COWLES COMMISSION DISCUSSION PAPER: STATISTICS No. 352

Bayes and Minimax Interpretation of the Maximum Likelihood Estimation Criterion

by Leonid Hurwicz

December 15, 1950

0. This paper is largely inspired by Wald's recent book [1] (especially sections 5.1.1 and 5.1.3), as well as the papers by Hodges and Lehman [2], Wolfowitz [3], and Stein and Wald [4]. To a large extent the paper is expository in nature. One reason for writing it was the desire to see in a relatively simple setting how the game interpretation of the statistical decision problem works out in detail. The writer suspects that most of the (correct) results in this paper could be found, perhaps in a more general form, in the existing literature of the subject.

1. Nature's "strategy" known; no observations available.

1.1. Let Θ denote the space of all possible hypotheses (this corresponds to Wald's Ω), e.g., all points in the parameter space. Denote by $H = H(\theta)$ a cumulative distribution in Θ (this corresponds to the use on Nature's part a "mixed" strategy). The symbol $h = h(\theta)$ will denote the probability of a particular point in Θ when H is discrete and the probability density function when H is absolutely continuous.

1.2. When the true hypothesis is θ while the estimate is $\bar{\theta}$, we denote the ("penalty") weight function by $W(\theta, \bar{\theta})$. $W_{\alpha\beta}(\theta, \bar{\theta})$ in the following weight function: if $d(\theta, \bar{\theta}) \leq \beta$, then, $W_{\alpha\beta} = 0$, but if $d(\theta, \bar{\theta}) > \beta$, $W_{\alpha\beta} = [d(\theta, \bar{\theta})]^{\alpha}$ where $d(\theta, \bar{\theta})$ is the Euclidean distance. For a given estimator (decision function, statistician's "strategy") S, and given H, α, β, the average (with respect to H) risk may be written as

$$(1.2.1) \quad r_{\alpha\beta}(H,S) = \int [d(\theta, \bar{\theta}_S)]^{\alpha} \, dH(\theta),$$

where the integration extends overall $\theta \in \Theta$ such that $d(\theta, \bar{\theta}) > \beta$. ($\bar{\theta}_S$ denotes the estimate obtained by the statistician who has adopted the strategy S.) We shall denote by $S^{(\alpha\beta)}(H)$ a strategy minimizing $r_{\alpha\beta}(H,S)$ and we call it optimal for $W_{\alpha\beta}$ (given H).[1] When H is absolutely continuous we define $S^{(\infty)}(H) = \lim_{\beta \to 0} S^{(\alpha\beta)}(H)$ and we shall say that $S^{(\infty)}(H)$ is optimal for W_{∞}. For brevity's sake we shall write $S^{(\alpha 0)}(H) = S^{(\alpha)}$, $W_{\alpha 0} = W_{\alpha}$. Hence, in particular, $S^{(\infty)}(H) = S^{(0)}(H)$ is said to be optimal for W_0.[2]

Now let Θ be one-dimensional, so that $d(\theta, \bar{\theta}) = |\theta - \bar{\theta}|$. It is clear that for $\alpha = 0, 1, 2$ we have[3] $S^{(\alpha)}(H) = \mu_H^{(\alpha)}$ where $\mu_H^{(0)}, \mu_H^{(1)}, \mu_H^{(2)}$ denote respectively the mode, median, and mean of H. [Strictly speaking, $S^{(\alpha)}(H)$ is a probability distribution assigning to $\mu_H^{(\alpha)}$ the probability 1 of becoming $\bar{\theta}$.] Since in what follows $S^{(0)}(H)$ will be of primary interest, it may be noted that $S^{(0)}(H) = \mu_H^{(0)}$ regardless of the dimensionality of Θ.

2. Nature's "strategy" known; observations available.

Let X be the sample space and denote by $F(x \mid \theta)$ the cumulative distribution of the observations while $f(x \mid \theta)$ will again stand for the density function or probability of a specified value of x depending on whether F is discrete or absolutely continuous. As the notation indicates $f(x \mid \theta)$ is regarded as a conditional distribution of x given θ. Nature's strategy, as before, is denoted by H and will be called the a priori distribution of θ; it is the marginal distribution of θ. The joint probability (density) function is $f(x \mid \theta)\, h(\theta)$ and hence the a posteriori (conditional) distribution of θ

---

1. When Nature's strategy is regarded as known, the optimal strategy is called "Bayes optimal," or, more briefly, a "Bayes strategy." "Bayes optimal" is to be distinguished from "minimax optimal," i.e., optimal in the sense of the theory of games where H is not known.

2. For the absolutely continuous case W_0 is defined by the limiting process; for the discrete case $W_0(\theta, \bar{\theta}) = 1$ when $\theta \neq \bar{\theta}$, $W_0(\theta, \bar{\theta}) = 0$ when $\theta = \bar{\theta}$.

3. Since here $W_1(\theta, \bar{\theta}) = |\theta - \bar{\theta}|$, $W_2(\theta, \bar{\theta}) = (\theta - \bar{\theta})^2$.

(given x) is

$$(2.1) \qquad g_H(\theta \mid x) = \frac{f(x \mid \theta)\, h(\theta)}{\int_{\theta \in \Theta} f(x\,\theta)\, d\, H(\theta)}.$$

For the statistician who knows H and has observed x, the situation is analogous to that described in §1 except that $g_H(\theta \mid x)$ as defined in (2.1) must be substituted for $h(\theta)$ used in §1. More explicitly, we have $S^{(\alpha)}(G_H(\theta \mid x)) = \mu_{G_H}^{(\alpha)}$ where $G_H(\theta \mid x)$ is the cumulative distribution corresponding to $g_H(\theta \mid x)$.

3. Nature's "strategy" rectangular and known; observations available; weight function S_o.

3.1 When the weight function is W_o, the mode $\mu_G^{(o)}$ of the distribution given by (2.1) yields the optimal strategy for the statistician. Since the denominator of (2.1) does not depend on θ, the mode of (2.1) is that value $\theta^{(o)}$ of θ which will maximize the product $f(x \mid \theta)\, h(\theta)$. Now let $h(\theta)$ be the rectangular distribution, i.e., $h(\theta) =$

$$\begin{cases} \text{const. for } \theta \in \Theta_o \subseteq \Theta & [\Theta_o \text{ is here assumed compact.}] \\ 0 \quad \text{for } \theta \notin \Theta_o \subseteq \Theta. \end{cases}$$

Clearly in that case the mode $\theta^{(o)}$ of $g_H(\theta \mid x)$ is obtained by maximizing $f(x \mid \theta)$ with regard to θ, so that the mode $\theta^{(o)}$ is the maximum likelihood estimate $\hat{\theta}$ of θ in $f(x \mid \theta)$.[4][4a] We have therefore:

(I) Let the weight function be W_o and let the a priori distribution $h(\theta)$ be rectangular over a compact subset Θ_o of the parameter space. Then the maximum likelihood estimate $\hat{\theta}$ of θ in $f(x \mid \theta)$ is the minimum risk (Bayes) estimate.[5]

---

4. If the maximum likelihood estimate is not unique it does not matter which (in this Bayesian approach) is chosen. This is not so in the minimax case. [Cf. [1], pp. 128-9.]

5. As for the converse, consider the case where $H(\theta)$ is absolutely continuous and $f(x|\theta)$ has the necessary differentiability properties with respect to θ, and let θ be multi-dimensional $\theta = (\theta_1, \theta_2, \ldots)$. Then for the location θ' of the maximum of the product $f(x \mid \theta)\, h(\theta)$ to coincide with that of $f(x|\theta)$, i.e., $\hat{\theta}$ is necessary that $\left.\frac{\partial h}{\partial \theta_i}\right|_{\theta_i = \hat{\theta}_i} = 0$, since in that case $\left.\frac{\partial}{\partial \theta_i}[f(x \mid \theta)\, h(\theta)]\right|_{\theta_i = \theta' = \hat{\theta}} =$

$$= \left[f(x \mid \theta) \frac{\partial h}{\partial \theta_i} \right]_{\theta_i = \hat{\theta}_i}.$$

4a. That $\hat{\theta}$ is the mode of g_H was pointed out by R. A. Fisher in his initial use of the maximum likelihood criterion (cf. [5]) and indeed seems to have been the original motivation for proposing the criterion. See below, § 5, for a discussion of the historical aspects of the problem.

When the hypothesis space Θ is not compact it is possible to define a sequence of a priori distributions which produce the effect analogous to the rectangular distribution in the compact case. For instance, let (one-dimensional) Θ be the whole real axis from $-\infty$ to $+\infty$. Define $h_\tau(\theta) = \frac{1}{\sqrt{2\pi}\,\tau} \exp\left\{-\frac{\theta^2}{2\tau^2}\right\}$.

Hence

$$(3.1) \qquad g_{H\tau}(\theta \mid x) = \frac{f(x \mid \theta)\frac{1}{\sqrt{2\pi}\,\tau}\exp\left\{-\frac{\theta^2}{2\tau^2}\right\}}{\int f(x \mid \theta)\frac{1}{\sqrt{2\pi}\,m}\exp\left\{-\frac{\theta^2}{2\tau^2}\right\} d\theta}$$

$$= \frac{f(x \mid \theta)\exp\left\{-\frac{\theta^2}{2\tau^2}\right\}}{\int f(x \mid \theta)\exp\left\{-\frac{\theta^2}{2\tau^2}\right\} d\theta}$$

Maximizing $\log g_{H\tau}(\theta \mid x)$ we have

$$(3.2) \qquad \frac{\partial \log g_{H\tau}(\theta \mid x)}{\partial \theta} = \frac{\partial \log f(x \mid \theta)}{\partial \theta} - \frac{\theta}{\tau^2}$$

Denote by $\hat{\theta}_\tau$ the solution of

$$(3.3) \qquad \frac{\partial \log f(x \mid \theta)}{\partial \theta} - \frac{\theta}{\tau^2} = 0.$$

It is clear that $\lim_{\tau \to \infty} \hat{\theta} = \hat{\theta}$ where $\hat{\theta}$ is the maximum likelihood estimate in $f(x \mid \theta)$.

The foregoing techniques, used in [3] and [4], can be applied to justify (from the Bayesian viewpoint) the maximum likelihood method of estimation of the mean μ of a normal distribution with a known variance σ^2 [to be denoted by $N(\mu, \sigma^2)$ where the underlined parameter is assumed known].

When (one-dimensional) Θ is noncompact but bounded on one side [as is the case in estimating $\sigma^2 \geq 0$ in, say, $N(\mu, \sigma^2)$] we may use as $h_\tau(\theta)$ the χ^2 type distribution with one degree of freedom with the density function $\frac{1}{2\sqrt{2\pi}\,\tau\sqrt{\theta}} \exp\left\{-\frac{\theta}{2\tau^2}\right\}$ and, again, let $\tau \to \infty$.

Analogous methods can be used when Θ is multidimensional. We shall say that $\lim_{\tau \to \infty} \hat{\theta}_\tau$ is optimal for $h_\infty(\theta)$ and regard $h_\infty(\theta)$ as the (analogue of a) rectangular

distribution when Θ is noncompact. With this understanding the statement (I) in § 3.1 above holds whether Θ is or is not compact.

3.3 When the mean, median, and mode of $g_H(\theta \mid x)$ are close together, we may expect that estimates optimal, say, for W_0 would also be fairly good for W_1 or W_2. Furthermore, it may so happen that $g_H(\theta \mid x)$ is almost invariant with regard to H. Under such circumstances we may expect that the maximum likelihood estimates will be close to (minimax) optimal ("asymptotically efficient") with regard to W_2 which is the classical result for "large samples," when $f(x \mid \theta) = \prod_{t=1}^{T} f(x^{(t)} \mid \theta)$ where $x^{(t)}$ denotes the t-th observation and $T \longrightarrow \infty$.

4. Nature's "strategy" unknown; observations available; weight function W_0; and x finite.

4.0 When no a priori distribution for θ is available, the minimax approach often makes it possible to define an optimal strategy for the statistician. When Θ is finite, it is found that there exists a (not necessarily unique) least favorable distribution $\tilde{h} \equiv \tilde{h}(\theta)$, i.e., a distribution $h(\theta)$ that is a minimax strategy for Nature. It is known from the theory of games that the (minimax) optimal strategy $\mathring{S}$ for the statistician must be a Bayes solution for $\tilde{h}$.[6] In general, $\tilde{h}$ depends on the weight function W. Now suppose that for W_0, it so happens that, for a class of $f(x \mid \theta)$, $\tilde{h}(\theta)$ is rectangular and that, for all x, $f(x \mid \theta)$ has a unique maximum with regard to θ. It then follows that the (unique) Bayes solution for $\tilde{h}$ is the maximum likelihood estimate for $f(x \mid \theta)$ and, hence, that the maximum likelihood estimate is (minimax) optimal. This is so since we know from § 3 that for W_0 and h rectangular the maximum likelihood estimate is the Bayes solution and if the Bayes solution corresponding to the least favorable distribution is unique, it is a minimax solution.

The crucial question, therefore, is the following: under what conditions, with W_0, is there a rectangular $\tilde{h}$?

---

6. This implies that if the statistican, through his spies found that Nature did, in fact, use h, he still could not do any better than use $\mathring{S}$.

4.1 A simple case: Θ and X with 2 points each.

4.1.1 Before considering the more general cases it seems useful to examine a simple case where a complete solution is obtained without difficulty. Let $\Theta = (\theta_1, \theta_2)$ where θ_i is a binomial distribution (for one throw) with the probability of x_1 ("heads") equal to p_{i1} (while the probability of x_2 ("tails") equals $p_{i2} = 1 - p_{i1}$). Thus the statistician is asked after one throw of one of the two coins which of the two coins he thinks it was.

Since $p_{i1} \neq p_{i2}$ (otherwise there would be no estimation problem), we may assume without loss of generality that $p_{11} > p_{21}$, hence $p_{22} > p_{12}$. This specification may be presented in the following table:

(4.1-1)

	x_1	x_2
θ_1	(**p_{11}**)	p_{12}
θ_2	p_{21}	(**p_{22}**)

[The encircled element is largest in its column.]

The weight function may be written as

(4.1-2)

Nature \ Stat.	$\bar{\theta}_1 = \theta_1$	$\bar{\theta}_2 = \theta_2$
θ_1	0	1
θ_2	1	0

(It will be noted that since there are only two points in Θ, one cannot distinguish W_0 from W_1 or W_2. Similarly when Θ consists of three points one cannot distinguish W_1 from W_2.)

A "pure" (i.e., nonrandomized) strategy used by the statistician can be thought of as a rule assigning one of the decisions (estimates) to each possible observation. We shall write $S_{i_1 i_2}$ to denote the following rule ("pure" strategy):

"if x_1 is observed, accept the estimate $\bar{\theta}_{i_1} = \theta_{i_1}$; if x_2 is observed, accept the

estimate $\bar{\theta}_{i_2} = \theta_{i_2}$." Clearly, there are only four pure strategies for the statistician: S_{11}, S_{12}, S_{21}, S_{22}. Since $p_{11} > p_{21}$, S_{12} is the maximum likelihood strategy and is sometimes written simply as $\hat{S}$. (S_{21} is the "minimum likelihood" strategy.)

In order to find out which strategies are (respectively) optimal for Nature and for the statistician we must evaluate the risk function, in this case the risk matrix, R. The rows of R correspond to Nature's pure strategies, the columns to those of the statistician. Hence the structure of the matrix R is

N \ S	S_{12}	S_{21}	S_{11}	S_{22}
θ_1				
θ_2				

The entry corresponding to θ_k and $S_{i_1 i_2}$ may be denoted by $r_{k,i_1 i_2} \equiv r(\theta_k, S_{i_1 i_2})$. We have (since W_o is the weight function used)

(4.1-3) $$r_{k,i_1 i_2} \equiv r(\theta_k, S_{i_1 i_2}) = (1 - \delta_{ki_1}) p_{k1} + (1 - \delta_{ki_2}) p_{k2}$$

where $$\delta_{pq} = \begin{cases} 1 \text{ when } p = q \\ 0 \quad " \quad p \neq q. \end{cases}$$

Hence R becomes

(4.1-4)

N \ S	S_{12}	S_{21}	S_{11}	S_{22}
θ_1	p_{12}	p_{11}	0	1
θ_2	p_{21}	p_{22}	1	0

In general, the strategies, both Nature's and the statistician's, are "mixed." A mixed strategy available to Nature may be represented by a vector $\alpha = (\alpha_1, \alpha_2)$, $\alpha_i \geq 0$, $\sum_i \alpha_i = 1$, where $\alpha_i = \text{Prob}(\theta_i)$. When $\alpha_1 = \alpha_2 = 1/2$ we call Nature's strategy <u>rectangular</u> and denote it by $\alpha^{\square} = (1/2, 1/2)$ where the symbol attached to α

is supposed to be reminiscent of a rectangle. A (minimax) optimal strategy for Nature given the risk matrix R will be denoted by $\alpha^{o(R)}$. $O_{\alpha R}$ is the (never empty) set of all $\alpha^{o(R)}$. [Wald's term for $\alpha^{o(R)}$ is a "least favorable" distribution.]

A mixed strategy available to the statistician may be represented by the vector $\tau = (\tau_{12}, \tau_{21}, \tau_{11}, \tau_{22}), \tau_{ij} \geq 0, \sum_{i,j} \tau_{ij} = 1$, where $\tau_{ij} = \text{Prob}(S_{ij})$. Since $S_{12} \equiv \hat{S}$ is the maximum likelihood strategy, we also call the maximum likelihood strategy the vector $\hat{\tau} \equiv (1,0,0,0)$ representing $\hat{S}$. A (minimax) optimal strategy for the statistician given the risk matrix R will be denoted by $\tau^{o(R)}$. $O_{\tau R}$ is the (never empty) set of all $\tau^{o(R)}$. A pair $(\alpha^{o(R)}, \tau^{o(R)})$ is called a solution of R.

Let $\phi_R(\alpha, \tau) \equiv \sum_k \sum_{i,j} \alpha_k r_{k,ij} \tau_{ij}$ be the expected value of the risk for (α, τ). Then it is known from the theory of games that, for any solution,

$$(4.1\text{-}5) \quad \max_{\alpha} \phi_R(\alpha, \tau^{o(R)}) = \min_{\tau} \phi_R(\alpha^{o(R)}, \tau) = \phi_R(\alpha^{o(R)}, \tau^{o(R)}).$$

The common value of the above three expressions is called the value of the game and is denoted by v_R.

We shall use the following result from the theory of games.[7]

(II) If (α, τ) is such that

(II.1) $\phi_R(\alpha, \tau) = \min_{\tau'} \phi_R(\alpha, \tau')$ [i.e., τ is (Bayes) optimal with regard to α],

and

(II.2) $\phi_R(\alpha', \tau) = \phi_R(\alpha'', \tau)$ for all (α', α'') [i.e., τ is a constant risk strategy],

then

(II.3) $\tau \in O_{\tau R}$ [i.e., τ is a (minimax) optimal strategy].

That (II.1,2) imply (II.3) can be seen as follows. (II.1) implies $v_R \equiv \max_{\alpha'} \min_{\tau'} \phi_R(\alpha', \tau') \geq \min_{\tau'} \phi_R(\alpha, \tau') = \phi_R(\alpha, \tau)$. But from (II.2) we get

---

7. Cf. corollary to Theorem 2.1 in [2].

$v_R = \min_{\tau'} \phi_R(\alpha^{o(R)}, \tau') \leq \phi_R(\alpha^{o(R)}, \tau) = \phi_R(\alpha, \tau)$. Hence $\phi(\alpha, \tau) = v_R$ and τ is (minimax) optimal since $\max_{\alpha} \phi_R(\alpha', \tau) = v_R$.

We shall now prove the following result.

(III) Let the risk matrix R be that given in (4.1-0). Then

(III.A):

(III.1) $p_{12} = p_{21}$

implies

(III.2) $\hat{\tau} \in O_{\tau_R}$ [maximum likelihood is (minimax) optimal]

(III.3) $\tau' \neq \tau$ implies $\tau' \notin O_{\tau_R}$ [no other (minimax) optimal strategies exist]

(III.4) $\alpha^0 \in O_{\alpha_R}$ [the rectangular strategy is a least favorable distribution]

and

(III.B):

(III.1') $p_{12} \neq p_{21}$

implies

(III.2') $\hat{\tau} \notin O_{\tau_R}$ [Maximum likelihood is not (minimax) optimal]

(III.4') $\alpha^0 \notin O_{\alpha_R}$ [the rectangular strategy is not a least favorable distribution]

Proof of (III.A).

We first show that if (III.1) holds, then (II.2) holds for $(\alpha^0, \hat{\tau})$. (III.1) is equivalent to (II.2) for $\tau = \hat{\tau}$. That (II.1) is also satisfied follows from (I), but it seems useful to show this here in detail. Clearly

$$\phi_R(\alpha^0, \tau) = 1/2[p_{12}\tau_{12} + p_{11}\tau_{21} + 0\cdot\tau_{11} + 1\cdot\tau_{22}] + 1/2[p_{21}\tau_{12} + p_{22}\tau_{21} + 1\cdot\tau_{11} + 0\cdot\tau_{22}].$$

Since $p_{11} > p_{21}$, $p_{22} > p_{12}$, we have $p_{12} + p_{21} < p_{11} + p_{22}$
and also $p_{12} + p_{21} = p_{12} + (1-p_{22}) = 1-(p_{22} - p_{12}) < 1$.
Hence in $\phi_R(\alpha^0, \tau) = 1/2[(p_{12} + p_{21})\tau_{12} + (p_{11} + p_{22})\tau_{21} + 1\cdot\tau_{11} + 1\cdot\tau_{22}]$

τ_{12} is the only one among the τ_{ij} with a coefficient below 1 and $\phi_R(\alpha^0, \tau)$ reaches its minimum if and only if $\tau_{12} = 1$ while the other components of τ vanish.(8) Therefore $\phi_R(\alpha^0, \hat{\tau}) = \min_{\tau} \phi_R(\alpha^0, \tau)$ and (11.1) holds for $(\alpha^0, \hat{\tau})$ so that, by (11.3), $\hat{\tau} \in O_R$.

It is clear that $\hat{\tau}$ is a unique (minimax) optimal strategy for the statistician. For, clearly, $v_R = p_{12} = p_{21}$ while, as has just been shown, $\phi_R(\alpha^0, \tau') > v_R$ if $\tau' \neq \hat{\tau}$. Hence $\tau' \neq \hat{\tau}$ cannot be (minimax) optimal.

(III.4) follows from the fact $\min_{\tau} \phi(\alpha^0, \tau) = \phi_R(\alpha^0, \hat{\tau}) = v_R$. In fact, it can be shown that $O_{\alpha R}$ consists of all $\alpha^{o(R)^R}$ such that $p_{12} \leq \alpha_1 \leq 1-p_{12}$. [Since the statistician's (minimax) optimal strategy $\hat{\tau}$ is "pure" and unique while α^0 is (minimax) optimal and mixed, it was to be expected from the theorems on the dimensions of game solutions that there would be more than one (minimax) optimal strategy for Nature.]

Proof of (III.B.)

In order to prove (III.1') we shall show that, when (III.1') is satisfied, there is no pair $(\alpha, \hat{\tau})$ for which (4.1-5) holds; since the game must have a solution, it follows that $\hat{\tau}$ is not (minimax) optimal.

To show that no pair $(\alpha, \hat{\tau})$ satisfies (4.1-5) we proceed as follows. Suppose first that $\alpha = \alpha^{(i)}$ where $\alpha^{(i)}$ is a vector whose i-th component is zero. Then clearly $\min_{\tau} \phi_R(\alpha^{(i)}, \tau) = 0$ since each row of R has a zero [viz., $\phi_R(\alpha^{(i)}, \tau_{ii}) = 0$]; on the other hand, $\max_{\alpha} \phi_R(\alpha, \hat{\tau}) = \max(p_{12}, p_{21}) > 0$, hence (4.1-5) does not hold. Hence, if a pair $(\alpha, \hat{\tau})$ exists that satisfies (4.1-5), it must be one with $\alpha \neq \alpha^{(i)}$.(9) Let $(\bar{\alpha}, \hat{\tau})$ be such a pair where $\bar{\alpha} = (\bar{\alpha}_1, \bar{\alpha}_2)$ with $\bar{\alpha}_i \neq 0$, i = 1,2. Then $\phi_R(\bar{\alpha}, \hat{\tau}) = \bar{\alpha}_1 p_{12} + \bar{\alpha}_2 p_{21} < \max(p_{12}, p_{21})$. Hence $\phi_R(\bar{\alpha}, \hat{\tau}) \neq \max_{\alpha} \phi_R(\alpha, \hat{\tau})$ and (4.1-5) is violated. Hence $\hat{\tau}$ is not (minimax) optimal.

Finally, we show that (III.1') implies (III.4'). Suppose $\alpha^0 \in O_{\alpha R}$. It would then

---

8. The proof of this statement does not involve (III.1).
9. Cf [1], theorem 5.4, p. 127.

follow that $v_R = 1/2(p_{12} + p_{21})$, since $\min_{\tau} \phi_R(\alpha^0, \tau) = \phi_R(\alpha^0, \hat{\tau}) = 1/2(p_{12} + p_{21})$. Now it has just been shown that when (III.1') holds $\tau \in O_{\tau_R}$ and that[10] $\phi_R(\alpha^0, \tau) > \phi_R(\alpha^0, \hat{\tau})$. Hence $\phi_R(\alpha^0, \tau^{o(R)}) > 1/2(p_{12} + p_{21})$. But we have necessarily $v_R \equiv \phi_R(\alpha^{o(R)}, \tau^{o(R)}) \equiv \max_{\alpha} \phi_R(\alpha, \tau^{o(R)}) \geq \phi_R(\alpha^0, \tau^{o(R)})$. Hence $v_R > 1/2(p_{12} + p_{21})$. The contradiction can only be removed by discarding the supposition that $\alpha^0 \in O_{\alpha_R}$.

4.1.2 It is intuitively obvious that, when $p_{12} \neq p_{21}$, S_{12} may still be a good strategy if $|p_{12} - p_{21}|$ is not too great. One way of putting this is to say that, probably, τ_{12} has a fairly high value (though < 1) in $\tau^{o(R)}$.

A more useful measure, however, can be given by defining for a given strategy τ the ratio

$$(4.1\text{-}6) \qquad e_{\tau} = \frac{v_R}{\max_{\alpha} \phi_R(\alpha, \tau)} .$$

Clearly, $e_{\tau} \leq 1$ and $e_{\tau^{o(R)}} = 1$. It seems appropriate to call e_{τ} the (relative) risk-efficiency of τ.[11]

In the example we have been considering we have $e_{\hat{\tau}} = \dfrac{v_R}{\max(p_{12}, p_{21})}$.

When $p_{12} = p_{21} = v_R$, $e_{\hat{\tau}} = 1$. [This was bound to be so since in the case $p_{12} = p_{21}$ $\hat{\tau}$ is (minimax) optimal.] When $p_{12} \neq p_{21}$, we have[12] $\mathrm{av}(p_{12}, p_{21}) \leq v_R$ where $\mathrm{av}(p_{12}, p_{21}) = 1/2(p_{12} + p_{21})$. [The inequality follows from the fact that, even for $p_{12} \neq p_{21}$, $\min_{\tau} \phi_R(\alpha^0, \tau) = \phi_R(\alpha^0, \hat{\tau}) = \mathrm{av}(p_{12}, p_{21})$ and we must have $v_R \equiv \max_{\alpha} \min_{\tau} \phi(\alpha, \tau) \geq \min \phi(\alpha^0, \tau)$.] Hence

$$(4.3) \qquad e_{\hat{\tau}} \geq \frac{\mathrm{av}(p_{12}, p_{21})}{\max(p_{12}, p_{21})} .$$

---

10. Independently of (III.1), cf. the proof of (III.A).

11. The conventional (variance ratio) efficiency is a special case of e_{τ} when W_2 is used.

12. For an upper bound on v_R see footnote 13 in § 4.1.3.

When this ratio is close to 1, maximum likelihood estimates are bound to be close to the (minimax) optimal estimate.

4.1.3 Example.

Let $p_{12} = .8$, $p_{21} = .1$

Hence the $f(x|\theta)$ table is

θ \ x	x_1	x_2
θ_1	.2	.8
θ_2	.1	.9

and the risk matrix R is

N \ S	S_{12}	S_{21}	S_{11}	S_{22}
θ_1	.8	.2	0	1
θ_2	.1	.9	1	0

It can easily be shown that the unique minimax solution in this case is given by $\alpha_1 = \frac{9}{17}$, $\alpha_2 = \frac{8}{17}$, $\tau_{12} = \frac{10}{17}$, $\tau_{21} = 0$, $\tau_{11} = \frac{7}{17}$, $\tau_{22} = 0$. The value of the game $v_R = \frac{8}{17} = .470$. The (relative) risk-efficiency of the maximum likelihood estimate is $e_{\hat{\tau}} = \frac{\left(\frac{8}{17}\right)}{.8} = \frac{10}{17} = .588$. If we had wanted to put bounds on $e_{\hat{\tau}}$ without computing the exact value of v_R we could have noted that $v \leq .5$, since this could always be accomplished by letting $\tau_{ii} = 1/2 (i=1,2)$[13], and also $v_R \geq .45 = \min(p_{12}, p_{21})$. Hence the bounds obtained are $\frac{.45}{.8} \leq e_{\hat{\tau}} \leq \frac{.5}{.8}$, i.e., $.5625 \leq e_{\hat{\tau}} \leq .625$.

4.2 Generalization of 4.1.

4.2.1. (Θ and X with K points each). Let $\Theta = (\theta_1, \ldots, \theta_K)$ and let the sample space be $X = (x_1, \ldots, x_K)$. [The case treated in § 4.1 was that of $K = 2$.]

We write

$$(4.2\text{-}1) \qquad p_{kj} = \text{Prob}(x = x_j \mid \theta = \theta_k)$$

---

13. In general, when R is given by (4.1-4), $v_R \leq \min[\max(p_{12}, p_{21}), 1/2]$.

so that the $f(x \mid \theta)$ table is

(4.2-2)

θ \ x	x_1	x_2	...	x_K
θ_1	$\textcircled{p_{11}}$	p_{12}	...	p_{1K}
θ_2	p_{21}	$\textcircled{p_{22}}$	...	p_{2K}
...	...	...	...	...
θ_K	p_{K1}	p_{K2}	...	$\textcircled{p_{KK}}$

where $p_{jj} > p_{kj}$ for all $k \neq j$. The maximal element in each column is encircled. [It will be noted that we have restricted ourselves, for the sake of simplicity, to the case where there are unique maximum likelihood solutions for each x_j.[14]]

Denote by $S_{i_1 i_2,\ldots,i_K}$ the (pure) strategy consisting in choosing θ_{i_k} as estimate whenever x_k is observed. The risk $r(\theta_k, S_{i_1 i_2,\ldots,i_K})$ associated with using $S_{i_1 i_2,\ldots,i_K}$ when θ_k is true is given by the following expression [$\epsilon_{kj} = 0$ if $k = j$, $\epsilon_{kj} = 1$ if $k \neq j$]:

$$(4.2\text{-}3) \qquad r(\theta_k, S_{i_1 i_2,\ldots,i_K}) = \sum_{j=1}^{K} \epsilon_{ki_j} p_{kj}$$

where the j-th term is the probability p_{kj} of observing x_j when θ_k is true multiplied by 0 or 1 depending on whether i_j does or does not equal k.

The average risk corresponding to the use by Nature of a rectangular strategy $\alpha^0 = (\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K})$ when $S_{i_1 i_2,\ldots,i_K}$ is used by the statistician is

$$(4.2\text{-}4) \quad r(\alpha^0, S_{i_1 i_2,\ldots,i_K}) = \frac{1}{K} \sum_{k=1}^{K} r(\theta_k, S_{i_1 i_2,\ldots,i_k}) = \frac{1}{K} \sum_{k=1}^{K} \sum_{j=1}^{K} \epsilon_{ki_j} p_{kj}$$

$$= \frac{1}{K} \sum_{j=1}^{K} \sum_{k=1}^{K} \epsilon_{ki_j} p_{kj} = \frac{1}{K} \sum_{j=1}^{K} r_j$$

where

$$(4.2\text{-}5) \qquad r_j = \sum_{k=1}^{K} \epsilon_{ki_j} p_{kj} .$$

---

14. Wald's example in [1], pp. 128-9, violates this condition.

Of the K entities $\epsilon_{1i_j}, \epsilon_{2i_j}, \ldots, \epsilon_{Ki_j}$ one and only one will have equal subscripts and vanish. Hence r_j is minimized when $k = i_j$ for the largest among $p_{1j}, p_{2j}, \ldots, p_{Kj}$, i.e., according to our notation r_j is minimized when $i_j = j$, i.e., for the pure strategy $S_{12 \ldots K}$ which is the maximum likelihood strategy $\hat{S}$.[15]

Let $\hat{\tau} = (1,0,0,\ldots,0)$ where the first component corresponds to $\hat{S}$. Then we may formulate the result (IV) below of which (III) is the special case for $K = 2$. An examination of the proof of (III) will show that it goes through for any K, since (II) holds for R of any dimensions and we have just shown that $\hat{\tau}$ is (Bayes) optimal with regard to α'' for any K. Hence both (III.A) and (III.B) may be reformulated as (IV.A) and (IV.B) below respectively with (III.1) being replaced by

(4.2-6) $r(\theta_{k'}, \hat{S}) = r(\theta_{k''}, \hat{S})$ for all (k', k'')

or, more explicitly

(4.2-6') $\sum_{j \neq 1} p_{1j} = \sum_{j \neq 2} p_{2j} = \ldots = \sum_{j \neq K} p_{Kj}$

and with (III.1') being replaced by the negation of (4.2-6).

We therefore have

(IV) Let the risk matrix R be that given by (4.2-3). Then

(IVA):

(IV.1) $\sum_{j \neq 1} p_{1j} = \sum_{j \neq 2} p_{2j} = \ldots = \sum_{j \neq K} p_{Kj}$

implies

(IV.2) $\hat{\tau} \in O_{\tau R}$ [maximum likelihood is (minimax) optimal]

(IV.3) $\tau' \neq \hat{\tau}$ implies $\tau' \notin O_{\tau R}$ [no other (minimax) optimal strategies exist]

(IV.4) $\alpha'' \in O_{\alpha R}$ [the rectangular strategy is a least favorable distribution]

and

(IV.B): if

(IV.1') $\sum_{j \neq 1} p_{1j} = \sum_{j \neq 2} p_{2j} = \ldots = \sum_{j \neq K} p_{Kj}$ does not hold,

this implies

---

15. The foregoing is essentially a repetition of the proof of (I) in §3 using methods which generalize those used in §4.1. It is given here because of its direct relationship to the risk matrix.

(IV.2') $\hat{\tau} \notin O_{\tau_R}$ [maximum likelihood is not (minimax) optimal]

(IV.4') $\alpha^0 \notin O_{\alpha_R}$ [the rectangular strategy is not a least favorable distribution]

4.2.2. As in § 4.1 we define the (relative) risk-efficiency of a strategy (represented by) τ as

$$(4.2\text{-}7) \qquad e_\tau = \frac{\max_\alpha \min_\tau \phi(\alpha, \tau')}{\max_\alpha \phi(\alpha, \tau)} .$$

We have

$$(4.2\text{-}8) \qquad \frac{\frac{1}{K}\sum_{k=1}^{K} r(\theta_k, \hat{\tau})}{\max_k r(\theta_k, \hat{\tau})} \leq e_{\hat{\tau}} \leq \frac{\min[\max_k r(\theta_k, \hat{\tau}), \frac{K-1}{K}]}{\max r(\theta_k, \hat{\tau})} .$$

4.3 Generalization of 4.2.

In this section we consider the case which is a generalization of that treated in 4.2. Let $\Theta = (\theta_1, \ldots, \theta_K)$ as before while

$X = (x_{11}, \ldots, x_{1m_1}; x_{21}, \ldots, x_{2m_2}; \ldots; x_{K1}, \ldots, x_{Km_K})$ where $m_k \geq 1$ $(k = 1, 2, \ldots, K)$.

[The case in § 4.2 is obtained when $m_k = 1$ for all k.]

4.3.1. The $f(x \mid \theta)$ table is

(4.3-1)

	x_{11}	x_{12}	...	x_{1m_1}	x_{21}	x_{22}	...	x_{2m_2}	x_{31}	...
θ_1	$p_{1,11}$	$p_{1,12}$	...	$p_{1,1m_1}$	$p_{1,21}$	$p_{1,22}$	...	$p_{1,2m_2}$	$p_{1,31}$	...
θ_2	$p_{2,11}$	$p_{2,12}$	...	$p_{2,1m_1}$	$p_{2,21}$	$p_{2,22}$	...	$p_{2,2m_2}$	$p_{2,31}$	...
:	:	:	:	:	:	:	:	:	:	:

where it is again assumed that each column has a unique maximum, indicated by the circle.

I.e., we have

$$(4.3\text{-}2) \qquad p_{k,js_j} = \text{Prob}(x = x_{js_j} \mid \theta = \theta_k)$$

and

$$(4.3\text{-}3) \qquad p_{j,js_j} > p_{k,js_j} \quad \text{for all } k \neq j.$$

We denote by $S_{i_{11}i_{12},\ldots,i_{21},\ldots,i_{Km_K}}$ the strategy consisting in choosing the estimate $\theta_{i_{pq}}$ whenever x_{pq} is observed.

The risk matrix R is given by

$$(4.3\text{-}4)\quad r(\theta_k, S_{i_{11}\ldots i_{Km_K}}) = \sum_{j=1}^{K} \sum_{s_j=1}^{m_j} \epsilon_{ki_{js_j}} p_{k,js_j}$$

where, as before, $\epsilon_{ki_{js_j}} = 0$ or 1 depending on whether $k = i_{js_j}$ or not.

Hence

$$(4.3\text{-}4.2)\quad \sum_k r(\theta_k, S_{i_{11}\ldots i_{Km_K}}) = \sum_{j=1}^{K} \sum_{s_j=1}^{m_j} r_{js_j}$$

where

$$(4.3\text{-}4.3)\quad r_{js_j} = \sum_{k=1}^{K} \epsilon_{ki_{js_j}} p_{k,js_j}$$

so that r_{js_j} is minimized for

$$(4.3\text{-}4.4)\quad i_{js_j} = j,$$

i.e., for $S_{1\ldots1;2\ldots2;\ldots;K\ldots K} = \hat{S}$ (maximum likelihood).

It can therefore be shown again that (IV) holds with only the constant risk condition appropriately modified, i.e.,

$$(4.3\text{-}5)\quad r(\theta_{k'}, \hat{S}) = r(\theta_{k''}, \hat{S}) \qquad \text{for all } (k', k'')$$

where $S = S_{11\ldots1;2\ldots2;\ldots;K\ldots K}$ (i.e., $i_{js_j} = j$) or, more explicitly,

$$(4.3\text{-}5')\quad \sum_{j\neq1} \sum_{s_j} p_{1,js_j} = \sum_{j\neq2} \sum_{s_j} p_{2,js_j} = \ldots = \sum_{j\neq K} \sum_{s_j} p_{K,js_j}.$$

Hence we have (V) below [of which (IV) is a special case for all $m_k = 1$].

(V) Let the risk matrix R be that given by (4.3-4). Then

(V.A):

$$(V.1)\quad \sum_{j\neq1} \sum_{s_j} p_{1,js_j} = \sum_{j\neq2} \sum_{s_j} p_{2,js_j} = \ldots = \sum_{j\neq K} \sum_{s_j} p_{K,js_j}$$

implies

(V.2) $\hat{\tau} \in O_{\tau_R}$ [maximum likelihood is (minimax) optimal]

(V.3) $\tau' \neq \hat{\tau}$ implies $\tau' \notin O_{\tau_R}$ [no other (minimax) optimal strategies exist]

(V.4) $\alpha'' \in O_{\alpha_R}$ [the rectangular strategy is a least favorable distribution]

and

(IV.B): if

(V.1') $\sum_{j \neq 1} \sum_{s_j} p_{1,js_j} = \sum_{j \neq 2} \sum_{s_j} p_{2,js_j} = \cdots = \sum_{j \neq K} \sum_{s_j} p_{K,js_j}$ does not hold,

this implies

(V.2') $\pi \in O_{\tau_R}$ [maximum likelihood is <u>not</u> (minimax) optimal]

(V.4') $\alpha'' \notin O_{\alpha_R}$ [the rectangular strategy is <u>not</u> a least favorable distribution].

4.4 <u>The Continuous Case Corresponding to the Discrete Case Treated in § 4.2.</u>

We shall now consider the case where the domains of both x and θ are continuous. Let

(4.4-1) $$\max_{\theta \in \Theta} f(x_{\theta_o} \mid \theta) = f(x_{\theta_o} \mid \theta_o)$$

and assume that this defines a unique element $x_{\theta_o} \in X$. It is also assumed that there is a unique $\theta_x \in \Theta$ such that

(4.4-2) $$\max_{\theta \in \Theta} f(x \mid \theta) = f(x \mid \theta_x).$$

Hence there is a one-to-one correspondence between the elements of Θ and of X.

Now if the maximum likelihood method of estimation is used, the risk for a given value θ of the unknown parameter will vary (inversely) with the density $f(x_\theta \mid \theta)$. Hence the <u>constant risk requirement</u> becomes[16]

(4.4-3) $f(x_{\theta'} \mid \theta') = f(x_{\theta''} \mid \theta'')$ for all $\theta' \in \Theta$, $\theta'' \in \Theta$.

The following argument[17] shows that (4.4-3) implies the <u>constancy of the density at the mode</u> for all θ. Let the mode of $f(x \mid \theta')$ be at x' while $x_{\theta''} = x''$. Also, let

---

16. This corresponds to (IV.1) above and follows from Theorem 2.1 in [2].

17. Due to Morton Slater.

$\theta_{x'} = \theta''$. Then $f(x'|\theta') \geq f(x''|\theta)$ and $f(x'|\theta'') > f(x'|\theta')$ of $\theta' \neq \theta''$. Hence $f(x'|\theta'') > f(x''|\theta')$ if $\theta' \neq \theta''$. But from the constant risk condition (4.4-3) it follows that $f(x'|\theta'') = f(x''|\theta')$, hence $\theta' = \theta''$, so that $f(x'|\theta')$ is the density for the mode x' of $f(x|\theta')$ and also $f(x'|\theta')$ is the maximal value (with regard to θ) of $f(x'|\theta)$. Hence, denoting by x^M_θ the mode of $f(x|\theta)$, we have

$$(4.4\text{-}4) \qquad f(x^M_{\theta'}|\theta') = f(x^M_{\theta''}|\theta'') \qquad \text{for all } \theta' \in \Theta, \ \theta'' \in \Theta .$$

As an example, we shall now consider the problem of estimating μ in $N(\mu, \sigma^2)$ [i.e., where the variance is known] in a sample of one observation. [It can be shown that mathematically the situation is not changed by increasing the size of the sample since the sample mean $\bar{X}$ is a sufficient statistic.]

By using the device described in §3, we find that the observation x is the (Bayes) optimal estimate for W_0 with regard to $h_\infty(\mu)$.[18] Now it is easily seen that x is a constant risk estimate since in this case

$$(4.4\text{-}5) \qquad f(x_\mu|\mu) = f(\mu|\mu) = \frac{1}{\sqrt{2\pi}\,\sigma}$$

which is independent of the true value of μ.[19] Hence in this case the maximum likelihood method is the (minimax) optimal estimate for W_0.

This is of importance, since a similar argument seems to go through for the normal regression theory, hence also for the (non-auto-regressive) "just identified" simultaneous equation systems arising in econometric problems.

5. Historical Notes. The following remarks are devoted to some aspects of the history of the problem. Since the writer has made only a cursory inspection of a few selected

---

(18) It would seem that a sequence of rectangular distributions with zero mean and a range 2τ, $\tau \to \infty$, could have been used instead of the Gaussian sequence $h_\tau(\mu)$.

(19) This is not the case when σ is being estimated in $N(\mu, \sigma^2)$, say $N(0, \sigma^2)$. Here $f(x_\sigma|\sigma) = f(\sigma|\sigma) = (1/\sqrt{2\pi}\sigma)\exp(-1/2)$ which is not independent of σ. Hence the constant risk requirement is not satisfied.

references, completeness of treatment cannot be expected.

5.1 Initially, R.A.Fisher introduced the maximum likelihood method [5] on the basis of (Bayesian) "inverse probability" theory, i.e., of the general a posteriori formula $g(\theta|x) = f(x|\theta)h(\theta)$ combined with the "Bayes postulate" (= "equal ignorance" argument, terminology used in [7]) that $h(\theta)$ is rectangular.[20]

In subsequent papers (see references [6]) Fisher has taken pains to disclaim the existence of any relationship between the Bayesian approach and the method of maximum likelihood. It seems worthwhile to quote from [6b](pp.[22.531-2]):

> " If we make the assumption that $\psi(\theta_1,\theta_2,\theta_3,\dots)$ = constant, and if then we ignore everything about the inverse probability distribution so obtained except its mode or point at which the ordinate is greatest, we have to maximise
>
> $$\prod_{r=1}^{n}\{\phi(x_r,\theta_1,\theta_2,\theta_3,\dots)\}$$
>
> for variations of $\theta_1,\theta_2,\theta_3,\dots$; and the result of this process will be the same whether we use the parameters $\theta_1,\theta_2,\theta_3,\dots$ or any functions of them, $\theta_1',\theta_2',\theta_3',\dots$. Two wholly arbitrary elements in this process have in fact cancelled each other out, the non-invariant process of taking the mode, and the arbitrary assumption that ψ is constant. The choice of the mode is thinly disguised as that of "the most probable value," whereas had the inverse probability distribution any objective reality at all we should certainly, at least for a single parameter, have preferred to take the mean or the median value. In fact neither of these two processes has a logical justification, but each is necessary to eliminate errors introduced by the other.
>
> " The process of maximising $\prod(\phi)$ or $S(\log\phi)$ is a method of estimation known as the "method of maximum likelihood"; it has in fact no logical connection with inverse probability at all."

The following comments pertain to the two "processes" questioned by Fisher in the above quotation.

---

20. This is more clearly stated, though at the same time repudiated, in subsequent papers (see reference [6]).

(a) The use of the mode.

In the light of the present paper it seems that the use of the mode can be justified in terms of the weight function W_o.

Admittedly, W_o may be an unreasonable weight function to use in many cases, especially for continuous θ. This, however, probably is an argument against using the maximum likelihood method (except where [as in large samples, or for $N(\mu, \sigma^2)$] it coincides with estimates optimal with regard to other weight functions). It is not an argument against the Bayesian interpretation of the maximum likelihood criterion.

(b) Rectangular a priori distribution.

As for the rectangular a priori distribution, Fisher and others are undoubtedly right that the "Bayes postulate" ("equal ignorance" argument in favor of the rectangularity assumption) lacks foundation and breaks down in the continuous case.

As shown in the present paper, however, it is possible to introduce the a priori rectangular distribution in the parameter space as a "least favorable" strategy for nature and this has nothing to do with the "Bayes postulate" approach. It is true, of course, that the case of a rectangular least favorable strategy is a special one.[21] But, again, this seems to be an argument against the use of the maximum likelihood method (at least in finite samples) rather than against its minimax interpretation.

A case of particular interest is that of the estimation of the mean of a normal distribution (see 4.2.3 above). It is seen that the maximum likelihood estimate is (minimax) optimal.[22]

5.2 The idea of the minimax justification of the maximum likelihood approach and of

---

21. However, even when the constant risk requirement is not satisfied and hence no least favorable distribution is rectangular, the maximum likelihood method may be quite efficient [in the sense defined in eq. (4.1-6)] and may still be the (minimax) best among "pure" strategies.

22. In proving this, a large class of $h_\tau(\theta), \tau \to \infty$, can be used.

a rectangular least favorable strategy is suggested by the example given in [1], pp. 128-9, although the example itself[23] leads to a mixed minimax strategy.[24]

5.3 In an early paper ([9] Theorem 6, p. 321) Wald gave conditions, different from those of the present paper, for the maximum likelihood case to be (minimax) optimal when Θ is one-dimensional and applied these conditions (ibid. p. 322) to the case of estimating μ in $N(\mu,1)$.

---

23. A sample of two observations from a binomial universe with the hypothesis space given by $p = \frac{1}{3}, \frac{2}{3}$.

24. This is due to the fact that for one of the points in the sample space there is no unique maximum likelihood estimate.

[1] A Wald, Statistical Decision Functions, John Wiley and Sons, New York, 1950.

[2] J. L. Hodges, Jr. and E. L. Lehman, "Some Problems in Minimax Point Estimation," Annals of Mathematical Statistics, Vol. 21, No. 2, June 1950, p. 182.

[3] J. Wolfowitz, "Minimax Estimates of the Mean of a Normal Distribution with Known Variance," Annals of Mathematical Statistics, Vol. 21, No. 2, June 1950, p. 218 ff.

[4] Charles Stein and Abraham Wald, "Sequential Confidence Intervals for the Mean of a Normal Distribution with Known Variance," Annals of Mathematical Statistics, Vol. XVIII, No. 3, September 1947, p. 427.

[5] R. A. Fisher, "On an Absolute Criterion for Fitting Frequency Curves," p. 155-60, The Messenger of Mathematics, Vol. XLI, 1912, Cambridge.

[6] R. A. Fisher, Contributions to Mathematical Statistics, New York 1950

[6b] Paper 22, "Inverse Probability," esp. pp. [22.531-2], (1930)

[6a] Paper 10, "On the Mathematical Foundations of Statistics," §§6, pp. [10.323-7], (1922).

[6c] Paper 24, "Two New Properties of Maximum Likelihood," esp. pp.[24.285-7], (1934).

[6d] Paper 27, "Uncertain Inference," esp. pp. [27.246-248,249], (1936).

[7] Kendall, M. G., The Advanced Theory of Statistics, Volume I, London 1945, Chapter 7, "Probability and Likelihood," §§7.24-7.30(pp.175-180) and bibliographical references, pp. 183-4.

[8] Kendall, M. G., "On the Method of Maximum Likelihood," Journal of the Royal Statistical Society, Volume 103, (1940) Part III, pp. 388-399.

[9] A. Wald, "Contributions to the Theory of Statistical Estimation and Testing Hypotheses," Annals of Mathematical Statistics, Volume X, No. 4, December 1939.