MAXIMUM LIKELIHOOD ESTIMATES OF MONOTONE PARAMETERS[1]

BY H. D. BRUNK

University of Missouri

1. Summary. The maximum likelihood estimators of distribution parameters subject to certain order relations are determined for simultaneous sampling from a number of populations, when (*i*) the order relations may be specified by regarding the distribution parameters, of which one is associated with each population, as values at specified points of a function of n variables (n a positive integer), monotone in each variable separately; (ii) the distributions of the populations from which sample values are taken belong to the exponential family defined below. This family includes, in particular, the binomial, the normal with fixed standard deviation and variable mean, the normal with fixed mean and variable standard deviation, and the Poisson distributions.

2. Monotone parameters, and the exponential family of distributions. Let m, n be positive integers. Let t denote an n-tuple, $t = (t^1, t^2, \cdots, t^n)$, of real numbers, and let $t_k (k = 1, 2, \cdots, m)$ denote one member of a set E consisting of m such n-tuples. Let there be m populations from which sample values are to be taken, and let the distribution of the k-th be completely specified by the knowledge of a single distribution parameter $\theta_k (k = 1, 2, \cdots, m)$. Let the parameters θ_k be known to satisfy the following monotonicity condition: there is a real-valued function $\theta(t)$, monotone non-decreasing in each of the separate variables $t^i (i = 1, 2, \cdots, n)$, such that $\theta_k = \theta(t_k)$, $k = 1, 2, \cdots, m$. (If $\theta(t)$ should be monotone non-increasing in some or all of the variables t^i, a change in direction of the corresponding axis in the n-space $\mathcal{R}_n$ would serve to render $\theta(t)$ non-decreasing in each variable.)

Let each population be a population of values of a stochastic variable, $\mathbf{x}$, such that there exist a finite or infinite interval I, a number $c \varepsilon I$, a probability measure ν on the Borel sets of the real line, a Baire function $g(x)$, and a strictly decreasing, continuous function $\tau(v)$ for $v \varepsilon I$ such that the density function of $\mathbf{x}$ with respect to ν is given by $\exp\{-F[g(x), \theta]\}$ for some $\theta \varepsilon I$, where

$$(1) \qquad F(u, v) = \int_c^v (u - z)\, d\tau(z) = \gamma(v) + u[\tau(v) - \tau(c)],$$

and where

$$\gamma(v) = -\int_c^v z\, d\tau(z).$$

Received October 25, 1954.

[1] This research was supported by the United States Air Force through the Office of Scientific Research of the Air Research and Development Command.

Then if A is a Borel set on the real line,

$$\Pr\{\mathbf{x} \in A\} = \int_A \exp\{-F[g(x), \theta]\}\, d\nu.$$

Let the distributions of the stochastic variables associated with the m populations be obtained by holding I, c, ν, g, and τ fixed, and letting θ take on the values $\theta_1, \theta_2, \cdots, \theta_m$.

As particular instances of such distributions, one obtains the normal distribution with mean θ, variance 1, on setting $I = (-\infty, \infty)$, $c = 0$, $g(x) \equiv x$, $\tau(v) \equiv -v$, $d\nu \equiv \exp(-x^2/2)/\sqrt{2\pi}\, dx$; the binomial with parameter θ on setting $I = [0, 1]$, $c = \frac{1}{2}$, $g(x) \equiv x$, $\tau(v) \equiv \log[(1 - v)/v]$, ν the measure assigning measure $\frac{1}{2}$ to each of the two points 0 and 1; the normal with mean 0, variance θ on setting $I = (0, \infty)$, $c = 1$, $g(x) \equiv x^2$, $\tau(v) \equiv (1 - v)/2v$, $d\nu \equiv \exp(-x^2/2)/\sqrt{2\pi}\, dx$; and the Poisson with mean θ on setting $I = [0, \infty)$, $c = 1$, $g(x) \equiv x$, $\tau(v) \equiv -\log v$, ν the measure assigning measure $e^{-1}/x!$ to each non-negative integer x.

The distributions described above belong to the exponential family ([5], p. 65). It is known that in sampling from a single population whose distribution depends on a single parameter, under conditions implying sufficient regularity, only distributions having density functions of the form

$$\exp\{-[\gamma(\theta) + g(x)\tau(\theta)]\}$$

with respect to a measure admit sufficient estimators ([6], [7]) and efficient estimators (as defined in [4], pp. 479–481), and that the maximum likelihood estimator is sufficient and efficient. It is further clear that, given sufficient regularity, if the density function is of the form $\exp\{-[\chi(\rho) + g(x)\psi(\rho)]\}$ then the change of parameter $\theta = -\chi'(\rho)/\psi'(\rho)$ yields a density function of the form

$$\exp\{-[\gamma(\theta) + g(x)\tau(\theta)]\} = \exp\{-F[g(x), \theta]\},$$

where $\gamma'(\theta) = -\theta\tau'(\theta)$, hence

$$F[g(x), \theta)] = \int_c^\theta [g(x) - z]\, d\tau(z)$$

(except perhaps for an additive function of x which may be absorbed in the measure ν).

The population whose distribution is determined by the distribution parameter $\theta_k = \theta(t_k)$ will be referred to as the population *at* $t_k (k = 1, 2, \cdots, m)$. Let there be drawn from the population at t_k, $N(k)$ sample values, $x_{1k}, x_{2k}, \cdots, x_{N(k),k}$. The sample stochastic variable $\mathbf{x}_{jk}$ has the density function $\exp\{-F[g(x), \theta_k]\}$ with respect to the measure ν, and for distinct j the stochastic variables $\mathbf{x}_{jk}$ are independent $(j = 1, 2, \cdots, N(k); k = 1, 2, \cdots, m)$.

The problem is to determine the maximum likelihood estimators of the parameters $\theta_1, \cdots, \theta_m$, subject to the monotonicity requirement.

If A is a Borel set in the sample space of points $(x_{1k}, x_{2k}, \cdots, x_{N(k),k})$, then

$$\Pr\{(\mathbf{x}_{1k}, \cdots, \mathbf{x}_{N(k),k}) \varepsilon A\} = \int_A \exp\left\{-\sum_{j=1}^{N(k)} F[g(x_{jk}), \theta_k]\right\} d(\nu \times \cdots \times \nu),$$

where $\nu \times \cdots \times \nu$ denotes the direct product of the $N(k)$ measures ν. From (1) it then follows that

$$\Pr\{(\mathbf{x}_{1k}, \cdots, \mathbf{x}_{N(k),k}) \varepsilon A\}$$
$$= \int_A \exp\left\{-N(k)F\left[\sum_{j=1}^{N(k)} g(x_{jk})/N(k), \theta_k\right]\right\} d(\nu \times \cdots \times \nu).$$

Similarly, if A is a Borel set in the sample space of points

$$(x_{11}, \cdots, x_{N(1),1}; \cdots; x_{1m}, \cdots, x_{N(m),m}),$$

then

$$\Pr\{(\mathbf{x}_{11}, \cdots, \mathbf{x}_{N(1),1}; \cdots; \mathbf{x}_{1m}, \cdots, \mathbf{x}_{N(m),m}) \varepsilon A\}$$
$$= \int_A \exp\left\{-\sum_{k=1}^{m} N(k)F\left[\sum_{j=1}^{N(k)} g(x_{jk})/N(k), \theta_k\right]\right\} d(\nu \times \cdots \times \nu),$$

where the integration is with respect to the direct product of the $\sum_{k=1}^{m} N(k)$ measures ν.

The numbers x_{jk}, as well as the functions g, F and the numbers $N(k)$, are to be regarded as given; the maximum likelihood estimator of the parameters $\theta_k = \theta(t_k)$, subject to the monotonicity requirement that $\theta(t)$ be non-decreasing in each t^i $(i = 1, 2, \cdots, n)$, is the set of m numbers for which the exponential in the integrand assumes its maximum value in the class of all sets $(\theta_1, \cdots, \theta_m)$ satisfying the monotonicity requirement. The same set will also minimize the negative logarithm of the exponential:

$$\sum_{k=1}^{m} N(k)F[\bar{y}_k, \theta_k], \tag{2}$$

where $\bar{y}_k = \sum_{j=1}^{N(k)} g(x_{jk})/N(k)$ $(k = 1, 2, \cdots, m)$. This sum can be rewritten in the form

$$\int F[\bar{y}(t), \theta(t)]\, d\mu, \tag{3}$$

where the measure μ assigns the measure $N(k)$ to the point $t_k (k = 1, 2, \cdots, m)$ and where $\bar{y}(t_k) = \bar{y}_k = \sum_{j=1}^{N(k)} g(x_{jk})/N(k)$. The problem of determining a function $\Theta(t)$ minimizing, for given $\bar{y}(t)$, the integral (3) in the class of functions $\theta(t)$ monotone non-decreasing in each $t^i (i = 1, 2, \cdots, n)$ is a special instance of the problem discussed in [3]. The problem discussed in [1] is again a special instance of the problem of the present paper, obtained on setting $n = 1$ and supposing that each population has a binomial distribution.

The results of the present paper as regards existence, representations, and uniqueness of the maximum likelihood estimators are contained in corresponding results obtained by Ewing, Reid, Utz, and Brunk [2], [3]. The purpose of the present paper is to point out the application of those results to the sampling situation described above. Further, the fact that in the problem of this paper the measure μ has all its mass concentrated in a finite number of points makes it possible to simplify very considerably, for this situation, the proofs of the results mentioned above. In the interest of completeness these simplified proofs are sketched in Section 4.

3. Example: a problem in perception. The following example, in which $n = 2$ and the populations have binomial distributions, has been suggested by G. M. Ewing. Given a population of people, to be termed observers, let $\theta(t) = \theta(t^1, t^2)$ denote the probability that an observer O will fail to see a certain object which is idealized into a line segment, when the distance OM from O to the center, M, of the segment is t^1, and the orientation of the segment is specified by the angle t^2, $0 \leqq t^2 \leqq \pi/2$, from the perpendicular to OM to the line segment. Imagine a series of experiments in which t^1 and t^2 are controlled and in which corresponding to each pair $t_k = (t_k^1, t_k^2)$, $k = 1, 2, \cdots, m$, interpreted as a point in the t^1, t^2 plane, $N(k)$ observers independently report whether they do or do not see the object. Let $\bar{y}(t_k)$ denote the fraction of the total number, $N(k)$, of observers who record not seeing the segment at distance t_k^1 with orientation t_k^2. Let $\mathbf{x}_{jk}(j = 1, 2, \cdots, N(k); k = 1, 2, \cdots, m)$ denote the stochastic variable which assumes the value 1 if the j-th observer at t_k fails to see the object, otherwise the value 0; then

$$\bar{y}(t_k) = \left[\sum_{j=1}^{N(k)} x_{jk}\right] / N(k).$$

It seems reasonable to suppose that $\theta(t)$ is monotone non-decreasing in t^1 and t^2 separately.

If $\mathbf{x}$ denotes the number of successes in a single trial of an event with probability θ, then its probability function is $\theta^x(1 - \theta)^{1-x}(x = 0, 1)$. In the present notation, $\Pr\{\mathbf{x} \in A\} = \int_A (2\theta)^x[2(1 - \theta)]^{1-x} \, d\nu$, where ν attributes measure $\frac{1}{2}$ to each of the points 0 and 1. The distribution of $\mathbf{x}$ is the binomial distribution.

4. Existence, representations, and uniqueness of maximum likelihood estimators. For points t, t^* in $\mathcal{R}_n$, write $t \prec t^*$ if no coordinate of t is greater than the corresponding coordinate of t^*, and $t \prec\prec t^*$ if each coordinate of t is less than the corresponding coordinate of t^*. The monotonicity condition on the parameters $\theta_1, \cdots, \theta_m$ can then be described as requiring that there exist a real-valued function $\theta(t)$ defined for $t \in \mathcal{R}_n$ such that $\theta_k = \theta(t_k)$ and such that $\theta(t) \leqq \theta(t^*)$ whenever $t \prec t^*$.

The set E of points t_k corresponding to populations from which sample values are taken is a finite subset of n-space, $\mathcal{R}_n$. Let E now be regarded as a fundamental space of points t.

If $t \in E$, determine k so that $t = t_k$ and set $Q(t) = N(k)$, so that $Q(t)$ is the number of sample values taken from the population at t. For an arbitrary function $\theta(t)$ defined on E, let

$$J[\theta(t)] = \sum Q(t)F[\bar{y}(t), \theta(t)] \tag{4}$$

where the summation is over all $t \varepsilon E$, and where

$$\bar{y}(t) = \sum_{j=1}^{N(k)} g(x_{jk}) \,/\, N(k) \quad \text{if} \quad t = t_k . \tag{5}$$

The problem is to minimize $J(\theta(t)]$ in the class $\mathfrak{M}$ of functions $\theta(t)$ such that $\theta(t) \leqq \theta(t^*)$ whenever $t \prec t^*$.

It is clear from (1) that for fixed u, $F(u, v)$ is strictly decreasing as a function of v for $v < u$, and strictly increasing for $v > u$. It will be supposed that the given function $\bar{y}(t)$, defined for $t \varepsilon E$, is such that $F[\bar{y}(t), \bar{y}(t)] > -\infty$ for $t \varepsilon E$, from which it follows that $F[\bar{y}(t), \theta(t)] > -\infty$ for $t \varepsilon E$, for every function $\theta(t)$. Since $F[\bar{y}(t), c] \equiv 0$, and since the function $\theta(t) \equiv c$ is in $\mathfrak{M}$, the infimum of $J[\theta(t)]$ for $\theta(t) \varepsilon \mathfrak{M}$ is finite (and non-positive).

For a function $\theta(t)$ defined on E, set $\theta_k = \theta(t_k)$, $t_k \varepsilon E$. It is a consequence of the above property of F that the search for numbers $\theta_1, \cdots, \theta_m$ such that $\theta(t)$ is in $\mathfrak{M}$ and minimizes $J[\theta(t)]$ in $\mathfrak{M}$ can be restricted to the closed, bounded m-dimensional interval.

$$\min_{1 \leqq k \leqq m} \bar{y}_k \leqq \theta_j \leqq \max_{1 \leqq k \leqq m} \bar{y}_k , \qquad j = 1, 2, \cdots, m,$$

where $\bar{y}_k = \bar{y}(t_k)$, $k = 1, 2, \cdots, m$. The problem thus is to choose $(\theta_1, \cdots, \theta_m)$ in that closed part of this interval determined by the monotonicity requirement, so as to minimize the sum in (2). The continuity of this sum in $(\theta_1, \cdots, \theta_m)$ assures the existence of the minimum.

The following definitions represent slight modifications of corresponding definitions introduced in [3]. Each point of E determines a *lower interval*, the set of all points t^* in E such that $t^* \prec t$, and an *upper interval*, the set of all points t^* in E such that $t^* \succ t$. A union of lower intervals is a *lower layer*; a lower layer L is characterized by the property: if t is in L and t^* is not in L then it is not the case that $t \succ t^*$. A union of upper intervals is an *upper layer*, admitting an analogous characterization. The complement in E of a lower layer is an upper layer, and conversely. The union and intersection of two lower (upper) layers is a lower (upper) layer. If $A \subset E$, $\tilde{A}$ denotes the complement in E of A. If R, S are lower layers, then the set $\tilde{R} \cap S$ is called a *layer*, and is denoted by $\{R, S\}$. We note that if $\theta(t)$ is monotone non-decreasing in each of the variables $t^1, t^2, \cdots, t^n$, and if a is real, then the subset of E on which $\theta(t) < a$ is a lower layer, the subset of E on which $\theta(t) > a$ is an upper layer, and the subset of E on which $\theta(t) = a$ is a layer.

If $A \subset E$, let

$$J[\theta(t); A] = \sum_{t \varepsilon A} Q(t)F[\bar{y}(t), \theta(t)], \tag{6}$$

and, if A is not empty ($A \neq \varphi$; φ is the void set), set

$$(7) \qquad M(A) = \sum_{t \varepsilon A} Q(t)\bar{y}(t) \,/\, \sum_{t \varepsilon A} Q(t).$$

Thus $M(A)$ is a weighted mean of the numbers $\bar{y}(t)$ at points $t \varepsilon A$.

We note the following properties of the function F, consequences of the definition, (1).

If $A \subset E$, $A \neq \varphi$, and if θ is a fixed number in I, then

$$(8) \qquad \sum_{t \varepsilon A} Q(t)F[\bar{y}(t), \theta] = [\sum_{t \varepsilon A} Q(t)]F[M(A), \theta].$$

(9) For fixed $u \varepsilon I$, $F(u, v)$ as a function of v is strictly decreasing for $v < u$ and strictly increasing for $v > u$.

It follows from (8) and (9) that

$$(10) \qquad \min_{\theta} \sum_{t \varepsilon A} Q(t)F[\bar{y}(t), \theta] = [\sum_{t \varepsilon A} Q(t)]F[M(A), M(A)],$$

that is, the minimum of the left member of (8) is attained for $\theta = M(A)$.

Let $\Theta(t)$ denote the maximum likelihood estimate, the function of the class $\mathfrak{M}$ minimizing J. Let $\theta^1 < \theta^2 < \cdots < \theta^r$ be the r distinct values $\Theta(t)$ assumes, and let E^i denote the subset of E on which $\Theta(t) = \theta^i (i = 1, 2, \cdots, r)$.

Lemma 1. $M(E^i) = \theta^i$ $(i = 1, 2, \cdots, r)$. We have $\theta^i \geqq M(E^i)$, for it is clear from (8)–(10) that if $\theta^i < M(E^i)$ then it would be possible to decrease J by increasing $\Theta(t)$ on E^i while preserving the monotonicity of $\Theta(t)$. A similar remark shows that $\theta^i \leqq M(E^i)$.

Let S^i denote that subset of E (a lower layer) on which $\Theta(t) \leqq \theta^i$, and set $S^0 = \varphi$ (the void set). We have $E^i = \{\tilde{S}^{i-1}, S^i\} \subset S^i (i = 1, 2, \cdots, r)$.

Lemma 2. *If* $\{R, S^i\} \neq \varphi$ *then* $M\{R, S^i\} \leqq M(E^i)$. *If* $\{\tilde{S}^{i-1}, S\} \neq \varphi$, *then* $M\{\tilde{S}^{i-1}, S\} \geqq M(E^i)(i = 1, 2, \cdots, r)$.

For if $M\{R, S^i\}$ were greater than $M(E^i)$, it would be possible to decrease J by increasing $\Theta(t)$ on $\{R, S^i\}$ while preserving the monotonicity of $\Theta(t)$. A similar remark establishes the second statement.

Theorem 1. *For* $t_0 \varepsilon E$ *we have*

$$(11) \qquad \Theta(t_0) = \max_{R} \min_{S; t_0 \varepsilon \{R,S\}} M\{R, S\},$$

$$(12) \qquad \Theta(t_0) = \min_{S} \max_{R; t_0 \varepsilon \{R,S\}} M\{R, S\},$$

$$(13) \qquad \Theta(t_0) = \max_{R; t_0 \varepsilon \tilde{R}} \min_{S} M\{R, S\},$$

$$(14) \qquad \Theta(t_0) = \min_{S; t_0 \varepsilon S} \max_{R} M\{R, S\}.$$

Proof. Determine i so that $t_0 \varepsilon E^i$; then $\Theta(t_0) = \theta^i$. We have

$$\min_{S; t_0 \varepsilon \{R,S\}} M\{R, S\} \leqq M\{R, S^i\}$$

if $t_0 \varepsilon \tilde{R}$, therefore

$$\max_{R} \min_{S;t_0\varepsilon\{R,S\}} M\{R, S\} \leqq \max_{R;t_0\varepsilon\tilde{R}} M\{R, S^i\} = M(E^i) = \theta^i$$

by Lemmas 2 and 1. Also

$$\max_{R} \min_{S;t_0\varepsilon\{R,S\}} M\{R, S\} \geqq \min_{S;t_0\varepsilon\{S^{i-1},S\}} M\{S^{i-1}, S\} = M(E^i) = \theta^i$$

by Lemmas 2 and 1. The two inequalities yield (11), and the proofs of (12)–(14) are similar. (The above proof of Theorem 1 through Lemmas 1 and 2 is a further simplified proof suggested by the referee; it provides an alternative proof of Theorem 2.2 in the preceding paper.)

Since any of the formulas in Theorem 1 determines the value of a minimizing function $\Theta(t)$ at an arbitrary point of R, the minimizing function is unique on E.

5. Calculation of minimizing function.

LEMMA 3. *If* $S \supset S^i$, $\{S^i, S\} \neq \varphi$, *then* $M\{S^{i-1}, S\} > M\{S^{i-1}, S^i\} = M(E^i) = \theta^i$; *if* $R \subset S^{i-1}$, $\{R, S^{i-1}\} \neq \varphi$, *then*

$$M\{R, S^i\} < M(E^i) = \theta^i \qquad (i = 1, 2, \cdots, r).$$

For by Lemma 2 $M\{S^{i-1}, S\} \geqq M(E^i)$; if $M\{S^{i-1}, S\} = M(E^i)$ then J could be decreased by setting $\Theta(t) = M(E^i) = \theta^i$ on $\{S^{i-1}, S\}$. The proof of the second statement is similar.

It follows from Lemmas 2 and 3 that $E^1 = S^1$ is the *maximal lower layer of minimal mean*: that is, the union of those lower layers over which the mean is the minimum assumed by the means of all lower layers. The layers $E^i (i = 2, 3, \cdots, r)$ may be determined successively by the use of the criterion: $E^i = \{S^{i-1}, S^i\}$ is the *maximal layer of minimal mean among layers* $\{S^{i-1}, S\}$ $(i = 2, 3, \cdots, r)$. Similarly, $E^r = \{S^{r-1}, S^r\} = \{S^{r-1}, E\}$ is the *maximal upper layer of maximal mean*, and the layers $E^i (i = r - 1, r - 2, \cdots, 1)$ may be determined successively by the use of the criterion: $E^i = \{S^{i-1}, S^i\}$ is the *maximal layer of maximum mean among layers* $\{R, S^i\}$.

For the one-dimensional case ($n = 1$), the method of calculation discussed in [1] is also available.

6. Consistency. In order to discuss the consistency of the maximum likelihood estimator, let again $\mathfrak{R}_n$ be the fundamental space over which t ranges, and let the distribution parameter $\theta = \theta(t)$ be defined (though unknown to the investigator) on an open subset $\mathfrak{O}$ of $\mathfrak{R}_n$, and be monotone non-decreasing in each of the variables $t^i (i = 1, 2, \cdots, n)$. We shall suppose that the set E of points $t_k (k = 1, 2, \cdots, m)$ corresponding to populations from which sample values are to be taken is in $\mathfrak{O}$. For $t \varepsilon E$, let $\Theta(t)$ denote, as in Sections 4 and 5, the maximum likelihood estimator of $\theta(t)$. Let $\Theta(t)$ also denote an extension to $\mathfrak{O}$ of $\Theta(t)$ defined on E which preserves the monotonicity property; for the sake of definiteness, let $\Theta(t) = \max \Theta(t_k)$ for $t_k \varepsilon E$, $t_k \prec t$. For fixed t, the value of

$\Theta(t)$ depends on the sample values from populations at points of E; accordingly this value is a stochastic variable, which we denote by $\boldsymbol{\Theta}(t)$.

The content of the consistency theorem for the one-dimensional case ($n = 1$) may be stated roughly as follows: at a point, t_0, of continuity of $\theta(t)$, the estimator $\boldsymbol{\Theta}(t_0)$ will with high probability be near $\theta(t_0)$, provided enough sample values are taken at points t near t_0 such that $t \prec\prec t_0$, and enough sample values are taken at points t near t_0 such that $t \succ\succ t_0$. This consistency theorem generalizes that given in [1] for the binomial case.

As in Section 2 we suppose that $\mathbf{x}$ is a random variable having a density function $\exp\{-F[g(x), \theta]\}$ with respect to the measure ν, independent of θ, on the class of Borel sets, where $g(x)$ is a Baire function. The density function of the stochastic variable $\mathbf{y} = g(\mathbf{x})$ with respect to the induced measure $\chi = \nu g^{-1}$ is $\exp\{-F(y, \theta)\}$; for a Borel set B on the y axis,

$$\Pr\{\mathbf{y} \varepsilon B\} = \int_B \exp\{-F(y, \theta)\}\, d\chi$$

([6], p. 163). In particular,

$$\int_{-\infty}^{\infty} \exp\{-y[\tau(\theta) - \tau(c)]\}\, d\chi = \exp[\gamma(\theta)],$$

where $\gamma(\theta) = -\int_c^\theta z\, d\tau(z)$. With the change of parameter $\tau = \tau(\theta)$, this integral becomes a bilateral Laplace transform, representing an analytic function of τ in its strip of convergence ([9], p. 240; [5], p. 67). The above hypotheses on $\mathbf{x}$ in effect imply the convergence of the integral for $\theta \varepsilon I$ and hence the analyticity of $\gamma(\theta)$ and $\tau(\theta)$ on the interior of I. Straightforward calculations then yield

$$E(\mathbf{y}) = \theta, \qquad \operatorname{Var} \mathbf{y} = -1/\tau'(\theta) \tag{15}$$

Let $-1/\tau'[\theta(t)]$ be bounded:

$$-1/\tau'[\theta(t)] < C \qquad \text{for } t \varepsilon \Theta. \tag{16}$$

For the binomial distribution ($-1/\tau'(v) = v(1 - v)$, $0 \leqq v \leqq 1$) and the normal with mean θ and standard deviation $1(-1/\tau' = 1)$ for example, this is no restriction. For the Poisson ($-1/\tau'(v) = v$, $0 < v < \infty$) and the normal with mean 0, variance θ ($-1/\tau'(v) = 2v^2$), (16) requires that $\theta(t)$ be bounded in Θ.

For given $\epsilon > 0$ and $\eta > 0$, let $K = K(\epsilon, \eta)$ be a positive integer such that

$$\sum_{\nu=K}^{\infty} 1/\nu^2 + 1/4K < \varepsilon^2\eta/128C.$$

THEOREM 2. *Let $n = 1$. Let the above conditions on the family of populations and on the function $\theta(t)$ be satisfied. Let t_0 be a continuity point in Θ of $\theta(t)$. Let t', t'' be so chosen that $t' < t_0 < t''$ and that $|\theta(t) - \theta(t_0)| < \epsilon/2$* for $t' \leqq t \leqq t''$. Then

$$\Pr\{|\boldsymbol{\Theta}(t_0) - \theta(t_0)| < \epsilon\} > 1 - \eta$$

provided that at least $K = K(\epsilon, \eta)$ sample values are taken from populations at points in $[t', t_0]$ and at least K from populations at points in $[t_0, t'']$.

The proof requires only minor modifications in the proof of the special case in [1], and is omitted.

It is not difficult to see that for $n > 1$ the maximum likelihood estimator does not have precisely the consistency property which is the obvious analog of the consistency for $n = 1$. It is not true that, provided enough sample values are taken from populations at points near t_0 and "above" t_0, and enough from populations at points near t_0 and "below" t_0, the estimator $\Theta(t_0)$ will assume a value near $\theta(t_0)$ with high probability. This can be illustrated with an example in which the populations are binomial: each of the sample stochastic variables $\mathbf{y}_{jk}(j = 1, 2, \cdots, N(k))$ corresponding to a population at t_k assumes the value 1 with probability $\theta(t_k) = \frac{1}{2}$ and the value 0 with probability $1 - \theta(t_k) = \frac{1}{2}$; let $\theta(t) \equiv \frac{1}{2}$. Let t_0 be a point of $\mathfrak{O}$. Let E be an arbitrary finite set of points in $\mathfrak{O}$, and let the total number of sample values to be taken from populations at points of E be H. Let S_0 be a union of n-dimensional intervals, each of the form $\{t:t^1 \leqq a^1, t^2 \leqq a^2, \cdots, t^n \leqq a^n\}$, where $a = (a^1, a^2, \cdots, a^n) \varepsilon \mathfrak{R}_n$. Let the boundary of S_0 contain no segment parallel to a coordinate axis. Let $t_0 \varepsilon S_0$, and let S_0 be so chosen that $t_k \varepsilon \tilde{S}_0$, for every point $t_k \varepsilon E$ such that $t_k \succ\succ t_0$. (S_0 can be chosen, for example, so that its boundary is a hyperplane separating t_0 from points $t_k \succ\succ t_0$.) Now add to E points $t^* \succ\succ t_0$, *on the boundary of* S_0, obtaining a set E^* containing E, and take one sample value from each of the points added. By adding enough such points to E, the probability can be made arbitrarily near 1 that at least $9H$ of the corresponding sample values will be 0. Let S_1 be a lower layer (with reference to the extended set E^*) containing those points, and only those among the points added, where the sample value is 0. Then if $t_0 \varepsilon \{R, S_1\}$ we have $M\{R, S_1\} \leqq H/10H = 1/10$, hence

$$\Theta(t_0) = \max_{R} \min_{S:t_0\varepsilon\{R,S\}} M\{R, S\} \leqq 1/10.$$

It is clear that adding points to E near t_0 need not bring $\Theta(t_0)$ near $\theta(t_0) = \frac{1}{2}$ with high probability, but can, indeed, if the points are added in a suitable way, bring $\Theta(t_0)$ arbitrarily near 0 with arbitrarily high probability.

It appears not unreasonable to expect that the consistency property will hold if the points of E are required, for example, to constitute a rectangular array. To decide whether or not this is the case would appear to require a study of properties of maximum means over layers in rectangular arrays, with a view to obtaining an analogue of the theorem of Kolomogorov on which the proof of Theorem 2 is based (cf. [1]).

The above example suggests the possible desirability of grouping together observations made at closely adjacent points, to avoid the paradox of the example in which more observations may be made to yield less precise results.

It should perhaps be mentioned that if the points of E are held fixed while the size of the sample from the population at each is increased indefinitely, then it follows from the strong law of large numbers that, with probability 1, $\Theta(t)$ will approach $\theta(t)$ at each point of E.

REFERENCES

[1] MIRIAM AYER, H. D. BRUNK, G. M. EWING, W. T. REID, AND EDWARD SILVERMAN, "An empirical distribution function for sampling with incomplete information," *Ann. Math. Stat.*, Vol. 26 (1955), pp. 641–647.

[2] H. D. BRUNK, G. M. EWING, AND W. T. REID, "The minimum of a certain definite integral suggested by the maximum likelihood estimate of a distribution function," (Abstract) *Bull. Amer. Math. Soc.*, Vol. 60 (1954), p. 535.

[3] H. D. BRUNK, G. M. EWING, AND W. R. UTZ," Minimizing integrals in certain classes of monotone functions," unpublished.

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

[5] M. A. GIRSHICK AND L. J. SAVAGE, "Bayes and minimax estimates for quadratic loss functions," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1951, pp. 53–73.

[6] P. R. HALMOS, *Measure Theory*, Van Nostrand, New York, 1950.

[7] B. O. KOOPMAN, "On distributions admitting a sufficient statistic," *Trans. Amer. Math. Soc.*, Vol. 39, (1936), pp. 399–409.

[8] E. J. G. PITMAN, "Sufficient statistics and intrinsic accuracy," *Proc. Camb. Phil. Soc.*, Vol. 32 (1936), p. 567–579.

[9] D. V. WIDDER, *The Laplace Transform*, Princeton University Press, 1941.