Research partially supported by the National Science Foundation Grants GU-2059 and GJ-306.

# NONPARAMETRIC PROBABILITY DENSITY ESTIMATION

by

EDWARD J. WEGMAN
Department of Statistics

University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 638

JULY 1969

# Nonparametric Probability Density Estimation

by Edward J. Wegman

1. Introduction and Historical Review. This paper is intended to be an expository review of efforts made in nonparametric estimation of probability densities. We also include some Monte Carlo simulation in Section 5 to give the reader a feeling for the appearance of the estimates as well as a feeling for their relative merits.

Nonparametric density estimation is a type of procedure which selects the estimate from a class of densities which has the cardinality of the set of real-valued real functions. In contrast, the class of densities from which the estimate is selected in the parametric case has the cardinality of the real line or less. In this case, the densities are usually indexed by elements from a set whose cardinality is less than or equal to that of the real line. The elements of the indexing set are called the parameters. In either case, we shall usually require some prior information about the density in order to estimate successfully (consistently) the density function.

The first attempts to estimate probability density functions appear to have been made by Karl Pearson in a series of papers culminating with [23] and [24]. The Pearsonian system of densities is the set of solutions to the differential equation

$$\frac{df}{dx} = \frac{(x - a)f}{b_0 + b_1 x + b_2 x^2},$$

Research partially supported by the National Science Foundation Grants GU-2059 and GJ-306.

where $a$, $b_0$, $b_1$, and $b_2$ are constants. The solutions are unimodal with mode $a$ and have smooth contact with the $x$ axis as $f$ tends to zero. It may be shown that the four constants are expressible in terms of the first four moments of the density $f$. Hence, the estimation procedure is to compute the sample moments from which the estimates of $a$, $b_0$, $b_1$, and $b_2$ may be found, and, then, to solve the differential equation. Pearson identified many types of densities, some of which are of little interest today. What is remarkable, however, is that the Pearsonian system contained almost all distributions known in his era. Consequently, the Pearsonian system enjoyed great popularity. The standard account of the Pearsonian system may be found in Elderton [7].

The Gram-Charlier approach is to consider a density representable by the series

$$f(x) = \alpha(x) \sum_{j=0}^{\infty} c_j H_j(x)$$

where $\alpha(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp(-\frac{1}{2}x^2)$ and $H_j(x) = (-1)^j \cdot \alpha^{-1}(x) \cdot \frac{d^j \alpha(x)}{dx^j}$.

$H_j$ is called the $j$-th Hermite polynomial and $c_j = \frac{1}{j!} \int_{\infty}^{\infty} f(x) H_j(x) dx$. By substituting the actual polynomial and integrating, one may express $c_j$ in terms of the moments of $f$. Using estimates of the moments, one may obtain an estimate of the density. The flaw in the scheme is the relatively slow convergence of higher moments. A closely related type of estimate is found in Schwartz [31]. We shall deal with this estimate later. An equation identical with the Gram-Charlier series was obtained by Edgeworth [6] in 1904! For a more complete discussion of these and other methods, see Kendall and Stewart [14]. The predecessor of the latter volume, Kendall [13], is commended to the reader's attention for a bibliography on early attempts at density estimation.

To close this section, we point out the paper by R.A. Fisher [10] which proposes the maximum likelihood procedure for parametric density estimation. This method is, of course, very powerful and continues to enjoy great popularity.

2. General Results and Types of Consistency. Perhaps the earliest published paper on modern nonparametric density estimation was that of Rosenblatt [29]. Fix and Hodges [11] consider nonparametric density estimates in connection with nonparametric discrimination. An interesting, but somewhat disappointing result of the Rosenblatt paper was a result concerning unbiasedness. If $X_1, \ldots, X_n$ are iid random variables whose continuous density function is $f$, then any estimate $f(y; x_1, \ldots, x_n)$, which is both jointly Borel measurable in $(y, x_1, \ldots, x_n)$ and nonnegative, cannot be unbiased for all $y$. It is clear that the measurability and nonnegativity are most desirable properties for the estimate. In practice, to obtain consistency results, most writers require continuity of the density function, at least. Hence it appears that any estimate we consider will not be unbiased for all $y$.

Farrell [9] considers a set $C_\alpha$ of distribution functions with the property that the density has a continuous derivative on the real line and that the density is bounded by $\alpha$. For the estimates considered in this paper, Farrell's results imply that "no uniformly consistent sequence of estimators exists relative to the class $C_\alpha$." By consistent is meant that the mean square error, $E(\hat{f}_n(y) - f(y))^2$, converges to 0,

where $\hat{f}_n(y) = f(y; x_1, \ldots, x_n)$. Farrell's result is stated in terms of a sequential procedure whose stopping variable has finite expectation.

The question of consistency is one which must be carefully examined, for "consistent" takes on a wide variety of meanings. The type of consistency stated above is called pointwise consistency in quadratic mean. A related type of consistency is uniform consistency in quadratic mean which occurs when $E(\hat{f}_n(y) - f(y))^2$ converges to 0 uniformly in y. A third type is called integratedly uniformly consistent in quadratic mean. This occurs when $E\int_{-\infty}^{\infty} (\hat{f}_n(y) - f(y))^2 dy$ converges to 0. The term $E\int_{-\infty}^{\infty} (\hat{f}_n(y) - f(y))^2 dy$ is frequently called mean integrated square error and is abbreviated M.I.S.E. A final type of mean consistency is functionally uniformly consistent in quadratic mean; when $E(\sup_x |\hat{f}_n(y) - f(y)|^2)$ converges to 0.

Similar types of consistency may be defined in probability and with probability one. Thus by functionally uniformly consistent in probability, we mean $\Pr[\sup_y |\hat{f}_n(y) - f(y)| < \epsilon]$ converges to 1 for all $\epsilon > 0$. This type of consistency is sometimes called simply uniformly consistent in probability. By pointwise consistent with probability one, we shall mean $|\hat{f}_n(y) - f(y)| \to 0$ with probability one and by uniformly consistent with probability one, $\sup_y |\hat{f}_n(y) - f(y)| \to 0$ with probability one. One other variant shall appear, that is almost uniformly with probability one which occurs when $\sup_{y \in A} |\hat{f}_n(y) - f(y)|$ converges to 0 with probability one where A is the complement of a set of arbitrarily small measure.

Thus it is clear that all other things being equal, "uniform" consistency is to be preferred over "pointwise", and "with probability one" is to be preferred over "in probability" or "in quadratic mean". A

special note should be taken of the advantage of consistency "with probability one" over the consistency "in quadratic mean" or "in probability". The latter two offer no assurances of convergence once one sample has been selected, whereas with the former we are assured of convergence with certainty.

3. Kernal Estimates. Let $X_1, \ldots, X_n, \ldots$ be a sequence of iid random variables with probability density function $f$. We shall focus our attention on estimates of the form

$$(3.1) \qquad \hat{f}_n(x) = \int_{-\infty}^{\infty} K_n(x, y)dF_n(y) = \frac{1}{n}\sum_{j=1}^{n} K_n(x, X_j) .$$

Here, $F_n$ is the empirical distribution function based on the first $n$ observations.

Rosenblatt [29] considers a "naive" estimate

$$\hat{f}_n(x) = \frac{F_n(x+h) - F_n(x - h)}{2h} .$$

This estimate is the special case of (3.1) when $K_n(x,y)$ is $\frac{1}{2h}$ for $|x - y| \le h$ and $0$ elsewhere. Of course, no assumptions are necessary on $f$ to form this estimate. However, Rosenblatt goes on to consider some asymptotic properties. He shows

$$E|\hat{f}_n(x) - f(x)|^2 \sim \frac{f(x)}{2hn} + \frac{h^4}{36}|f''(x)|^2 + o(\frac{1}{hn} + h^4),$$

under the assumption that the first three derivatives of $f$ exist at $x$. Of course, the problem is to choose the sequence of $h = h(n)$ converging to $0$ at an appropriate rate. If $h = kn^{-\alpha}$, $\alpha > 0$, the optimum choice

of $\alpha$ is $\frac{1}{5}$ and the optimum value of $k$ is

$$\left[\frac{9}{2}\frac{f(x)}{|f''(x)|^2}\right]^{1/5}.$$

In a similar manner, if one is concerned with M.I.S.E., the optimum choice of $\alpha$ remains $\frac{1}{5}$, but the optimum choice of $k$ is

$$\left[\frac{9}{2\int_{-\infty}^{\infty}|f''(x)|^2dx}\right]^{1/5}.$$

Of course, since we are attempting to estimate $f$, it is unlikely that we will know enough to choose optimum $k$. Nonetheless, with a satisfactory choice of the constant $k > 0$, we should still have either pointwise consistency in quadratic mean or integrated consistency in quadratic mean. Rosenblatt also proposes estimates with $K_n(x,y) = \frac{1}{h(n)}K(\frac{x-y}{h(n)})$ where $K$ is a nonnegative function such that

(3.2) $$\int_{-\infty}^{\infty}K(u)du = 1$$

(3.3) $$\int_{-\infty}^{\infty}[K(u)]^2du < \infty$$

(3.4) $$\int_{-\infty}^{\infty}[K(u)]|u|^3du < \infty$$

(3.5) $K(u)$ is symmetric about 0.

We shall call these types of estimates algebraic estimates. Under these conditions and the condition that $f$ have derivatives of the first 3 orders, the optimum choice of $h$ leads to mean square error, $E(\hat{f}_n(x) - f(x))^2$ no smaller than $0(n^{-4/5})$.

Parzen [22] considers algebraic estimates and requires that the nonnegative, even, Borel function $K$ satisfy (3.2) and

$$(3.6) \qquad \sup_{-\infty < x < \infty} |K(x)| < \infty$$

$$(3.7) \qquad \int_{-\infty}^{\infty} |K(x)|dx < \infty$$

$$(3.8) \qquad \lim_{|x| \to \infty} |xK(x)| = 0.$$

He denotes such functions as weighting functions. Table 1 gives some suggested weighting functions. The asymptotic variance of $\hat{f}_n(x)$ is given by $[f(x) \cdot \int_{-\infty}^{\infty} K^2(y)dy] \cdot [nh(n)]^{-1}$, at continuity points of $f$. Hence, we desire to minimize $\int_{-\infty}^{\infty} K^2(y)dy$ and if the variance is to converge to zero, we require $\lim_{n \to \infty} nh(n) = \infty$. If the bias, $E(\hat{f}_n(x) - f(x))$, is to converge to zero at continuity points, $x$, of $f$, we require that $h(n) \to 0$ and $f \in L_1$, the set of integrable functions. Since the mean square error may be written as the sum of the variance and the square of the bias, under these conditions we obtain pointwise consistency in quadratic mean. In addition, Parzen shows the sequence of estimates are asymptotically normal.

By requiring $k(u) = \int_{-\infty}^{\infty} e^{-iuy}K(y)dy$, the Fourier transform of $K$ to be absolutely integrable, $\lim_{n \to \infty} nh^2(n) = \infty$ and the density $f$ to be uniformly continuous, Parzen obtains uniform consistency in probability. In addition, the mode of the estimate (the value maximizing $\hat{f}_n$) converges in probability to the mode of the density $f$, where the mode is that unique value of $x$ for which $f(x)$ is maximized.

The sequence $h(n)$ which minimizes mean square error is

$$h(n) = \{f(x) \int_{-\infty}^{\infty} K^2(y)dy\}/\{n^{2r}|k_r f^{(r)}(x)|^2\}^{\frac{1}{(2r+1)}}$$

Table 1

Suggested Weighting Functions

1. $$K(y) = \begin{cases} \frac{1}{2} & |y| \le 1 \\ 0 & |y| > 1 \end{cases}$$

2. $$K(y) = \begin{cases} 1 - |y| & |y| \le 1 \\ 0 & |y| > 1 \end{cases}$$

3. $$K(y) = \begin{cases} (\frac{4}{3}) - 8y^2 + 8|y|^3 & |y| < \frac{1}{2} \\ \frac{8}{3}(1 - |y|)^3 & \frac{1}{2} \le |y| \le 1 \\ 0 & |y| > 1 \end{cases}$$

4. $$K(y) = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}y^2}$$

5. $$K(y) = \tfrac{1}{2}e^{-|y|}$$

6. $$K(y) = (\frac{1}{\pi})\ (1 + y^2)^{-1}$$

7. $$K(y) = \frac{1}{2\pi}\ (\frac{\sin\ (y/2)}{y/2})^2$$

where $r$ is a positive number, if it exists, such that $k_r = \lim_{n\to 0} \{[1 - k(u)]/|u|^r\}$ is not zero. Of course, this sequence also depends on the density which is unknown. However, pointwise consistency in quadratic mean does not depend on an optimal value.

M.S. Bartlett [2] discusses the optimization of the estimates in terms of both probability density and spectral density functions. For an interesting comparison, this work is recommended.

The results of Parzen have been extended in several directions. V.K. Murthy [19] considers estimates of form (3.1) where $K_n(x,y) = B_n \cdot K((x-y) \cdot B_n)$. This is the algebraic estimate with $B_n = h(n)^{-1}$. Murthy allows the distribution function to assign positive mass to a countable set of points. Hence the distribution may be decomposed into a pure step function and an absolutely continuous distribution. If $x_0$ is a continuity point of both the distribution and the density, and if the weighting function, $K$, and the sequence, $B_n = h(n)^{-1}$, satisfy the conditions of Parzen, then $\hat{f}_n(x_0)$ is consistent in quadratic mean. Furthermore, $\hat{f}_n(x_0)$ is asymptotically normal under these conditions.

É. A. Nadaraya [20] and [21] is among several authors who strengthen the conditions on the density to obtain stronger consistency results. Nadaraya considers estimates of the algebraic type. If, in addition, $f$ is uniformly continuous, $K$ is bounded variation and $\Sigma_{n=1}^{\infty} e^{-\gamma n h^2(n)}$ is finite for every $\gamma > 0$, then $\hat{f}_n$ is uniformly consistent (in $x$) with probability one. In addition, the mode of the estimate is a consistent (with probability one) estimate of the mode of $f$.

P.K. Bhattacharya [2] and E.F. Schuster [30] consider the algebraic estimates and their derivatives. Bhattacharya shows that if the density $f$ and its first $(r + 1)$ derivatives are bounded, then, if

$h(n) = n^{-1/2r+4}$ and $k_n = n^u$, $u > 0$ and $0 < c < \frac{1}{2r+4}$, $\lim_{n\to\infty} \sup_{|x| \le k_n} n^c |\hat{f}_n^{(r)}(x) - f^{(r)}(x)| = 0$ with probability one. He modifies the estimate to obtain uniform consistency with probability one. Schuster extends this work. If $h(n) \to 0$ and $h^*(n) \to \infty$ with $h(n)h^*(n) = o(1)$ and with $\Sigma_{n=1}^{\infty} \exp(-\gamma n[h(n)]^{2r+2}/[h^*(n)]^2)$ finite for every $\gamma > 0$, then $\lim_{n\to\infty} \sup_x h^*(n)|f_n^{(r)}(x) - f^{(r)}(x)| = 0$ with probability one whenever $f$ and its first $(r+1)$ derivatives are bounded. For this result, the weighting function, $K$, must be chosen as a probability density function with $\int_{-\infty}^{\infty} |u|K(u)du$ finite and such that $K^{(s)}$ is a continuous function of bounded variation for $s = 0,1,2,\ldots,r$. If we require the weighting function $K$ to be a density satisfying (3.8) as well as satisfying:

(3.9) $K$ is continuous and of bounded variation on $(-\infty, \infty)$.

(3.10) There is a $\delta$ in $(0,1)$ such that

$$u \cdot \left[ \underset{-\infty}{\overset{-(u^\delta)}{V}}(K) + \underset{u^\delta}{\overset{\infty}{V}}(K) \right] \to 0 \quad \text{as} \quad u \to \infty$$

where $\underset{a}{\overset{b}{V}}(K)$ is the variation of $K$ over $(a, b)$.

(3.11) $\int_{-\infty}^{\infty} |u|dK(u)$, the integral with respect to the signed measure determined by $K$, is finite.

and if $h(n)$ satisfies $\Sigma_{n=1}^{\infty} e^{(-cnh^2(n))} < \infty$, then a necessary and sufficient condition for $\lim_{n\to\infty} \sup_x |\hat{f}_n(x) - f(x)| = 0$ with probability one is that $f$ is uniformly continuous.

M. Woodroofe [41] considers more general kernal functions in estimates of the form (3.1). He lets $K_n(x,y) = \tau(n) \cdot G(x, \tau(n) \cdot (x-y))$ where $\tau(n) \to \infty$ and $\tau(n) = o(n)$ as $n \to \infty$. $G$ is a nonnegative, bounded, measurable function on the Euclidean plane such that

$$(3.12) \qquad \int G(x,y)dy = 1 \quad \text{for each} \quad x$$

$$(3.13) \qquad \sup_x \int_{|y| \geq t} |y| G(x,y)dy \to 0 \quad \text{as} \quad t \to \infty.$$

The true density $f$ is assumed bounded, and positive and continuous on some neighborhood of $[-1, 1]$. If $G$ satisfies a uniform Lipschitz condition of order $\beta$ on the Euclidean plane with $0 < \beta \leq 1$ and if $(\tau(n))^{\delta} = o(n)$ and $n = o((\tau(n))^{\delta})$ as $n \to \infty$ with $1 < \gamma < \delta$, then $\max_{|x| \leq 1} |\hat{f}_n(x) - f(x)| \to 0$ with probability one as $n \to \infty$. If, in addition, $f$ satisfies a uniform Lipschitz condition of order $\alpha$ with $\frac{1}{2} < \alpha \leq 1$ and if $\gamma > 2$, then $\max_{|x| \leq 1} |\hat{f}_n(x) - f(x)| = o(\tau^{\frac{1}{2}})$ with probability one as $n \to \infty$. In effect, the results are stated for densities with finite support. Woodroofe also considers limiting distributions.

Watson and Leadbetter [35] consider estimates of the form (3.1) with $K_n(x,y) = \delta_n(x-y)$. In [35], they consider the M.I.S.E. as a measure of goodness and find the $\delta_n$ which minimizes the M.I.S.E. In general, the optimal $\delta_n$ is a complicated expression which depends on the explicit form of the probability density function. Watson and Leadbetter show that the behaviour of the estimate depends on the behaviour of the characteristic function $\Phi_f$ corresponding to $f$. $\Phi_f$ is said to decrease algebraically of degree $p > 0$ if $\lim_{|t| \to \infty} |t|^p |\Phi_f(t)| = c > 0$. An estimator has algebraic form if $\delta_n(x) = h(n)^{-1} \cdot K(h(n)^{-1} \cdot x)$,

which is the type of estimate considered by Parzen. If $\Phi_f(t)$ is known to decrease algebraically of degree $p$, an estimator with algebraic form may be found whose M.I.S.E. $J_n$ satisfies $\lim_{n\to\infty}(J_n/J_n^*) = 1$ where $J_n^*$ is the M.I.S.E. corresponding to the optimal $\delta_n$. Hence, integratedly uniform consistency in quadratic mean follows when we know the characteristic function decreases algebraically. The characteristic function $\Phi_f$ decreases exponentially of coefficient $\rho > 0$ if $|\Phi_f(t)| \le Ae^{-\rho|t|}$ for some constant $A$ and all $t$, and if $\lim_{v\to\infty} \int_0^1 [1 + e^{2\rho v}|\Phi_f(vt)|^2]^{-1}dt = 0$. An estimate $\hat{f}_n$ formed from $\delta_n$ has exponential form if the characteristic function of $\delta_n$ is given by $\Phi_{\delta_n}(t) = \Phi(A_n e^{\alpha|t|})$ where $A_n \to 0$ as $n \to \infty$, $\alpha > 0$, and $\Phi$ is bounded and square integrable. As in the algebraic case, an exponential estimate may be found which has asymptotically the minimum M.I.S.E. Originally, [35] was intended to have a sequel. Unfortunately, the second paper was never published, but the results may be partially found in Leadbetter [17]. The unpublished paper dealt with the choice of $\delta_n$ with mean square error as the measure of goodness instead of M.I.S.E. See also the paper of Pickands [25] in connection with the optimal choice of an estimating sequence.

Whittle [40] approaches the problem in a unique manner. The form of Whittle's estimate is (3.1) with $K_n(x,y) = w_x(y)$. The sample size $n$ is assumed to be a Poisson variable with mean $M$. One may equally well estimate the unnormalized density $\phi = M \cdot f$. The form of the estimate then becomes

$$\hat{\phi}_n(x) = n \cdot \int_{-\infty}^{\infty} w_x(y)dF_n(y).$$

In the absence of any additional information on the densities for $\phi$,

Whittle assumes the ordinates, $\phi(y)$, have a prior distribution with mean $E_p[\phi(x)] = \mu(x)$ and second moments $E_p[\phi(x)\phi(y)] = \mu(x,y)$. Here $E_p$ is the expectation with respect to the prior distribution. The weighting function, $w_x$, is obtained by minimizing

$$\Delta^2 = E_p E_s[\hat{\phi}_n(x) - \phi(x)]^2 ,$$

where $E_s$ is the expectation over sampling variations. The weighting function which minimizes $\Delta^2$ is given as the solution of

$$\mu(x)w_x(y) + \int \mu(y,z)w_x(z)dz = \mu(y,x) .$$

If $\hat{f}_n$ is the normalized estimate, the mean square error

$$D^2 = E_p E_s[\hat{f}_n(x) - f(x)]^2$$

decreases to zero, but more slowly than $n^{-1}$, as $n \to \infty$. Hence, we have consistency in quadratic mean. It is of interest to note that actually the prior distribution need not be known, only the first two moments. Whittle remarks that $\mu(x)$ and $\mu(x,y)$ are $O(M)$ and $O(M^2)$ respectively as $M$ becomes large so that $w_x(y)$ approaches 1 if $x = y$ and 0 if $x \neq y$. Other properties of the weighting function $w_x$ may be found in Whittle's paper.

We close this section by calling attention to related papers. Pickands [25] considers the problem of estimating efficiently. Efficiency is defined as the limit as the sample size approaches $\infty$ of the ratio of the M.I.S.E. of a given estimate to the M.I.S.E. of an optimal estimate.

Cacoullos [3] extends the algebraic-type of estimate to the multidimensional case. Loftsgaarden and Quesenberry [18] give a variant on

the procedure of Cacoullos and Elkins [8] studies the analogues of the "naive" estimate in two dimensions.

Van Ryzin [33] considers the multidimensional estimates of Cacoullos and strong consistency properties. The consistency properties, of course, carry over in the one dimensional case and, particularly, the suggested weighting functions of Parzen (see Table 1) allow pointwise consistency with probability one.

Craswell [5] points out that this type of procedure may be extended to estimation in a topological group.

We should notice that the papers in this section have not been arranged to appear in historical order. The reader should check publication dates to judge priorities.

4. Estimation by Orthogonal Series and Maximum Likelihood. Estimates of the form (3.1) have received most of the attention in the literature. Several other apparently unrelated methods fall into a common theoretical background.

Let $X$ be a random variable and let $R_X$ be the range of $X$. Let $\lambda$ be a measure on $R_X$ and let $P$ be the probability distribution on $R_X$. Thus, $f = \frac{dP}{d\lambda}$ is the probability density function. Let $r$ be some function and define a scalar product by

$$(g, h) = \int_{R_X} g(x)h(x)r(x)d\lambda(x) .$$

We thus have a Hilbert space, $L_2(r)$. Let us assume $f \in L_2(r)$. In the special case $r \equiv 1$, we have the ordinary set of square integrable

functions, $L_2$. Let us further consider a subspace, E, spanned by the orthonormal basis, $\{g_k\}$, $k \in I$, where I is some index set. Finally, let us consider the projection of the density f onto E. This is given by

$$f^*(x) = \sum_{k \in I} (g_k, f) g_k(x), \tag{4.1}$$

where $(g_k, f) = \int_{R_X} g_k(x) f(x) r(x) d\lambda(x)$. Let $a_k = (g_k, f)$. To estimate $f^*$ (hence f), one must estimate $a_k$, $k \in I$. In general, of course, if $x_1, \ldots, x_n$ is a set of observations,

$$\hat{a}_k = \frac{1}{n} \sum_{j=1}^{n} g_k(x_j) r(x_j) \tag{4.2}$$

is a strongly consistent estimate of $a_k$. Thus

$$\hat{f}_n(x) = \sum_{k \in I} \hat{a}_k g_k(x) \tag{4.3}$$

is an estimate of the probability density f. This general description is due largely to Čencov [4]. Čencov considers cases where the index set I is finite and chooses $||\hat{f}_n - f||^2$ as error criterion. One must "select a sufficiently good approximating space E and then make a sufficiently large number of observations".

Several authors have selected approximating subspaces or equivalently orthonormal bases. It will be recalled from Section 1 that the Gram-Charlier approach was to use a series of the form (4.1), where the orthonormal basis was the sequence of Hermite functions.

Schwartz [31] considers the sequence of normalized Hermite functions

$$g_k(x) = (2^k k! \pi^{\frac{1}{2}})^{-\frac{1}{2}} \cdot e^{-x^2/2} \cdot H_k(x) \qquad k = 0, 1, \ldots ,$$

where

$$H_k(x) = (-1)^k e^{x^2} (d^k/dx^k)(e^{-x^2}) .$$

If $r(x) \equiv 1$, and the density, $f$, is continuous, bounded variation, and $f \in L_1 \cap L_2$, then $f(x) = \Sigma_{j=0}^{\infty} a_j g_j(x)$ where $a_j = \int f(x) g_j(x) d\lambda(x)$. Schwartz considers as an estimate of the density,

$$(4.4) \qquad \hat{f}_n(x) = \sum_{j=0}^{q(n)} \hat{a}_j g_j(x) ,$$

where $\hat{a}_j = \frac{1}{n} \Sigma_{i=1}^{n} g_j(x_i)$. If $q(n)$ is chosen so that $q(n)/n \to 0$ as $q(n) \to \infty$, the estimate is integratedly consistent in quadratic mean. Schwartz gives convergence rates and compares this estimate with those in [35] and [22].

Kronmal and Tarter [16], Tarter, Holcomb, and Kronmal [32], and Kowalski and Tarter [15] consider an estimate based on trigonometric functions rather than the Hermite function as in Schwartz. The form of their estimate is (4.4) where $\{g_k\}$ are chosen as one of the orthogonal systems $\{\sin \pi kx\}$, $\{\cos \pi kx\}$ or $\{e^{i\pi kx}\}$. Kronmal and Tarter show that if we have a density, $f$, with finite support, $\{x: f(x) > 0\}$, and if the density $f$ may be represented by a Fourier cosine series, then for a choice of $q(n) = o(\sqrt{n})$, the mean square error and the M.I.S.E. converge to zero. An interesting point of contact with estimates of form (3.1) exists. One may express the Fourier cosine estimates as algebraic estimates of form (3.1) where the weighting function is chosen as 7 in Table 1. Estimates based on either Fourier functions or Hermite functions have one serious flaw, however, which is that since the trigonometric series and the Hermite series may, in certain cases, take on negative values, it is possible that the density estimate become negative.

Another type of estimate which can be fitted into this general framework but which arises in a different context is the maximum likelihood type of estimate found in Robertson [27], Prakasa Rao [26], and Wegman [36], [37], and [38]. In general, maximum likelihood estimates for density functions do not exist, but, if some appropriate type of restriction is placed on the class of densities from which the estimate may be selected, then a maximum likelihood estimate over that class may exist. The restriction considered here is that the estimates are measurable with respect to a σ-lattice. A σ-lattice, $L$, is a collection of subsets (of $R_X$) which is closed under countable unions and intersections and contains both the whole set, $R_X$, and the empty set. A function is measurable with respect to a σ-lattice, $L$, if $\{x: f(x) > a\} \in L$ for all real $a$. Under nominal restrictions, the set of densities which are measurable with respect to a σ-lattice will form a closed convex cone in $L_2$. Hence just as we projected the density onto the subspace spanned by the orthogonal functions, we may also project the density onto the cone. This projection is called a conditional expectation with respect to the σ-lattice. See Brunk [43] for a detailed treatment. Robertson [27] considers estimates which are measurable with respect to a σ-lattice, $L$. An interesting case, considered by Robertson, is the case of a unimodal density with known mode. A unimodal density with known mode may be characterized as measurable with respect to the σ-lattice, $L$, of intervals containing the mode. (It should be noted that a unimodal density in this sense is monotone on either side of the mode. This definition is somewhat different from that of Parzen [22] and others.) The estimate of Robertson has the form

$$\hat{f}_n(x) = \sum_{j=1}^{k} \frac{n_j}{n} \cdot \frac{I_{A_j}(x)}{\lambda(A_j)} .$$

Here $A_j$ is an interval determined from the lattice, $L$, and the particular set of observations, $x_1, \ldots, x_n$. The function $I_{A_j}$ is the indicator of $A_j$ and $n_j$ is the number of observations in $A_j$. Notice by letting $g_j = I_{A_j}/\lambda(A_j)$, we may define an orthonormal basis and $n_j/n$ is the estimate of $\int_{-\infty}^{\infty} g_j(x) \cdot f(x)d\lambda(x) = P(A_j)$ given in (4.2). Robertson shows pointwise consistency with probability one and under the assumption of continuity on $f$, he shows almost uniform consistency with probability one. The maximum likelihood estimate of the unimodal density is due to Grenander [12].

Prakasa Rao [26] considers the same case and derives an asymptotic distribution theory. Wegman [36] considers a related estimate with unknown mode estimated by strongly consistent estimates such as that of Venter [34]. Wegman also obtains pointwise consistency with probability one and under the assumption of continuity on $f$, he obtains almost uniform consistency with probability one. Robertson, Cryer and Hogg [28] also construct an estimate based on an estimate of the mode. A peculiar feature of the estimates in [27] and [36] is a peaking near the mode. Robertson, Cryer and Hogg [28] attempt to overcome this by smoothing their estimate near the mode. Wegman [37] obtains a maximum likelihood estimate with unknown mode which also allows a smoothing near the mode. Under assumptions of continuity on $f$, Wegman obtains almost uniform convergence with probability one to a density which is closely related to $f$. In particular, the density is identical with $f$ except on an interval of length $\epsilon > 0$, which is an arbitrarily small, predetermined number. One may also show that the estimates in [36] and [37] have asymptotic distributions as in Prakasa Rao [26].

A discussion of density estimation would not be complete without the venerable histogram. Wegman [38] points out that the histogram is measurable with respect to a fixed $\sigma$-algebra with a finite number of atoms (a special case of a $\sigma$-lattice). The histogram is a maximum likelihood estimate over the set of functions measurable with respect to the $\sigma$-algebra. Wegman generalizes the notion of histogram to the case where the $\sigma$-algebra is not fixed, but determined by the observations. He shows that these histograms enjoy a certain type of consistency (with probability one).

Again, we point out that the papers referred to in this section are not necessarily presented in historical order. Let us also point out a paper by Weiss and Wolfowitz [39] which is concerned not with estimating a density but estimating the density evaluated at a point.

5. Monte Carlo Simulation. In this section, we make use of Monte Carlo methods to study some of the density estimates. The multiplicative congruential method (see [44], pp.27-29) was used to obtain pseudorandom number distributed uniformly on $(0, 1)$. There are some difficulties with this method (see [45]), but they appear not to affect the present work. Pseudorandom numbers with a triangular distribution were obtained by adding two uniform pseudorandom numbers and dividing by two. Normal pseudorandom numbers were obtained approximately by adding 12 uniform pseudorandom numbers and subtracting 6. This results in a standard normal distribution approximately by virtue of the central limit theorem. Cauchy pseudorandom numbers were obtained by taking the ratio of two

standard normal pseudorandom numbers. Finally, exponential pseudorandom numbers were generated by taking the negative natural logarithm of uniform pseudorandom numbers.

As measures of goodness, we have two alternatives. In connection with the maximum likelihood estimates, it seemed natural to consider the likelihood product. The likelihood product, however, tends to be a quite small positive number, so that we actually considered $(\frac{1}{n}) \cdot \log$ (Likelihood Product), where $n$ is the sample size and the log is the natural logarithm. Another frequently used criterion is M.I.S.E. In view of the relative complexity of numerically obtaining the integrated square error and of the availability of the density estimate at the observations, we decided to form the average square error,

$$(5.1) \qquad \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_n(x_i) - f(x_i))^2$$

at the observations. This approximates the integrated square error with weight function, f. Since this error measure tends to put heavier weight on the "central" observations, it discriminates somewhat against the types of estimates found in [27] and [36] which tend to be quite inaccurate near the mode. (5.1) may be viewed as an integral with respect to the empirical distribution function. The discrete nature of (5.1) also tends to inflate the importance of "peaks" near the modes in terms of an estimate of mean square error since there may be quite a few observations in a very small interval near the mode. The reader should be forewarned when he views Tables 2 and 3. For each type of estimate, underlying density, and sample size, the estimate was formed for 25 different samples, except for the sample size of 500 where only 10 different samples seemed necessary. The average values of the average square error and of $(\frac{1}{n})\log$ (Likelihood Product) were formed. In view of the comments

on the estimators found in [27] and [36], the median values of the average square error are given also (in parenthesis). The various estimates were all formed from the same samples, so, for example, in Table 2, the error .3036 for the histogram is based on the same 25 samples of size 50 as the error .0531 for the naive estimate.

Six types of estimates were chosen. As a well-known standard against which others could be compared, we chose the histogram with 10 intervals. A variant of the histogram whose intervals are not of equal length but based on the percentiles is labelled Histogram II. A treatment of this estimate is found in [38]. The maximum likelihood unimodal estimate with known mode and a unimodal estimate with unknown mode found in [27] and [36], respectively, are representatives of the maximum likelihood procedures. As representative of the kernal estimates, the naive estimate was chosen. This is an algebraic estimate with weighting function 1 from Table 1 and is the estimate suggested by Rosenblatt [29]. Originally, other weighting functions were to be used, but the naive estimate did so well in terms of average square error that it was felt unnecessary to investigate any others. For this study, $h(n)$ was chosen as $.3118 \cdot n^{-1/5}$. The coefficient .3118 was arrived at after a small amount of experimentation with samples from a triangular density. Finally, as representative of the orthogonal series estimates, the cosine-based trigonometric estimate was used. This is the Fourier cosine estimate found in [16]. In the notation of Section 4, $q(n)$ is chosen as $n^{1/3}$. We may add that contrary to the experience of the authors of [16], negative values of the density estimate were encountered.

Tables 2 through 5 summarize the results. As might be expected, the representatives of the maximum likelihood procedures generally did well

in terms of likelihood product but fared poorly in terms of average square error when compared to the naive and the trigonometric estimates. The histograms seemed to occupy an intermediate position. It is interesting to note how favorably the standard histogram compares to other types of estimates.

Finally, the appearance of several estimates is illustrated by Figures 1 through 8. In addition to the six estimates found in Tables 2 through 5, we also include the maximum likelihood histogram as described in [38] and the Hermetian estimate as described in [31]. In Tables 2 through 5 and in Figures 4 and 5, unimodal estimate I is the maximum likelihood unimodal estimate with known mode and unimodal estimate II is the unimodal estimate with unknown mode estimated by the Venter estimate (see [27], [36] and [34]). The underlying density is the standard normal and each estimate is based on the same 100 observations.

Table 2: Average Square Error for Triangular Density.

| Number of Observations \ Type of Estimate | Histogram | Histogram II | Unimodal* Estimate I | Unimodal** Estimate II | Naive Estimate | Trigonometric Estimate |
|---|---|---|---|---|---|---|
| 50 | .3036 | .7578 | 2.2321 (1.2400) | 1.0138 (1.0070) | .0531 | .0559 |
| 100 | .1349 | .2743 | .7323 (.4521) | .7458 (.6088) | .0322 | .0363 |
| 175 | .0836 | .1297 | .7828 (.2979) | .6999 (.4350) | .0228 | .0369 |
| 250 | .0600 | .0953 | .7197 (.3700) | .3608 (.2972) | .0205 | .0262 |
| 500 | .0439 | .0536 | .4100 (.1985) | .1348 (.1248) | .0083 | .0132 |

* In this Table as well as Tables 3, 4, 5, Unimodal Estimate I is the Maximum Likelihood Estimate with Known Mode.

** In this Table as well as Tables 3, 4, 5, Unimodal Estimate II is the Unimodal Estimate with mode estimated by the Venter estimate.

Table 3: Average Square Error for several densities with a sample of size 100 from each.

| Under-lying Density \ Type of Estimate | Histogram | Histogram II | Unimodal Estimate I | Unimodal Estimate II | Naive Estimate | Trigonometric Estimate |
|---|---|---|---|---|---|---|
| Uniform on (0,1) | .1210 | .1544 | .3615* (.1213) | .5382* (.4301) | .0439 | .0700 |
| Triangular | .1349 | .2743 | .7323 (.4521) | .7458 (.6088) | .0322 | .0363 |
| Exponential | .0142 | .0474 | ** | .7191 (.1223) | .0238 | .0156*** |
| Standard Normal | .0057 | .0148 | .1622 (.0274) | .0767 (.0298) | .0124 | .0009*** |
| Cauchy with median 0 | .0209 | .0070 | .0857 (.0182) | .0599 (.0201) | .0075 | .0288*** |

* Mode is .5 for Unimodal Estimate I; there are no convergence properties for Venter Estimate in this case.

** Unimodal density program not applicable to this case.

*** Negative values of estimate occurred.

Table 4: $(\frac{1}{n}) \cdot \log$ (Likelihood Product) for Triangular Density.

| Number of Observations \ Type of Estimate | Histogram | Histogram II | Unimodal Estimate I | Unimodal Estimate II | Naive Estimate | Trigonometric Estimate |
|---|---|---|---|---|---|---|
| 50 | .3373 | .3454 | .4050 | .3896 | .1841 | .2522 |
| 100 | .2623 | .2239 | .3083 | .3496 | .1919 | .2434 |
| 175 | .2335 | .2002 | .2855 | .3024 | .1870 | .2276 |
| 250 | .2211 | .2073 | .2612 | .2933 | .1932 | .2119 |
| 500 | .2090 | .1856 | .2353 | .2340 | .1948 | .2299 |

Table 5: $(\frac{1}{n})$ · log (Likelihood Product) for several densities with a sample of size 100 from each.

| Type of Estimate / Underlying Density | Histogram | Histogram II | Unimodal Estimate I | Unimodal Estimate II | Naive Estimate | Trigonometric Estimate |
|---|---|---|---|---|---|---|
| Uniform on (0,1) | .0674 | .0672 | .0514* | .0610* | - .0693 | .0120 |
| Triangular | .2623 | .2239 | .3083 | .3496 | .1919 | .2434 |
| Exponential | - .9969 | -1.0172 | ** | - .8781 | - .9403 | *** |
| Standard Normal | -1.3463 | -1.3727 | -1.2086 | -1.2595 | -1.2930 | *** |
| Cauchy with median 0 | -3.3095 | -2.6301 | -2.2663 | -2.2664 | -1.8980 | *** |

*
**
***
} see corresponding notes in Table 3.

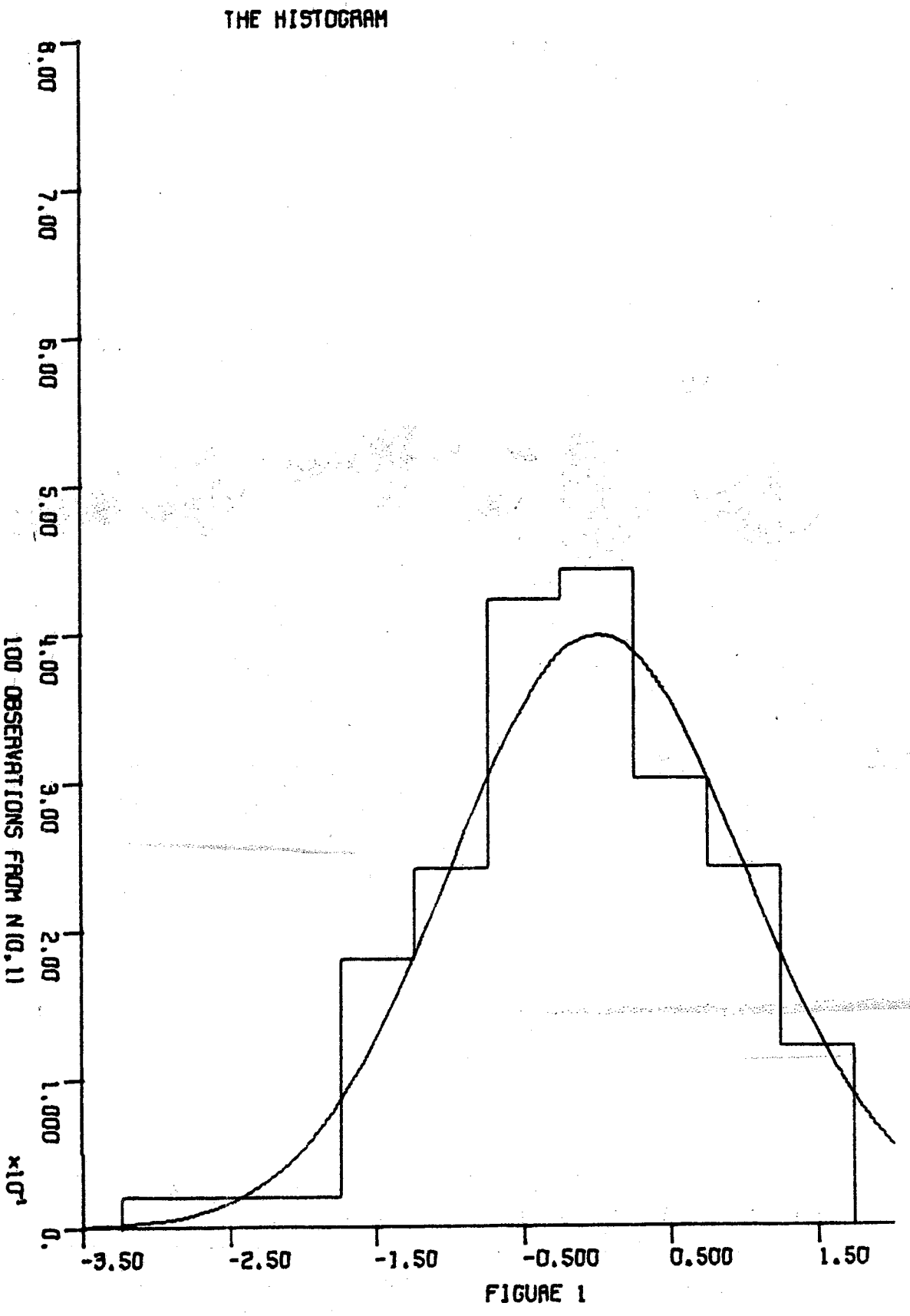
THE HISTOGRAM
8.00
7.00
6.00
5.00
4.00
3.00
2.00
1.000
0.
100 OBSERVATIONS FROM N(0,1)
×10⁻²
-3.50
-2.50
-1.50
-0.500
0.500
1.50

FIGURE 1

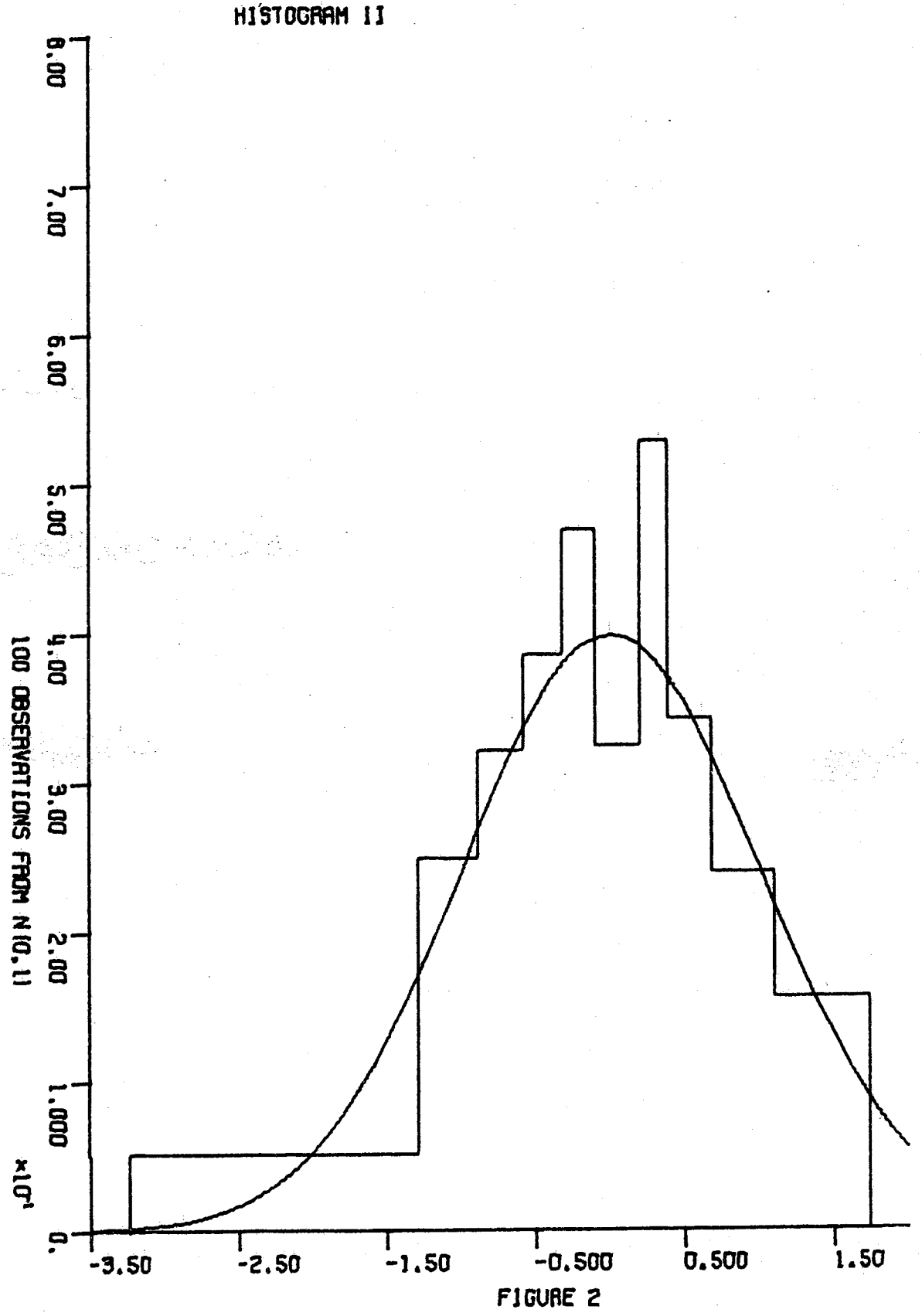
HISTOGRAM 11
8.00
7.00
6.00
5.00
4.00
3.00
2.00
1.000
0.
100 OBSERVATIONS FROM N(0,1)
×10¹
-3.50
-2.50
-1.50
-0.500
0.500
1.50

FIGURE 2

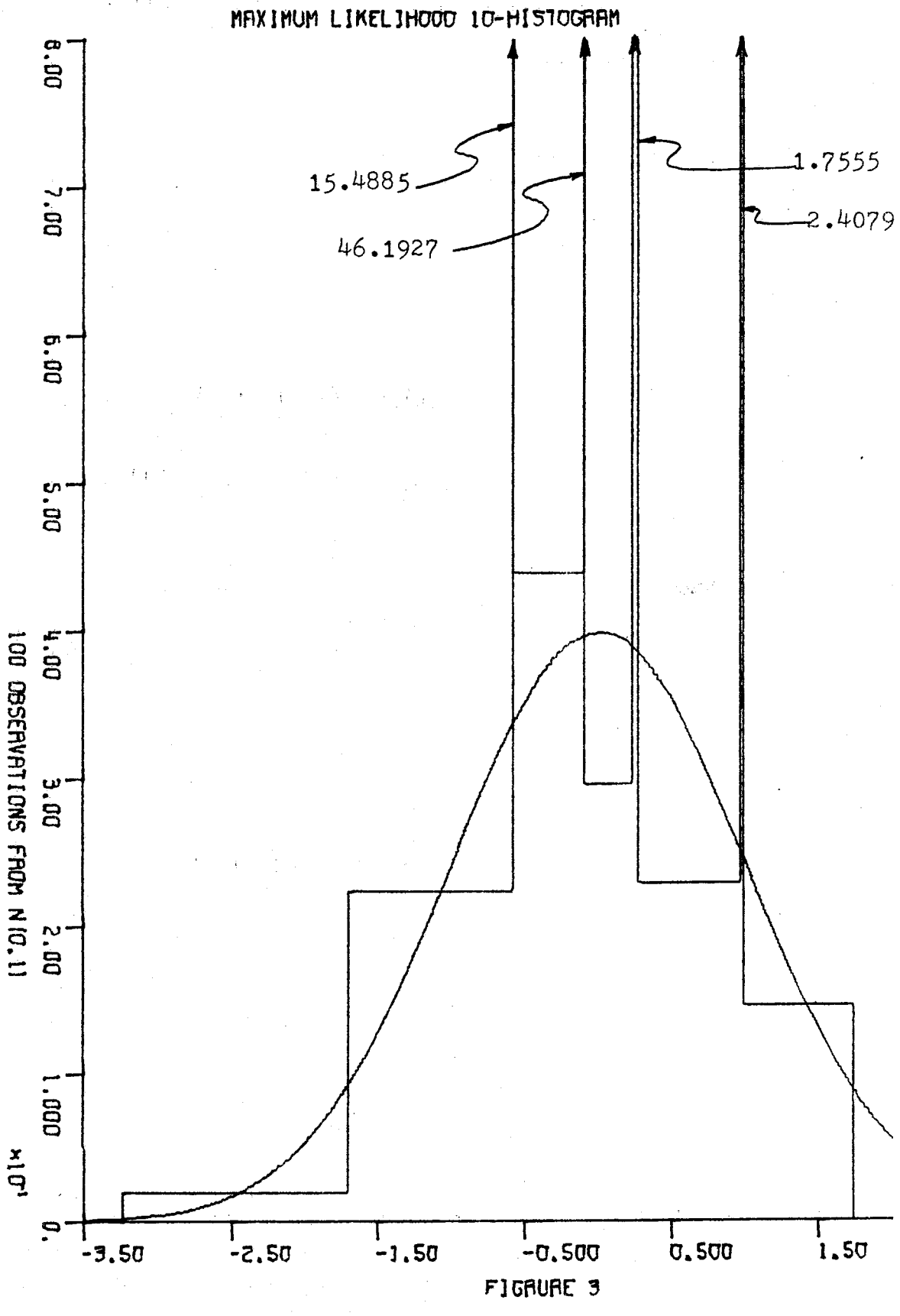
MAXIMUM LIKELIHOOD 10-HISTOGRAM
15.4885
46.1927
1.7555
2.4079
100 OBSERVATIONS FROM N(0,1)
×10¹
0.
1.000
2.00
3.00
4.00
5.00
6.00
7.00
8.00
-3.50
-2.50
-1.50
-0.500
0.500
1.50

FIGURE 3

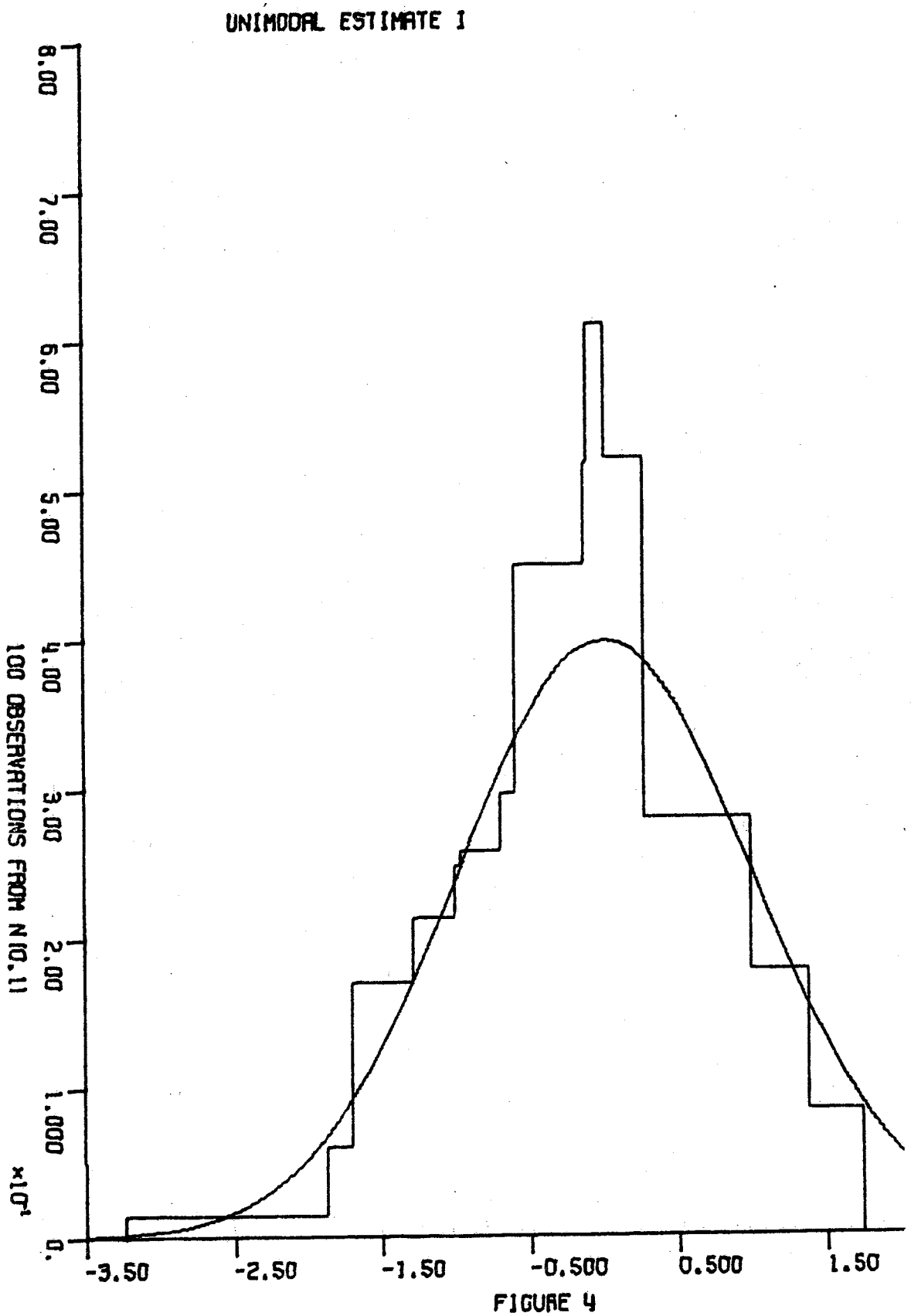
UNIMODAL ESTIMATE I
8.00
7.00
6.00
5.00
4.00
3.00
2.00
1.000
0.
×10⁻¹
100 OBSERVATIONS FROM N(0,1)
-3.50
-2.50
-1.50
-0.500
0.500
1.50

FIGURE 4

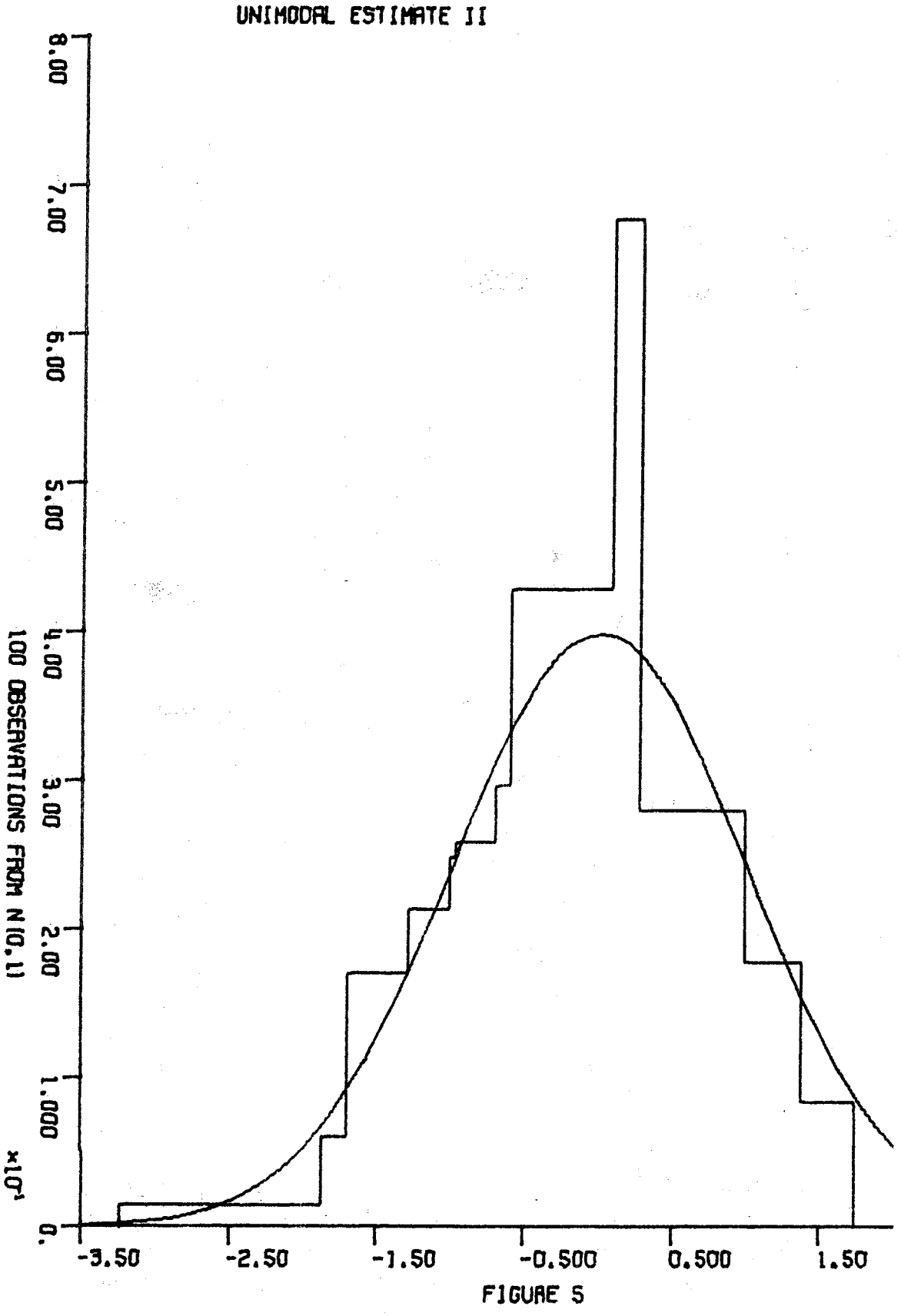
UNIMODAL ESTIMATE II
8.00
7.00
6.00
5.00
4.00
3.00
2.00
1.000
0.
×10⁻¹
100 OBSERVATIONS FROM N(0,1)
-3.50
-2.50
-1.50
-0.500
0.500
1.50

FIGURE 5

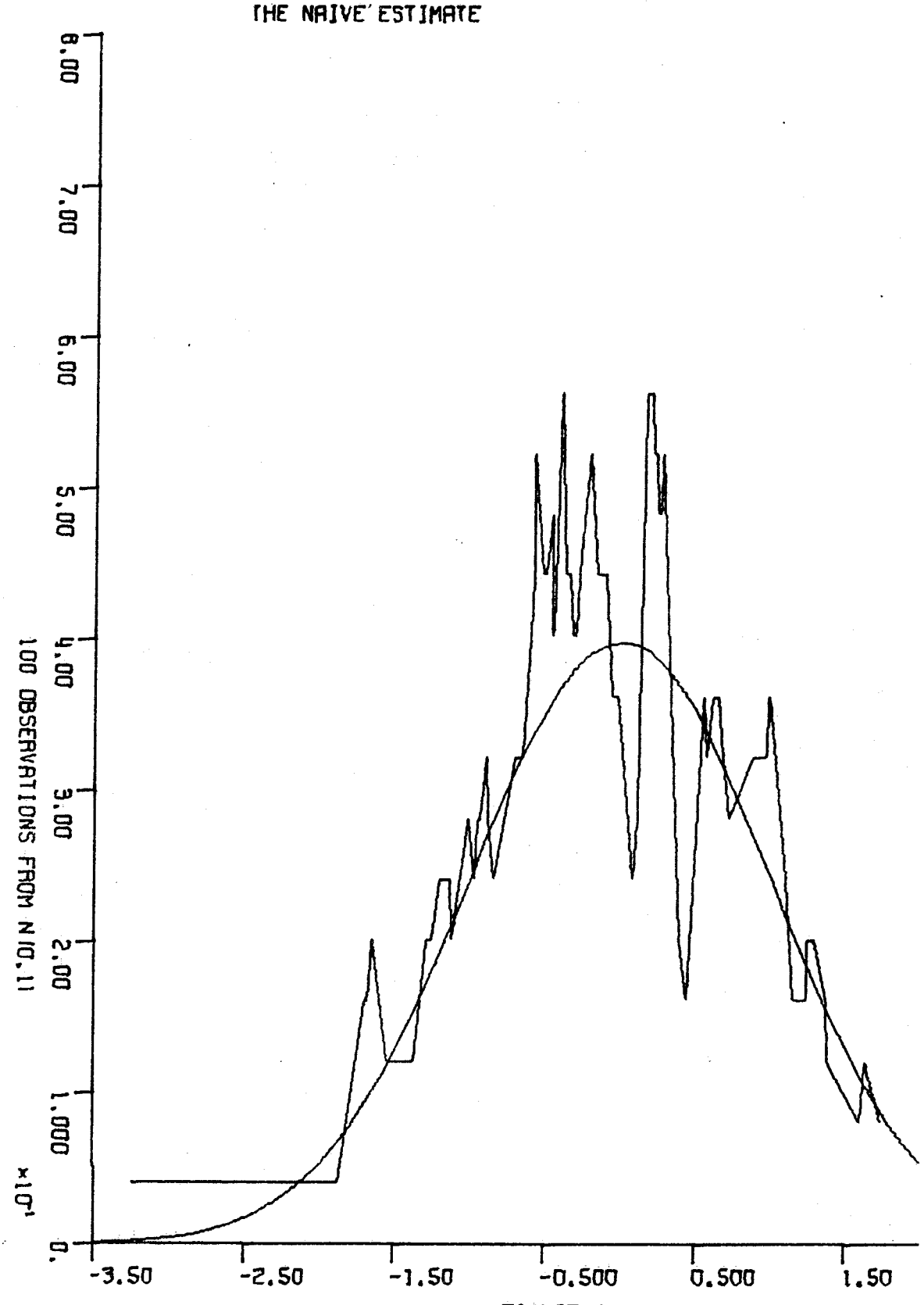
THE NAIVE ESTIMATE
100 OBSERVATIONS FROM N(0,1)
×10⁻¹
8.00
7.00
6.00
5.00
4.00
3.00
2.00
1.000
0.
-3.50
-2.50
-1.50
-0.500
0.500
1.50

FIGURE 6

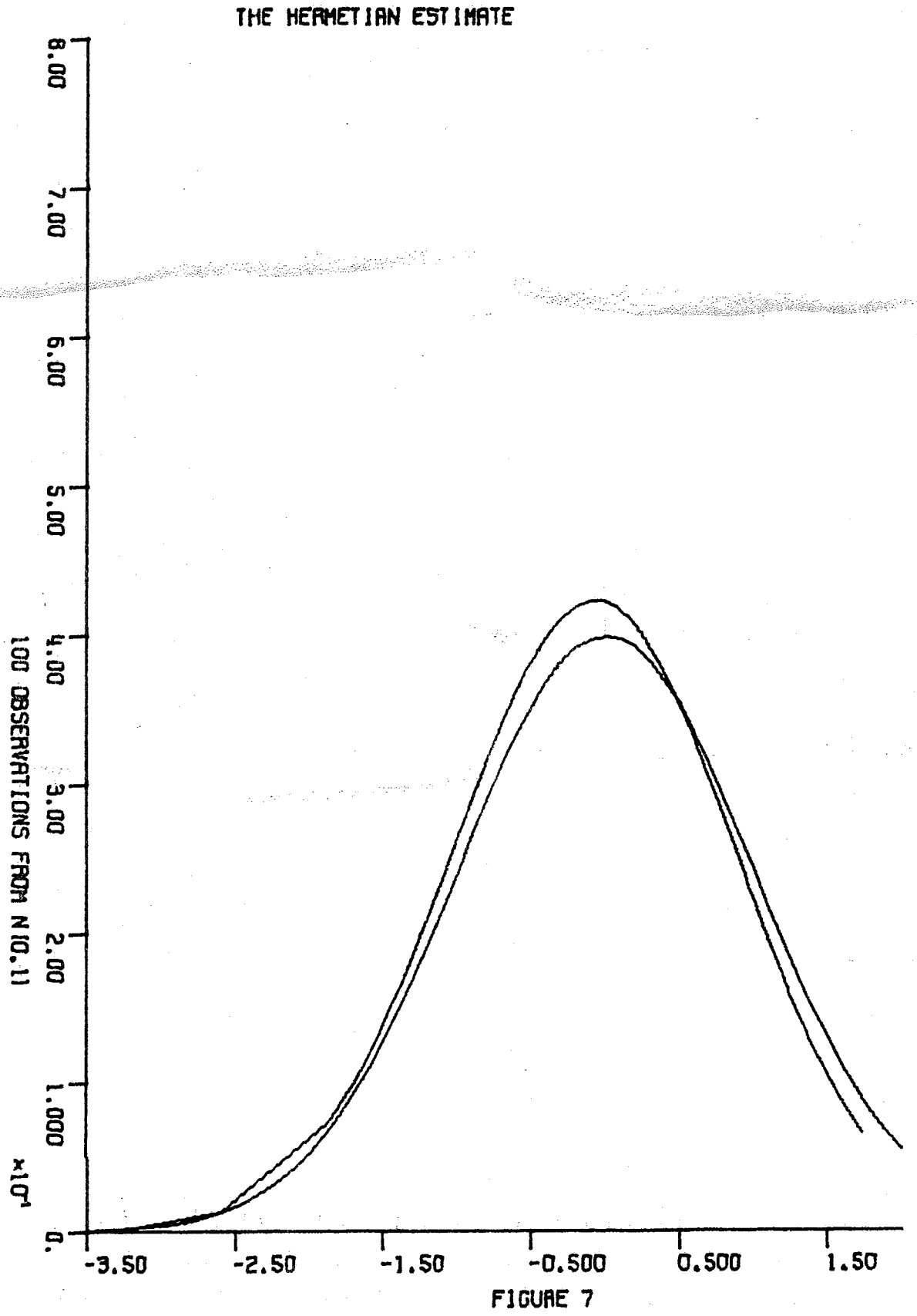
THE HERMETIAN ESTIMATE
8.00
7.00
6.00
5.00
4.00
3.00
2.00
1.000
0.
100 OBSERVATIONS FROM N(0,1)
×10⁴
-3.50
-2.50
-1.50
-0.500
0.500
1.50

FIGURE 7

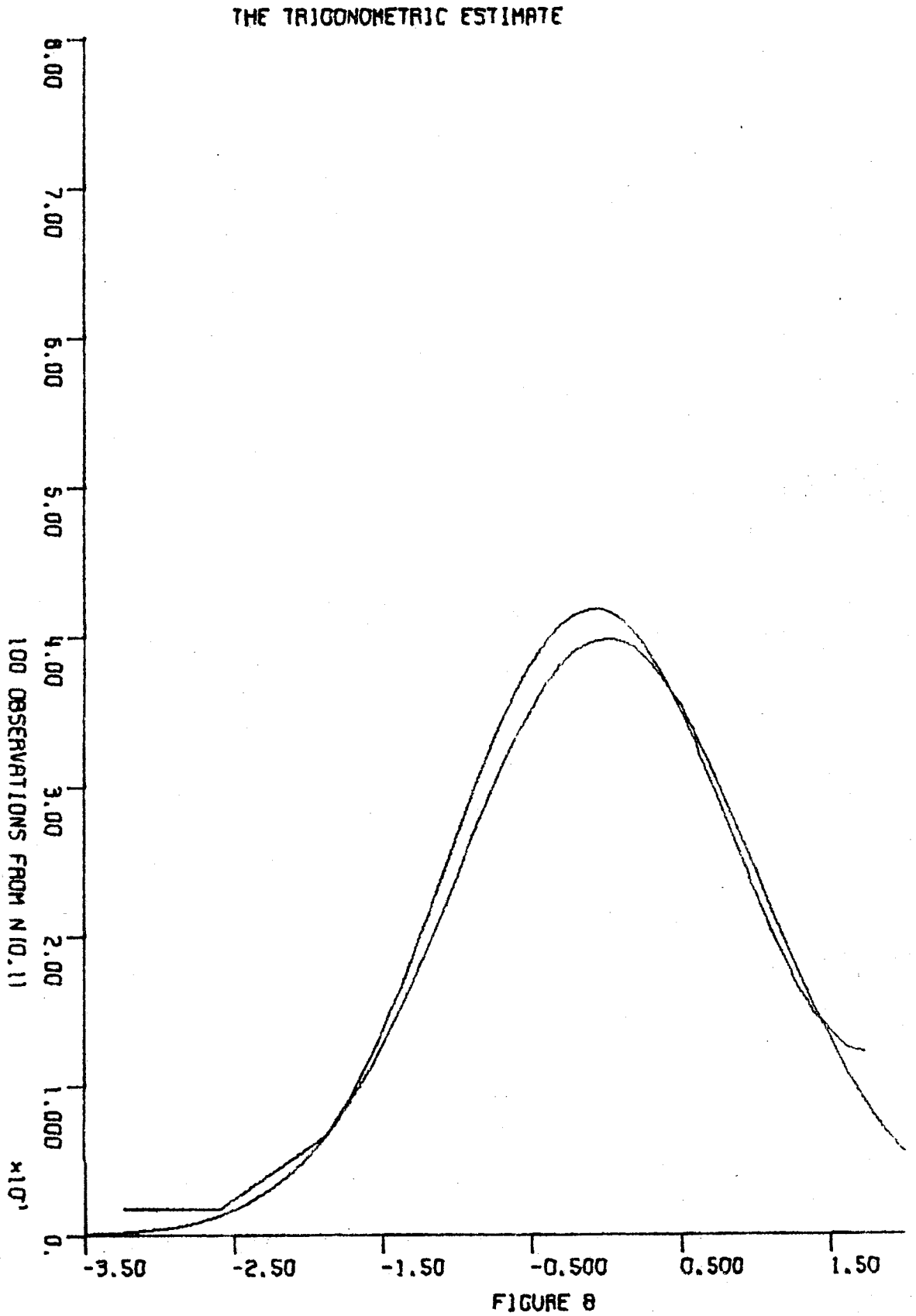
THE TRIGONOMETRIC ESTIMATE
8.00
7.00
6.00
5.00
4.00
3.00
2.00
1.000
0.
100 OBSERVATIONS FROM N(0,1)
x10'
-3.50
-2.50
-1.50
-0.500
0.500
1.50

FIGURE 8

6. Acknowledgment. The author thanks Mr. Joseph Austin, Jr., who assisted with the programming chores, especially with the calculation of the maximum likelihood histogram and with the computer drawn figures.

7. Bibliography on Probability Density Estimation.

[1] Bartlett, M.S. (1963), "Statistical estimation of density functions," Sankhyā (A), 25, pp. 245-254.

[2] Bhattacharya, P.K. (1967), "Estimation of a probability density function and its derivatives," Sankhyā (A), 29, part 4, pp. 373-382.

[3] Cacoullos, Theophiles (1966), "Estimation of a multivariate density," Annals of the Institute of Statistical Mathematics, 18, pp. 178-189.

[4] Čencov, N.N. (1962), "Evaluation of an unknown distribution density from observations," Soviet Math., 3, pp. 1559-1562.

[5] Craswell, W.J. (1965), "Density estimation in a topological group," Ann. Math. Statist., 36, pp. 1047-1048.

[6] Edgeworth, F.Y. (1904), "The law of error," Trans. Camb. Phil. Soc., 20, pp. 36 and 113.

[7] Elderton, W.P. (1938), Frequency Curves and Correlation, 3rd edition, Cambridge University Press.

[8] Elkins, T.A. (1968), "Cubical and Spherical estimation of a multivariate probability density," JASA, 63, pp. 1495-1513.

[9] Farrell, R.H. (1967), "On the lack of a uniformly consistent sequence of estimators of a density function in certain cases," Ann. Math. Statist., 38, pp. 471-474.

[10] Fisher, R.A. (1912), "On an absolute criterion for fitting frequency curves," Mess. of Math., 41, pp. 155-160.

[11] Fix, Evelyn and Hodges, J.L., Jr. (1951), "Nonparametric discrimination: consistency properties." Report Number 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, February, 1951.

[12] Grenander, Ulf (1956), "On the theory of mortality measurement, Part II," Skan. Ahtuarietidskr., 39, pp. 125-153.

[13] Kendall, M.G. (1948), The Advanced Theory of Statistics, Vol. 1, Charles Griffin and Co., Ltd., London.

[14] Kendall, M.G. and Stewart, A. (1958), The Advanced Theory of Statistics, Vol. 1, Charles Griffin and Co., Ltd., London.

[15] Kowalski, C. and Tarter, M. (1968), "On the simultaneous estimation of density and distribution functions with application to C-type density estimation," an unpublished manuscript.

[16] Kronmal, R. and Tarter, M. (1968), "The estimation of probability densities and cumulatives by Fourier Series methods," JASA, 38, pp. 482-493.

[17] Leadbetter, M.R. (1963), "On the non-parametric estimation of probability densities," Technical Report No. 11, Research Triangle Institute. (Doctoral dissertation at the University of North Carolina at Chapel Hill.)

[18] Loftsgaarden, D.O. and Quensenberry, C.P. (1965), "A non-parametric estimate of a nultivariate density function," Ann. Math. Statist., 38, pp. 1261-1265.

[19] Murthy, V.K. (1965), "Estimation of probability density," Ann. Math. Statist., 36, pp. 1027-1031.

[20] Nadaraya, É.A. (1963), "On estimation of density functions of random variables," Soobsch. Akad. Nauk. Grugin SSR, XXXII, 2, pp. 277-280. (In Russian.)

[21] Nadaraya, É.A. (1965), "On non-parametric estimates of density functions and regression curves," Theory Prob. Appl.,10, pp. 186-190.

[22] Parzen, E. (1962), "On the estimation of a probability density function and the mode." Ann. Math. Statist., 33, pp. 1065 - 1076.

[23] Pearson, K. (1902), "On the systematic fitting of curves to observations and measurements, I," Biometrika, 1, pp. 265 - 303.

[24] Pearson, K. (1902), "On the systematic fitting of curves to observations and measurements, II," Biometrika, 2, pp. 1-23.

[25] Pickands, J. (1969), "Efficient estimation of a probability density function," Ann. Math. Statist., 40, pp. 854-864.

[26] Rao, B.L.S.P. (1967), "Estimation of a unimodal density," an unpublished manuscript.

[27] Robertson, T. (1967), "On estimating a density which is measurable with respect to a σ-lattice," Ann. Math. Statist., 38, pp. 482-493.

[28] Robertson, T., Cryer, J.D., and Hogg, R.V. (1968), "On non-parametric estimation of distributions and their modes," an unpublished manuscript.

[29] Rosenblatt, M. (1956), "Remarks on some nonparametric estimates of a density function," Ann. Math. Statist., 27, pp. 832-837.

[30] Schuster, E.F. (1969), "Estimation of a probability density function and its derivatives," to appear Ann. Math. Statist., August, 1969.

[31] Schwartz, S.C. (1967), "Estimation of a probability density by an orthogonal series," Ann. Math. Statist., 38, pp. 1261 - 1265.

[32] Tarter, M.E., Holcomb, R.L., and Kronmal, R.A. (1967), "A description of new computer methods for estimating the population density," Proc. A.C.M., Thompson Book Company, 22, pp. 511-519.

[33] Van Ryzin, J. (1969), "On strong consistency of density estimates," an unpublished manuscript. (Abstracted, Ann. Math. Statist., 40, p. 1150.)

[34] Venter, J.H. (1967), "On estimation of the mode," Ann. Math. Statist., 38, pp. 1446-1455.

[35] Watson, G.S. and Leadbetter, M.R. (1963), "On estimating a probability density, I," Ann. Math. Statist., 34, pp. 480-491.

[36] Wegman, E.J. (1968), "A note on estimating a unimodal density function," Institute of Statistics Mimeo Series # 599, University of North Carolina at Chapel Hill. (To appear Ann. Math. Statist., October, 1969.)

[37] Wegman, E.J. (1969), "Maximum likelihood estimation of a unimodal density function," Institute of Statistics Mimeo Series # 608, University of North Carolina at Chapel Hill.

[38] Wegman, E.J. (1969), "Maximum likelihood histograms," Institute of Statistics Mimeo Series # 629, University of North Carolina at Chapel Hill.

[39] Weiss, L. and Wolfowitz, J. (1967), "Estimation of a density at a point," Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 7, pp. 327-335.

[40] Whittle, P. (1958), "On smoothing of probability density functions," JRSS (B), 20, pp. 334-343.

[41] Woodroofe, M. (1967), "On the maximum deviation of the sample density," Ann. Math. Statist., 38, pp. 475-481.

[42] Woodroofe, M. (1968), "On choosing a delta sequence," Technical Report No. 10, Department of Statistics, Carnegie-Mellon University, Pittsburgh, Pennsylvania. (Abstracted, Ann. Math. Statist., 39, p. 700.)

8. References.

[43] Brunk, H.D. (1965), "Conditional expectation given a σ-lattice and applications," Ann. Math. Statist., 36, pp. 1339-1350.

[44] Hammersley, J.M. and Handscomb, D.C. (1964), Monte Carlo Methods, Methuens' Statistical Monographs, John Wiley and Sons.

[45] Marsaglia, G. (1968), "Random numbers fall mainly in the planes," Proc. N.A.S., 61, pp. 25-28.