The Annals of Mathematical Statistics
1970, Vol. 41, No. 6, 2169–2174

MAXIMUM LIKELIHOOD ESTIMATION OF A UNIMODAL DENSITY, II[1]

By Edward J. Wegman

University of North Carolina

This paper is a sequel to the earlier paper, "Maximum Likelihood Estimation of a Unimodal Density Function." The MLE of a unimodal density with unknown mode is shown to agree, for sufficiently large n and on certain regions, with the MLE of a unimodal density with known mode. The asymptotic distributions of the MLE's then agree. Also a geometrical interpretation of the MLE of a unimodal density with unknown mode is given.

1. Introduction. Several authors, Grenander [3], Robertson [6], and Rao [4], have described the MLE for a unimodal density when the mode was known as well as some of the estimate's properties. A MLE for a unimodal density when the mode is unknown was described in [7]. Strong consistency was also established in [7]. We wish to describe some additional properties in this paper.

2. Asymptotic distribution. The estimates discussed in this paper rely heavily on the notions of σ-lattices and conditional expectations with respect to σ-lattices. A σ-lattice, $\mathscr{L}$, of subsets of a measure space, $(\Omega, \mathscr{A}, \mu)$, is a collection of subsets of Ω closed under countable unions and countable intersections and containing both the empty set ϕ and Ω. A function is measurable with respect to a σ-lattice, $\mathscr{L}$, if the set, $[f > a]$, is in $\mathscr{L}$ for every real a. If Ω is the real line, $\mathscr{A}$ is the collection of Borel sets and $\mu = \lambda$ is Lebesgue measure, let L_2 be the set of square-integrable functions and $L_2(\mathscr{L})$ be those members of L_2 which are also measurable with respect to $\mathscr{L}$.

Definition. If $f \in L_2$, then $g \in L_2(\mathscr{L})$ is equal to $E(f \mid \mathscr{L})$, the conditional expectation of f given $\mathscr{L}$, if and only if

$$\int f \cdot \theta(g)\, d\lambda = \int g \cdot \theta(g)\, d\lambda$$

for every θ, a real-valued function such that $\theta(g) \in L_2$ and $\theta(0) = 0$ and

$$\int (f-g)h\, d\lambda \leqq 0$$

for each $h \in L_2(\mathscr{L})$.

The collection of intervals about a fixed point, m, together with ϕ is a σ-lattice which we denote as $\mathscr{L}(m)$. A function, f, is unimodal at M by definition if f is measurable with respect to $\mathscr{L}(\cdot M)$. It is not difficult to see that this is equivalent to f non-decreasing at $x < M$ and f non-increasing at $x > M$. If f is unimodal at every point of an interval, I, then we call I the modal interval and write $\mathscr{L}(I)$ for the lattice

Received January 12, 1970; revised June 1, 1970.

[1] The research for this work was partially sponsored by the National Science Foundation under Grant No. GU-2059.

of intervals containing I. Clearly, f has modal interval I if and only if f is measurable with respect to $\mathscr{L}(I)$.

Now let $\hat{f}_n$ be the maximum likelihood estimate when the mode is unknown as described in [7], and let f_n^* be the maximum likelihood estimate when the mode is known. In defining $\hat{f}_n$, $\varepsilon \geqq 0$ was a predetermined number and $\hat{f}_n$ was chosen to be the maximum likelihood estimate over the class of densities whose modal interval is at least ε in length. The special case $\varepsilon = 0$ was discussed by Robertson [6], Example 1.2. As far as this author knows, consistency results exist only for $\varepsilon > 0$ (see [7]). We shall now describe $\hat{f}_n$.

If $[L, R]$ is any fixed interval of length ε and $y_1 < y_2 < \cdots < y_n$ are the ordered observations sampled according to the density f, let $A_1 = [y_1, y_2)$, $A_2 = [y_2, y_3), \cdots$, $A_{\ell(n)} = [y_{\ell(n)}, L)$, $A_{\ell(n)+1} = [L, R]$, $A_{\ell(n)+2} = (R, y_{r(n)}], \cdots$, $A_k = (y_{n-1}, y_n]$. Here $y_{\ell(n)}$ and $y_{r(n)}$ are respectively the largest observations smaller than L and the smallest observations larger than R. Now $\mathscr{L}([L, R])$ is the σ-lattice of intervals containing $[L, R]$ and the maximum likelihood estimate with modal interval $[L, R]$ is given by $E(\hat{g}_n \mid \mathscr{L}([L, R]))$ where

$$\hat{g}_n = \sum_{i=1}^{k} n_i \cdot [n\lambda(A_i)]^{-1} \cdot I_{A_i}.$$

Here n_i is the number of observations in A_i and I_{A_i} is the indicator of A_i. Thus we have the MLE given the interval. We need only let the interval vary over all possible intervals of length ε to find the MLE, $\hat{f}_n$. Fortunately, it is shown in [7], that we do not have to consider all possible intervals, but only those for which either L or R is an observation. Hence $\hat{f}_n$ is the density which has the largest likelihood product among the $2n$ candidates obtained by letting L range among $y_1, \cdots, y_n$ and by letting R range among $y_1, \cdots, y_n$. Let us denote the modal interval of $\hat{f}_n$ by $[L_n, R_n]$. In [7], it is shown that L_n and R_n have limits with probability one for which we shall henceforth reserve L and R respectively. See [7] and, in particular, Theorem 5.1 for the exact conditions for this convergence.

In a similar manner, let $A_1^* = [y_1, y_2), \cdots, A_{q(n)}^* = [y_{q(n)}, M)$, $A_{q(n)+1}^* = [M, y_{q(n)+1}]$, $A_{q(n)+2}^* = (y_{q(n)+1}, y_{q(n)+2}], \cdots, A_n^* = (y_{n-1}, y_n]$. Here M is the known mode and $y_{q(n)}$ is the largest observation smaller than M. Notice with probability one, $M \neq y_j$ for each j. If $\mathscr{L}(M)$ is the σ-lattice of intervals containing M, the maximum likelihood estimate, f_n^*, is given by $E(g_n^* \mid \mathscr{L}(M))$ where

$$g_n^* = \sum_{i=1}^{n} n_i^* \cdot [n\lambda(A_i^*)]^{-1} \cdot I_{A_i^*}.$$

Of course, n_i^* is the number of observations in A_i^*. In [7], it is shown that $M \in (L, R)$, hence $\hat{g}_n$ and g_n^* agree except possibly on $[y_{\ell(n)}, y_{r(n)}]$. A similar situation was the case in Lemma 5.4 in [7]. If we require only that some neighborhood of L, say N_L, is a set of points of increase of f and similarly some neighborhood of R, say N_R, is a set of points of decrease of f, we may use the arguments of Lemma 5.4 in [7] to obtain

LEMMA 2.1. *Let $\eta > 0$ be an arbitrary number such that $L-\eta$ and $R+\eta$ are elements of N_L and N_R respectively. Then with probability one, for sufficiently large n, $\hat{f}_n$ and $\hat{f}_n^*$ agree on $(L-\eta, R+\eta)^c$.*

Before we give the proof of Lemma 2.1, we reproduce part of a result of Robertson [5] upon which we shall rely. This result characterizes conditional expectation when the underlying measure space has finite total measure. (This will apply to our situation since we may restrict our attention to $[y_1, y_n]$.) If $\mathscr{L}$ is any σ-lattice and g is a function in L_2, then $E(g \mid \mathscr{L}) = f$ may be represented by

$$(2.1) \qquad f(y_0) = \inf_{L \in H_1(T_t)} [\lambda(T_t - L)]^{-1} \cdot \int_{T_t - L} g \, d\lambda \qquad \text{and}$$

$$(2.2) \qquad f(y_0) = \sup_{L \in H_2(P_t)} [\lambda(L - P_t)]^{-1} \cdot \int_{L - P_t} g \, d\lambda.$$

Here $t = f(y_0)$, $P_t = [f > t]$, $T_t = [f \geqq t]$, $H_1(T_t) = \{L' \in \mathscr{L} : \lambda(T_t - L') > 0\}$ and $H_2(P_t) = \{L' \in \mathscr{L} : \lambda(L' - P_t) > 0\}$. Robertson's theorem is more general than stated here, but this is sufficient for our needs.

PROOF OF 2.1. We shall first consider agreement to the left of L. Pick t_0 in $(L - \eta, L) \subset N_L$. Let $\delta \in (0, L - t_0)$. Since the underlying density f has a point of increase in $(t_0, L - \delta)$, $\hat{f}_n$ and f_n^* must both eventually have a jump in $(t_0, L - \delta)$ since they are both consistent estimates of $f(x)$ for $x < L$. Let y_i be the smallest observation greater than t_0 at which $\hat{f}_n$ has a jump and let y_{i^*} be the smallest observation greater than t_0 at which f_n^* has a jump. (A result of [6] is that jumps occur in f_n^* only at observations. Similarly in [7] it is shown that jumps in $\hat{f}_n$ occur only at observations or at L_n or R_n.) Without loss of generality, we may assume $y_{i^*} \geqq y_i$. Let y_{i^*} be the largest observation smaller than or equal to t_0 at which there is a jump in f_n^*. Thus we have $y_{j^*} \leqq t_0 < y_i \leqq y_{i^*}$. We wish to show equality holds in the last inequality. Assume $y_i < y_{i^*}$. Let $t = f_n^*(t_0)$ and $T_t = [f_n^* \geqq t]$, so that $T_t = [y_{j^*}, y_{k^*}]$. But $[y_i, y_{k^*}] \in \mathscr{L}(M)$, so by (2.1)

$$t \leqq (y_i - y_{j^*})^{-1} \cdot \int_{[y_{j^*}, y_i]} g_n^* \, d\lambda.$$

In a similar manner using (2.2) we can show

$$t \geqq (y_{i^*} - y_i)^{-1} \cdot \int_{[y_i, y_{i^*}]} g_n^* \, d\lambda.$$

Since g_n^* and $\hat{g}_n$ agree except possibly on $[y_{\ell(n)}, y_{r(n)}]$, (here, of course, $y_{\ell(n)}, y_{r(n)}$ and $\hat{g}_n$ are defined with respect to $[L_n, R_n]$) we have for sufficiently large n,

$$(2.3) \qquad (y_{i^*} - y_i)^{-1} \cdot \int_{[y_i, y_{i^*}]} \hat{g}_n \, d\lambda \leqq (y_i - y_{j^*})^{-1} \cdot \int_{[y_{j^*}, y_i]} \hat{g}_n \, d\lambda.$$

Now let $t = \hat{f}_n(t_0)$ and $P_t = [\hat{f}_n > t]$. Again by use of (2.2), we have

$$(2.4) \qquad \hat{f}_n(t_0) \geqq (y_i - y_{j^*})^{-1} \cdot \int_{[y_{j^*}, y_i]} \hat{g}_n \, d\lambda.$$

Finally letting $t = \hat{f}_n(y_i)$ and using (2.1), we have

$$(2.5) \qquad \hat{f}_n(y_i) \leqq (y_{i^*} - y_i)^{-1} \cdot \int_{[y_i, y_{i^*}]} \hat{g}_n \, d\lambda.$$

Using (2.3), (2.4) and (2.5), $\hat{f}_n(t_0) \geqq \hat{f}_n(y_i)$. But y_i is a jump point in $\hat{f}_n$, so $\hat{f}_n(t_0) < \hat{f}_n(y_i)$. Thus our assumption $y_i \neq y_{i^*}$ is false. Robertson [6] shows that there is an $L = [u, v]$ such that $L \in \mathscr{L}(M)$ and

$$f_n^*(t_0) = \lambda(L - P_t)^{-1} \cdot \int_{L - P_t} g_n^* \, d\lambda.$$

Rewriting this

$$f_n^*(t_0) = (y_i - u + v - y_{k^*})^{-1} \cdot \int_{[u,y_i] \cup [y_{k^*},v]} g_n^* \, d\lambda.$$

But by (2.2),

$$f_n^*(t_0) \geqq (v - y_{k^*})^{-1} \cdot \int_{[y_{k^*},v]} g_n^* \, d\lambda.$$

From these two displays we may obtain $f_n^*(t_0) \leqq (y_i - u)^{-1} \cdot \int_{[u,y_i]} g_n^* \, d\lambda$.

Hence for sufficiently large n

$$f_n^*(t_0) = \sup_{u \leqq y_i} \{(y_i - u)^{-1} \cdot \int_{[u,y_i]} g_n^* \, d\lambda\}.$$

Similarly, for sufficiently large n,

$$\hat{f}_n(t_0) = \sup_{u < y_i} \{(y_i - u)^{-1} \cdot \int_{[u,y_i]} \hat{g}_n \, d\lambda\}.$$

Since $\hat{g}_n$ and g_n^* agree eventually in this region, $\hat{f}_n$ and f_n^* must be equal eventually at t_0. For any $t < t_0$, by virtue of the fact that $\hat{f}_n(t_0) = f_n^*(t_0)$ and by use of (2.1) and (2.2), we obtain the desired conclusion to the left of L. We may obtain similar results to the right of R to complete the proof.

Hence, for any $x \notin [L, R]$, for sufficiently large n, $\hat{f}_n(x) = f_n^*(x)$. An immediate theorem follows

THEOREM 2.1. *For $x \notin [L, R]$, $\hat{f}_n(x)$ has the same asymptotic distribution as $f_n^*(x)$.*

Rao [4], through some very clever but rather tedious arguments, develops the asymptotic distribution of $f_n^*(x)$. Arguments similar to these could be applied to $\hat{f}_n(x)$, but are avoided by use of Lemma 2.1. Rao assumes a non-zero derivative of the density, f, at each point x where the asymptotic distribution is to be found.

3. A characterization of $\hat{f}_n$. Grenander [3] gives a characterization of the MLE of a strictly monotone density. Reid (see [1] and [2]) gave a geometrical interpretation of a conditional expectation with respect to a σ-lattice, $\mathscr{L}$, when $\mathscr{L}$ consists of intervals with the right (or left) endpoint fixed. If the σ-lattice is $\mathscr{L}(M)$, the conditional expectation may be characterized by applying Reid's method individually to the right and to the left of M. To find $E(h \mid \mathscr{L}(M))$, the conditional expectation of some function h with respect to $\mathscr{L}(M)$, determine $H(x) = \int_{(-\infty,x]} h \, d\lambda$. To the left of M, $E(h \mid \mathscr{L}(M))$ is given by the slope of the greatest convex minorant of H restricted to $(-\infty, M)$ and to the right of M, by the slope of the least concave majorant of H restricted to (M, ∞).

Let us assume that h has bounded support, $\{x : h(x) \neq 0\}$. Let L and R be fixed with $R - L = \varepsilon$. We want a geometrical interpretation of the conditional expectation of h with respect to $\mathscr{L}([L, R])$.

We shall use the theorem of Robertson [5] mentioned in Section 2. Recall that Robertson's theorem is stated for a finite measure space, hence the requirement here that we have bounded support. Let $E(h \mid \mathscr{L})(x_0) = y_0$ and $P_{y_0} = \{x : E(h \mid \mathscr{L})(x) > y_0\}$. Let $\mathscr{H} = \{L^* \in \mathscr{L} : \lambda(L^* - P_{y_0}) > 0\}$. Then recall

$$y_0 = \sup_{L^* \in \mathscr{H}} [\lambda(L^* - P_{y_0})]^{-1} \cdot \int_{L^* - P_{y_0}} h \, d\lambda.$$

If we let $\mathscr{L} = \mathscr{L}([L, R])$ it is clear that since $E(h \mid \mathscr{L})$ is by definition $\mathscr{L}$-measurable, it achieves its maximum throughout $[L, R]$. Hence if $x_0 \in [L, R]$, P_{y_0} is empty so that

$$y_0 = \sup_{L^* \in H} (\lambda(L^*))^{-1} \cdot \int_{L^*} h \, d\lambda. \tag{3.1}$$

In fact, this supremum is a maximum and $\mathscr{H} = \mathscr{L}([L, R])$. Let L^* be the maximizing interval so that

$$y_0 = (\lambda(L^*))^{-1} \cdot \int_{L^*} h \, d\lambda. \tag{3.2}$$

Since for any $x \in [L, R]$, $E(h \mid \mathscr{L})(x) = E(h \mid \mathscr{L})(x_0)$, we have

$$E(h \mid \mathscr{L})(x) = (\lambda(L^*))^{-1} \cdot \int_{L^*} h \, d\lambda.$$

Assume $x \in L^* - [L, R]$ and assume

$$E(h \mid \mathscr{L})(x) < (\lambda(L^*))^{-1} \cdot \int_{L^*} h \, d\lambda. \tag{3.3}$$

If $y = E(h \mid \mathscr{L})(x)$, then it is clear that P_y is not empty and by (2.2)

$$E(h \mid \mathscr{L})(x) \geqq (\lambda(L^* - P_y))^{-1} \cdot \int_{L^* - P_y} h \, d\lambda, \qquad \text{or}$$

$$[\lambda(L^* - P_y)]^{-1} \cdot \int_{L^* - P_y} h \, d\lambda < [\lambda(L^*)]^{-1} \cdot \int_{L^*} h \, d\lambda.$$

Let $P_y{}^* = P_y \cap L^*$. By some elementary algebraic manipulations, we obtain

$$[\lambda(L^*)]^{-1} \cdot \int_{L^*} h \, d\lambda < [\lambda(P_y{}^*)]^{-1} \cdot \int_{P_y{}^*} h \, d\lambda.$$

But since $P_y{}^* \in \mathscr{L}$, this is a contradiction to (3.1). Hence (3.3) cannot hold. So for any $x \in L^*$,

$$E(h \mid \mathscr{L})(x) = [\lambda(L^*)]^{-1} \cdot \int_{L^*} h \, d\lambda.$$

Let $a = \inf L^*$ and $b = \sup L^*$. As in the case of the conditional expectation with respect to $\mathscr{L}(M)$, it is not difficult to see we may apply Reid's method individually to the left of a and to the right of b. Thus we have,

THEOREM 3.1. *The conditional expectation of a function, h, with bounded support, with respect to a σ-lattice, $\mathscr{L}([L, R])$, is given by the following procedure.*

Find the interval $[a, b]$ containing $[L, R]$ such that $(H(b) - H(a))/(b - a)$ is maximized. On $[a, b]$, the conditional expectation is given by $(H(b) - H(a))/(b - a)$. To the left of a, it is the slope of the greatest convex minorant of H restricted to $(-\infty, a)$ and to the right of b, it is the slope of the least concave majorant of H restricted to (b, ∞).

The application of this theorem to the finding of MLE is of particular interest since this would give an algorithm for computing the MLE. Let $[L, R]$ be an arbitrary interval of length ε and the sets A_i and the function $\hat{g}_n$ be defined as in the first part of Section 2. Finally let $\hat{G}_n(x) = \int_{(-\infty, x]} \hat{g}_n \, d\lambda$.

COROLLARY 3.1. *If $h = \hat{g}_n$ in Theorem 3.1, $E(\hat{g}_n \mid \mathscr{L}([L, R]))$ may be computed by the procedure set forth in Theorem 3.1. Moreover, the function $\hat{G}_n$ may be replaced by F_n, the empirical distribution function, in the procedure of Theorem 3.1.*

Noting that $\hat{g}_n$ has bounded support, $[y_1, y_n]$, is sufficient to prove the first part of this corollary. To the left of L, $\hat{G}_n$ is a minorant of F_n and to the right of R, $\hat{G}_n$ is a majorant of F_n. Using these facts, it is not difficult to see that the procedure in Theorem 3.1 gives the same result whether it is applied to F_n or to $\hat{G}_n$.

It is interesting to note that Theorem 3.1 implies Theorem 3.1 of [7] if the condition of f being continuous is exchanged for f having bounded support. The author is indebted to the referee of [7] for pointing this out.

Finally, the author would like to point out that in [7] a printing error was made in Figure 1. Part of this figure was left out, the maximum values of the estimate, which illustrate the peaking of f_n*-type estimates. This is one reason for considering estimates $\hat{f}_n$ with $\varepsilon > 0$. Table 1, fortunately, reflects this peaking in tabular form.

REFERENCES

[1] Brunk, H. D. (1956). On an inequality for convex functions. *Proc. Amer. Math. Soc.* **7** 817–824.

[2] Brunk, H. D., Ewing, G. M. and Reid, W. T. (1954). The minimum of a certain definite integral suggested by the maximum likelihood estimate of a distribution function, (abstract). *Bull. Amer. Math. Soc.* 684.

[3] Grenander, Ulf. (1956). On the theory of mortality measurement, Part II. *Skand. Aktuarietidskr.* **39** 125–153.

[4] Rao, B. L. S. Prakasa (1969). Estimation of a unimodal density. *Sankyā Ser. A* **31** 23–36.

[5] Robertson, Tim (1966). A representation for conditional expectations given σ-lattices. *Ann. Math. Statist.* **37** 1279–1283.

[6] Robertson, Tim (1967). On estimating a density which is measurable with respect to a σ-lattice. *Ann. Math. Statist.* **38** 482–493.

[7] Wegman, E. J. (1970). Maximum likelihood estimation of a unimodal density function. *Ann. Math. Statist.* **41** 457–471.