] Clearly, $E \mid u \mid < \infty$. It i known that for

any chance variable u which is not equal to a constant (with probability one) and for which $E \mid u \mid < \infty$, we have[6]

$$Eu < \log Ee^u. \tag{4}$$

Since in our case

$$Ee^u \leqq 1, \tag{5}$$

and since u differs from zero on a set of positive probability (due to Assumption 4), we obtain from (4)

$$Eu < 0. \tag{6}$$

Thus, Lemma 1 is proved.

We shall now prove the following lemma.

LEMMA 2. $\lim_{\rho=0} E \log f(X, \theta, \rho) = E \log f(X, \theta)$.

PROOF. Let $f^*(x, \theta, \rho) = f(x, \theta, \rho)$ when $f(x, \theta, \rho) \geqq 1$, and $=1$ otherwise. Similarly, let $f^*(x, \theta) = f(x, \theta)$ when $f(x, \theta) \geqq 1$, and $=1$ otherwise. It follows from Assumption 3 that

$$\lim_{\rho=0} \log f^*(x, \theta, \rho) = \log f^*(x, \theta) \tag{7}$$

except perhaps on a set whose probability measure is zero. Since $\log f^*(x, \theta, \rho)$ is an increasing function of ρ, it follows from (7) and Assumption 2 that

$$\lim_{\rho=0} E \log f^*(X, \theta, \rho) = E \log f^*(X, \theta). \tag{8}$$

Let $f^{**}(x, \theta, \rho) = f(x, \theta, \rho)$ when $f(x, \theta, \rho) \leqq 1$, and $=1$ otherwise. Similarly, let $f^{**}(x, \theta) = f(x, \theta)$ when $f(x, \theta) \leqq 1$, and $=1$ otherwise. Clearly,

$$\mid \log f^{**}(x, \theta, \rho) \mid \leqq \mid \log f^{**}(x, \theta) \mid \tag{9}$$

and

$$\lim_{\rho=0} \log f^{**}(x, \theta, \rho) = \log f^{**}(x, \theta) \tag{10}$$

for all x except perhaps on a set whose probability measure is zero. The relation

$$\lim_{\rho=0} E \log f^{**}(X, \theta, \rho) = E \log f^{**}(X, \theta) \tag{11}$$

follows from (9) and (10) in both cases, when $E \log f^{**}(X, \theta)$ is finite and when $E \log f^{**}(X, \theta) = -\infty$. Lemma 2 is an immediate consequence of (8) and (11).

LEMMA 3. *The equation*

$$\lim_{r=\infty} E \log \varphi(X, r) = -\infty. \tag{12}$$

holds.

---

[5] It is of no consequence what value is assigned to u when $f(x, \theta)$ or $f(x, \theta_0)$ is zero, since the probability of such an event, because of (3), is zero.

[6] This is a generalization of the inequality between geometric and arithmetic means. See, for example, HARDY, LITTLEWOOD, POLYA, *Inequalities*, Cambridge 1934, p. 137, Theorem 184.

PROOF. It follows from Assumption 5 that

$$\lim_{r=\infty} \log \varphi(x, r) = -\infty, \tag{13}$$

for any x (except perhaps on a set of probability 0). Since according to Assumption 2,

$$E \log \varphi^*(X, r) < \infty, \tag{14}$$

and since $\log \varphi(x, r) - \log \varphi^*(x, r)$ and $\log \varphi^*(x, r)$ are decreasing functions of r, Lemma 3 follows easily from (13).

3. The main theorems. We shall now prove the following theorems.

THEOREM 1. *Let ω be any closed subset of the parameter space Ω which does not contain the true parameter point θ_0. Then*

$$\text{prob.} \left\{ \lim_{n=\infty} \frac{\operatorname{Sup}_{\theta \in \omega} f(X_1, \theta) f(X_2, \theta) \cdots f(X_n, \theta)}{f(X_1, \theta_0) f(X_2, \theta_0) \cdots f(X_n, \theta_0)} = 0 \right\} = 1. \tag{15}$$

PROOF. Let r_0 be a positive number chosen such that

$$E \log \varphi(X, r_0) < E \log f(X, \theta_0). \tag{16}$$

The existence of such a positive number follows from Lemma 3. Let ω_1 be the subset of ω consisting of all points θ of ω for which $|\theta| \leqq r_0$. With each point θ in ω_1 we associate a positive value ρ_θ such that

$$E \log f(X, \theta, \rho_\theta) < E \log f(X, \theta_0). \tag{17}$$

The existence of such a ρ_θ follows from Lemmas 1 and 2. Since the set ω_1 is compact, there exists a finite number of points $\theta_1, \cdots, \theta_h$ in ω_1 such that $S(\theta_1, \rho_{\theta_1}) + \cdots + S(\theta_h, \rho_{\theta_h})$ contains ω_1 as a subset. Here $S(\theta, \rho)$ denotes the sphere with center θ and radius ρ. Clearly,

$$0 \leqq \operatorname{Sup}_{\theta \in \omega} f(x_1, \theta) \cdots f(x_n, \theta) \leqq \sum_{i=1}^{h} f(x_1, \theta_i, \rho_{\theta_i}) \cdots f(x_n, \theta_i, \rho_{\theta_i}) + \varphi(x_1, r_0) \cdots \varphi(x_n, r_0).$$

Hence, Theorem 1 is proved if we can show that

$$\text{prob} \left\{ \lim_{n=\infty} \frac{f(X_1, \theta_i, \rho_{\theta_i}) \cdots f(X_n, \theta_i, \rho_{\theta_i})}{f(X_1, \theta_0) \cdots f(X_n, \theta_0)} = 0 \right\} = 1 \quad (i = 1, \cdots, h) \tag{18}$$

and

$$\text{prob} \left\{ \lim_{n=\infty} \frac{\varphi(X_1, r_0) \cdots \varphi(X_n, r_0)}{f(X_1, \theta_0) \cdots f(X_n, \theta_0)} = 0 \right\} = 1. \tag{19}$$

The above equations can be written as

$$\text{(20)}\quad \text{prob}\left\{\lim_{n=\infty}\sum_{\alpha=1}^{n}[\log f(X_\alpha, \theta_i, \rho_{\theta_i}) - \log f(X_\alpha, \theta_0)] = -\infty\right\} = 1 \qquad (i = 1, \cdots, h)$$

and

$$\text{(21)}\quad \text{prob}\left\{\lim_{n=\infty}\sum_{\alpha=1}^{n}[\log \varphi(X_\alpha, r_0(- \log f(X_\alpha, \theta_0)] = -\infty\right\} = 1.$$

These equations follow immediately from (16), (17) and the strong law of large numbers. This completes the proof of Theorem 1.

THEOREM 2. *Let* $\bar{\theta}_n(x_1, \cdots, x_n)$ *be a function of the observations* $x_1, \cdots, x_n$ *such that*

$$\text{(22)}\quad \frac{f(x_1, \bar{\theta}_n) \cdots f(x_n, \bar{\theta}_n)}{f(x_1, \theta_0) \cdots f(x_n, \theta_0)} \geqq c > 0 \text{ for all } n \text{ and for all } x_1, \cdots, x_n.$$

Then

$$\text{(23)}\quad \text{prob}\{\lim_{n=\infty} \bar{\theta}_n = \theta_0\} = 1.$$

PROOF. It is sufficient to prove that for any $\epsilon > 0$ the probability is one that all limit points $\bar{\theta}$ of the sequence $\{\bar{\theta}_n\}$ satisfy the inequality $|\bar{\theta} - \theta_0| \leqq \epsilon$. The event that there exists a limit point $\bar{\theta}$ of the sequence $\{\bar{\theta}_n\}$ such that $|\bar{\theta} - \theta_0| > \epsilon$ implies that $\operatorname{Sup}_{|\theta-\theta_0|\geqq\epsilon} f(x_1, \theta) \cdots f(x_n, \theta) \geqq f(x_1, \bar{\theta}_n) \cdots f(x_n, \bar{\theta}_n)$ for infinitely many n. But then

$$\text{(24)}\quad \frac{\operatorname{Sup}_{|\theta-\theta_0|\geqq\epsilon} f(x_1, \theta) \cdots f(x_n, \theta)}{f(x_1, \theta_0) \cdots f(x_n, \theta_0)} \geqq c > 0$$

for infinitely many n. Since, according to Theorem 1, this is an event with probability zero, we have shown that the probability is one that all limit points $\bar{\theta}$ of $\{\bar{\theta}_n\}$ satisfy the inequality $|\bar{\theta} - \theta_0| \leqq \epsilon$. This completes the proof of Theorem 2.

Since a maximum likelihood estimate $\hat{\theta}_n(x_1, \cdots, x_n)$, if it exists, obviously satisfies (22) with $c = 1$, Theorem 2 establishes the consistency of $\hat{\theta}_n(x_1, \cdots, x_n)$ as an estimate of θ.

4. Remarks on possible generalizations. The method given in this note can be extended to establish the consistency of the maximum likelihood estimates for certain types of dependent chance variables for which the strong law of large numbers remains valid.

The assumption that the parameter space Ω is a subset of a finite dimensional Cartesian space is unnecessarily restrictive. Let Ω be any abstract space. All of

our results can easily be shown to remain valid if Assumptions 3, 5 and 7 are replaced by the following one:

ASSUMPTION 9. *It is possible to introduce a distance* $\delta(\theta_1, \theta_2)$ *in the space* Ω *such that the following four conditions hold:*

(i) *The distance* $\delta(\theta_1, \theta_2)$ *makes* Ω *to a metric space*

(ii) $\lim_{i=\infty} f(x, \theta_i) = f(x, \theta)$ *if* $\lim_{i=\infty} \theta_i = \theta$ *for any* x *except perhaps on a set which may depend on* θ *(but not on the sequence* θ_i*) and whose probability measure is zero according to the probability distribution corresponding to the true parameter point* θ_0.

(iii) *If* θ_0 *is a fixed point in* Ω *and* $\lim_{i=\infty} \delta(\theta_i, \theta_0) = \infty$, *then* $\lim_{i=\infty} f(x, \theta_i) = 0$ *for any* x.

(iv) *Any closed and bounded subset of* Ω *is compact.*

REFERENCES

[1] J. L. DOOB, "Probability and statistics," *Trans. Amer. Math. Soc.*, Vol. 36 (1934).

[2] H. CRAMÉR, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, 1946.

[3] V. S. HUZURBAZAR, "The likelihood equation, consistency and the maxima of the likelihood function," *Annals of Eugenics*, Vol. 14 (1948).

ON WALD'S PROOF OF THE CONSISTENCY OF THE MAXIMUM LIKELIHOOD ESTIMATE

BY J. WOLFOWITZ

Columbia University

This note is written by way of comment on the pretty and ingenious proof of the consistency of the maximum likelihood estimate which is due to Wald and is printed in the present issue of the *Annals*. The notation of this paper of Wald's will henceforth be assumed unless the contrary is specified.

The consistency of the maximum likelihood estimate is a "weak" rather than a "strong" property, in the technical meaning which these words have in the theory of probability, i.e., it is a property of distribution functions rather than of infinite sequences of observations. Prof. Wald actually proves strong convergence, which is more than consistency. His proof uses the strong law of large numbers, and he remarks that his method "can be extended to establish consistency of the maximum likelihood estimates for certain types of dependent chance variables for which the strong law of large numbers remains valid." Below we shall use Wald's lemmas to give a proof of consistency which employs only the weak law of large numbers. Not only does this proof have the advantage of being expeditious, but it can be extended to a larger class of dependent chance variables.

The consistency of the maximum likelihood estimate follows from the following

THEOREM. *Let* η *and* ϵ *be given, arbitrarily small, positive numbers. Let* $S(\theta_0, \eta)$ *be the open sphere with center* θ_0 *and radius* η, *and let* $\Omega(\eta) = \Omega - S(\theta_0, \eta)$. *Let*