ESTIMATION OF A DENSITY FUNCTION WITH APPLICATIONS TO RELIABILITY

by

Thomas W. Jones

Dissertation submitted to the Graduate Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY
in
Statistics

APPROVED:

A. D. Hendrickson, Chairman	J. C. Arnold
R. H. Myers	G. K. Bennett
R. G. Krutchkoff	

May,

Blacksburg, Virginia

ACKNOWLEDGMENTS

I would like to express my appreciation:

to various members of the Department of Industrial Engineering and Operations Research who added greatly to my professional development and overall enjoyment of the industrial engineering and operations research fields. I would like to mention one person in particular, Dr. G. Kemble Bennett.

to the members of my committee who gave their time and effort in order that I may complete this study. Most notable of these members is my major professor, Dr. Arlo D. Hendrickson.

to Dr. Boyd Harshbarger for his constant encouragement and for the financial assistance that was given to me.

to the Center for Demographic Studies at Duke University where I was employed during the final months of preparation.

and finally to my wife, for just being my wife and for the drive she instilled in me, and also to my son.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The purpose of this dissertation is to examine the problem of estimation of a univariate probability density function. Let $Y_1, Y_2, \ldots, Y_n$ be a sample of n independent observations, each distributed according to an unknown continuous density function f(y). Given this sequence of observations, how can one estimate f(y)? Let the estimator be denoted by $f_n(y)$.

Several estimators of a univariate density function have been proposed in recent years. Estimators of the conditional probability density and of the conditional mean or regression line have also been investigated. While most of the results have been concerned exclusively with independent observations, a few have been developed when the observations are dependent. Many of these papers will be discussed briefly as we proceed.

Historically, Rosenblatt (1956) is credited with formulation of the problem of estimating a density function. However, it should be noted that estimation of the spectral density function when sampling a stationary sequence was developed before that of probability density estimation. The results obtained in both areas are similar but are much more simplified for the probability functions. Except for the last ten years, little research of this type considered

probability density estimation as indicated by a quote from Watson and Leadbetter (1963). "Many authors have considered the problems of estimating the spectral density of a stationary time series from observations of the series throughout a time T. The corresponding problem of estimating probability densities has received less attention in the literature."

Not long after Rosenblatt's article, the Royal Statistical Society's Symposium on the Spectral Approach to Time Series (1957) delved into the subject of estimating a spectral density. Whittle's (1957) contributed paper seemed to deal more with estimating a probability density. The symposium led to future papers by Whittle (1958), Daniels (1962), and Priestly (1962), while Rosenblatt's article led to papers by Parzen (1962) and Nadaraya (1963).

Whittle (1957, 1958) discusses the difficulty in constructing smooth curves. He introduces a Bayesian argument in order to accomplish the desired smoothing and to find an optimal weighting function.

The work originated by Rosenblatt and then extended by Parzen will be presented in a separate section since their papers have played such an important role in the many articles written on density estimation. Bartlett (1963) improves upon the order of the mean square error for the results proposed by Rosenblatt. He also considers spectral densities. Watson and Leadbetter (1963) use estimators of the same form

as Rosenblatt's kernel estimator. They minimize the mean integrated square error, which is considered to be a global error, as opposed to the local mean square error. In the paper, they showed that the optimal $f_n(y)$ would be obtained by inverting an expression involving the characteristic function of the true density $f(y)$. Loftsgaarden and Quesenberry (1965) estimate a d-dimensional density function by specifying the number of points which should lie within the radius of a hypersphere and then determining the minimum radius. This approach is from the opposite direction, since the usual process is to count the number of points in a hypersphere of a given radius. Cacoullos (1964, 1966) adapted Parzen's univariate estimator to provide estimators of a multivariate density. His results for asymptotic unbiasedness, consistency, and bounds for bias and mean square error of the estimate are straightforward extensions of the univariate results. Murthy (1965a) relaxed Parzen's assumption of an absolutely continuous distribution $F(y)$, and showed that the estimators still consistently estimate the true density at all points which are continuity points of both $F(y)$ and $f(y)$. He also proves the sequence of estimators to be asymptotically normally distributed. Murthy (1966) extended his own results to the multivariate case. Nadaraya (1965), using Parzen's kernel estimator, investigated the asymptotic behavior of the maximum deviation

of the estimator $f_n(y)$ from $f(y)$. This was also considered by Woodroofe (1967). Other articles were published by Nadaraya (1963, 1964, 1966, 1970) on the estimation of: i) a univariate density, ii) a bivariate density, and iii) a regression curve. Curiously, as it was noted by Gaskins (1972), Nadaraya (1964a) gives a subset of the Parzen paper without reference to Parzen.

Craswell (1965) proves that $f_n(y)$ and $f_n(z)$ are independent and normally distributed as $n \to \infty$. under certain regularity conditions. Van Ryzin (1966) introduces a Bayesian analysis for classification procedures and examines asymptotic properties based on the methods of Cencov (1962) and of Parzen (1962). Elkins (1968) forms two estimators of a multivariate density function by counting how many sample points lie within a d-dimensional cube and by counting how many sample points lie within a d-dimensional sphere. Elkins' method is a multi-dimensional extension of Rosenblatt's naive estimator (see section 1.1). Epanechnikov (1969) determines what the optimal kernel function should be in order to minimize the global error. He also indicates the insensitivity of the integral of the square of the kernel to the shape of the kernel for nonnegative weight functions.

Several other methods of density estimation exist which differ from the kernel estimation method. The first of these, to be discussed briefly, is the orthogonal series represen-

tation of a density function. Cencov (1962) developed a class of estimators of f(y) by expansion of an orthogonal set with respect to some weight function. Schwartz (1967), using Hermite functions, considers the univariate case. Kronmal and Tarter (1968) estimate the density function by means of generalized Fourier series. In particular, they chose a special case, namely the trigonometric systems {cos kπy}, {sin kπy}, {cos kπy, sin kπy}. Watson (1969) also discusses density estimation by orthogonal series. In addition, we mention that Kronmal and Tarter (1968) showed that it is possible to express the orthogonal series estimators in the form of the kernel estimators.

Another method is that put forth by Boneva, Kendall, and Stefanov (1971). Their method transforms the classical histogram into a smooth curve or histospline. While there exists a one-to-one relationship between histograms and histosplines, a definite disadvantage inherent in their procedure is that the curve is negative at some parts.

Another method similar to the kernel method is the rank kernel method of Hendrickson (1972). The kernels of this method depend on the ranks of the order statistics within the sample. This method is discussed more thoroughly in Chapters II and III.

Finally, we mention the modified maximum likelihood estimator of a probability density function, proposed by

Good (1971). A maximum likelihood estimate of a density function would give a Dirac delta function assigning mass of 1/n at each of the n observations. Because this is obviously too rough an estimate, Good suggests subtracting a roughness penalty (a function of f(y)) from the log-likelihood. Thus, he proposes to use an estimator which maximizes the function

$$w = \sum_i \log f(y_i) - \Phi(f)$$

where $\Phi(\cdot)$ is the roughness penalty.

Other articles which are of interest and which have much in common with the aforementioned literature are listed below in chronological order: Maniya (1961), Blaydon (1967), Lin (1968), Moore and Henrichon (1969), Pelto (1969), Pickands (1969), Schuster (1969, 1972), Van Ryzin (1969, 1972), and Specht (1971). A more extensive tabulation can be found in the bibliography.

1.1 The Rosenblatt Tradition (The Kernel Estimates of Rosenblatt and Parzen)

Let $Y_1, Y_2, \ldots, Y_n$ be independent identically distributed random variables with continuous density function f(y). Let $f_n(y)$ denote any estimator of f(y). Rosenblatt (1956) presents an obvious estimator of f(y) given by

$$f_n(y) = \frac{F_n(y+h) - F_n(y-h)}{2h} \qquad (1.1.1)$$

where $F_n(\cdot)$ is the sample distribution function and $h = h(n)$ is a function of n, which approaches zero as $n \to \infty$.

He shows the asymptotic mean square error to be

$$E[f_n(y) - f(y)]^2 \sim \frac{f(y)}{2hn} + \frac{h^4}{36}[f''(y)]^2 \qquad (1.1.2)$$

as $h \to 0$ and $n \to \infty$. Next, the question of an optimal choice for h is considered in order to minimize equation (1.1.2). This is found to be

$$h = k\, n^{-1/5} \qquad \text{where} \qquad k = \left\{ \frac{9}{2} \frac{f(y)}{[f''(y)]^2} \right\}^{1/5}$$

Rosenblatt also considers the minimization of the mean integrated square error or global error

$$\int_{-\infty}^{\infty} E[f_n(y) - f(y)]^2 dy \sim \frac{1}{2nh} + \frac{h^4}{36} \int_{-\infty}^{\infty} [f''(\xi)]^2 d\xi \ .$$

The optimal h for this error is no longer a function of y. The corresponding value of h is given as

$$h = k\, n^{-1/5} \qquad \text{where} \qquad k = \left\{ 9/2 \int_{-\infty}^{\infty} [f''(\xi)]^2 d\xi \right\}^{1/5} .$$

Rosenblatt suggests a more general class of estimators, which contains those given by (1.1.1). This is the class of kernel estimators defined by

$$f_n(y) = \frac{1}{n} \sum_{j=1}^{n} w_n(y - Y_j) \qquad (1.1.3)$$

where the nonnegative sequence of weights $w_n(\cdot)$ satisfies the following two conditions:

(a) $\int_{-\infty}^{\infty} w_n(u)du = 1$

(b) $\int_{|u|<\varepsilon} w_n(u)du \to 1$ for any $\varepsilon > 0$ as $n \to \infty$.

Note that estimators of this form are themselves a density.

He then states that if

$$w_n(u) = \frac{1}{h} w\left(\frac{u}{h}\right) \text{ and } w(z) = \begin{cases} \frac{1}{2} & |z| \leq 1 \\ 0 & \text{otherwise} \end{cases} \qquad (1.1.4)$$

then equation (1.1.3) reduces to equation (1.1.1) where $h \to 0$ as $n \to \infty$. Using the class of estimators given by equation (1.1.3), Parzen (1962) changes the notation and extends the results in much greater detail. His estimator is given as

$$f_n(y) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{y-Y_j}{h}\right) . \qquad (1.1.5)$$

Parzen shows that if the spreading coefficient $h = h(n)$ is chosen such that $h \to 0$ as $n \to \infty$ then the estimator given by equation (1.1.5) is asymptotically unbiased; that is,

$$\lim_{n\to\infty} E[f_n(y)] = f(y) .$$

The conditions on $K(\cdot)$ are the following:

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

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

(c) $\lim_{x\to\infty} |x\ K(x)| = 0$ and

(d) $\int_{-\infty}^{\infty} K(x)dx = 1$.

The estimator $f_n(y)$ is also shown to be consistent in quadratic mean in the sense that

$$E[f_n(y) - f(y)]^2 \to 0 \quad \text{as} \quad \begin{matrix} n \to \infty \\ h(n) \to 0 \\ nh(n) \to \infty \end{matrix} .$$

If it is assumed that $K(\cdot)$ is an even function (in this connotation called a weighting function), the sequence $\{f_n(y)\}$ is proven to be asymptotically normal in the sense that

$$\lim_{n\to\infty} P\left[\frac{f_n(y) - E[f_n(y)]}{\sqrt{Var[f_n(y)]}} \leq c\right] = \Phi(c)$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Parzen also proves that if the probability density $f(y)$ is uniformly continuous, then for every $\varepsilon > 0$,

$$P\left[\sup_{-\infty<y<\infty} |f_n(y) - f(y)| < \varepsilon\right] \to 1$$

as $n \to \infty$ and $nh^2 \to \infty$.

Finally, he presents results corresponding to those of Rosenblatt for determining the optimal h such that the mean square error (M.S.E.) is minimized. The M.S.E. is given by

$$E[f_n(y) - f(y)]^2 \sim \frac{f(y)}{nh}\int_{-\infty}^{\infty} K^2(x)dx + h^{2r}[k_r f^{(r)}(y)]^2$$

where r is the characteristic exponent of the Fourier transform k(u) of the weighting function $K(\cdot)$,

$$f^{(r)}(y) = -(2\pi)^{-1}\int_{-\infty}^{\infty} \exp\{-iuy\}|u|^r\phi(u)du ,$$

and $\phi(u)$ is the characteristic function of f(y). The optimal h is given by

$$h = \left[\{f(y)\int_{-\infty}^{\infty} K^2(x)dx\}/\{2nr|k_r f^{(r)}(y)|^2\}\right]^{1/(2r+1)}$$

Parzen does not consider an optimal h for the mean integrated square error but it would be given by

$$h = \left[\int_{-\infty}^{\infty} K^2(x)dx/2nr\int_{-\infty}^{\infty} |k_r f^{(r)}(y)|^2 dy\right]^{1/(2r+1)} .$$

One very important disadvantage to the generic form of

the optimal h derived by Rosenblatt and by Parzen should be noticed immediately (this is also true for Watson and Leadbetter). This is that the optimal h is necessarily a function of f(y), the true unknown density, which we are trying to estimate. The interpretation of these results would be that of a "bootstrap method". In other words, if we know f(y) then we would be able to optimally estimate it by $f_n(y)$. Epanechnikov (1969) attempted to eliminate the difficulty involved here, but was unable to do so. He found the optimal kernel form K(•) and then found the optimal spreading coefficient h(n). Significantly, K(•) is independent of the true probability density function, the sample size, and the dimensionality of the space; however, this is not true of the optimum h. Unfortunately, h is still a function of the unknown density f(y).

In closing this section, we again mention Cacoullos' (1964, 1966) multivariate extension of the material presented by Parzen. Also, Rosenblatt (1970) considered estimates of a density function when the observations are dependent. The asymptotic results are essentially the same as those in the case of independent observations.

1.2 Applications of Density Estimation

Various applications of density estimation can be found in the following articles: classification or discriminatory analysis (Fix and Hodges (1951, 1952), Stoller (1954), Gupta (1964)); derivatives of a density function (Bhattacharya (1967), Lin (1968), Schuster (1969)); empirical Bayes (Martz and Krutchkoff (1969), Bennett (1970)); hazard and reliability analysis (Watson and Leadbetter (1964a,b), Murthy (1965b), Martz and Hailey (1971)); regression (Nadaraya (1964c, 1965, 1970), Rosenblatt (1969), Schuster (1972)); and signal detection (Cooper (1964)).

Let us now consider a practical application of density estimation from reliability analysis. Reliability is defined as the probability of a component (or system) performing its purpose adequately under the operating conditions encountered for the period of time intended. Mathematically, the reliability function is represented by

$$R(t) = \Pr[T \geq t]$$

$$= \int_t^\infty f(\tau)d\tau$$

where the random variable T is the time to failure of the component and $f(\cdot)$ is the failure density. Consequently,

estimating f(t) enables us to estimate R(t) by performing the integration in equation (1.2.1),

$$\hat{R}(t) = \int_t^{\infty} \hat{f}(\tau)d\tau \qquad (1.2.1)$$

where the "roof" indicates an estimate.

Another fundamental problem of interest to reliability engineers is the estimation of the hazard function (sometimes called the failure rate). The hazard function, z(t), is the probability that an item, operating at time t, will fail in the interval [t, t+Δt]. Mathematically, the hazard function is represented as

$$z(t) = \frac{f(t)}{1-F(t)} = \frac{f(t)}{R(t)} \quad . \qquad (1.2.2)$$

Therefore, we can estimate the hazard function by

$$\hat{z}(t) = \frac{\hat{f}(t)}{\hat{R}(t)} \qquad (1.2.3)$$

where $\hat{R}(t)$ is obtained as indicated above. The theoretical failure rate curve is high initially, decreasing with time, followed by a time period where the function is relatively constant. Finally, the failure rate increases with time as components reach the end of their useful life. This curve is appropriately named the "Bathtub Curve" and the three time

periods correspond to the region of early failures (infant mortality), the region of random failures (useful life), and the region of wearout failures (adult mortality).

For this reason, it is easier to formulate models based on $z(t)$ rather than on $f(t)$; however, any analytical procedure that is applied to the hazard curve to determine a model directly produces models for both $f(t)$ and $R(t)$. By means of equation (1.2.2), it can be shown that

$$f(t) = z(t) \exp[-Z(t)]$$

where $Z(t) = \int_0^t z(\tau)d\tau$. In other words, to any nonnegative function $z(t)$ for which $Z(\infty) = \infty$ there corresponds a distribution and conversely.

Many existing techniques used to estimate the reliability of a device or component assume a specified form of the underlying failure density. Typically, these forms are one of a family which include the exponential, Weibull, and gamma distributions. The initial step in the treatment of failure data is to first compute a piecewise continuous failure density and hazard rate. The choice of the model which sufficiently explains the data is then determined from these graphs. Suppose N items are placed in operation at time $t = 0$, and let $n(t)$ be the number of survivors at time t. The empirical density function over the time interval

$t_i < t < t_i + \Delta t_i$ is given by the ratio of the number of failures occuring in this interval to the size of the original sample, divided by the length of the time interval; that is,

$$f_d(t) = \frac{[n(t_i) - n(t_i + \Delta t_i)]/N}{\Delta t_i} \quad .$$

Similarly, the empirical hazard rate over the same time interval is defined as the ratio of the number of failures occuring in the interval to the number of survivors at the beginning of the interval, divided by the length of the time interval; that is,

$$z_d(t) = \frac{[n(t_i) - n(t_i + \Delta t_i)]/n(t_i)}{\Delta t_i} \quad .$$

A discussion of the choice of t_i and Δt_i can be found in Shooman (1968).

Thus, the initial procedure in the choice of a failure model is to plot the histogram for $z_d(t)$. The general contour of this empirical function is perhaps the best indication of the model that should be chosen -- constant failure rate model, linearly increasing or decreasing failure rate model, Weibull model, et cetera. Then a suitable model is chosen by inspection of the $z_d(t)$ function. Finally, statistical techniques can be applied to efficiently process

the data and obtain "best" estimates of the parameters in the model. These techniques include least squares, moment estimation, and maximum likelihood. Notice that all of these methods require an underlying failure or hazard model to be assumed.

Several other methods of reliability estimation are presented in the articles by Taylor and Lochner (1965) and Schwartz, Seltzer, and Stehle (1965). Shooman (1968) discusses an estimate based on the sample distribution function. Each of these approximations of the true reliability function are variations of the basic histogram.

The foregoing material has briefly discussed some techniques for approximation of the reliability function and the hazard rate. In Chapters II-IV, we introduce some other estimation procedures which may be applied to reliability analysis.

Using any of the estimators presented in Chapters II-IV in conjunction with equations (1.2.1, 1.2.3), we can then form estimates of either the reliability function or the hazard function, respectively. An example is shown in Chapter V.

1.3 Outline of Succeeding Chapters

The estimator of the unknown density function developed in Chapter II is similar to those estimators of Rosenblatt and of Parzen. However, the kernel we consider is a function of the rank of each observation. We use, as our kernel, the asymptotic distribution of the order statistics of a sample. As is usually the case, we wish to determine the optimal h.

Chapter III uses the estimator of Chapter II as the basis for the development of another estimation procedure. The method employed is the method of successive substitution in which the solution at each iteration is used to generate the next solution until convergence is obtained.

In Chapter IV, a sequential procedure is developed for estimation of a probability density function. Initially, a normal density with mean $\bar{x}$ and variance s^2 is fitted to the data and a goodness of fit test is performed to ascertain the degree of approximation. This hypothesis rejected, a sequential procedure employing the concept of spline functions is used.

Several examples are given in Chapter V which compare the various methods of density estimation introduced in the preceeding three chapters. These examples have been selected to illustrate both good and bad features of the estimators. Two examples illustrate the application of the estimators to

reliability analysis.

Finally, relevant computer programs (Fortran) and descriptions of their utilization appear in the appendices. Appendix A contains the computer program for Chapters II and III while Appendix B contains the computer program for Chapter IV.

CHAPTER II

A RANK KERNEL ESTIMATOR OF A DENSITY FUNCTION

2.1 Introduction

In this chapter we are concerned with the estimation of a probability density function f(y) of a univariate random variable from a sample of size n. Suppose that we know nothing about the density f(y) except that it is continuous. Let Y be a univariate random variable and let $Y_1, Y_2, \ldots, Y_n$ be a sample of n independent observations, each being distributed according to the common continuous density function f(y). In order to eliminate repetition, we mention that the above discussion is also relevant to the introduction of each chapter and will not be given again.

The estimator discussed in this chapter was introduced by Hendrickson (1972) and is similar to one proposed by Rosenblatt and by Parzen (see equations (1.1.3), (1.1.5)). The estimator differs from theirs by having kernels which are functions of the ranks of the observations.

The order statistics are obtained by arranging the sample in ascending order,

$$Y_{k_1} \leq Y_{k_2} \leq \ldots \leq Y_{k_n} \quad (1 \leq k_i \leq n,\ k_i \neq k_j) ,$$

and setting $X_r = Y_{k_r}$ $(r = 1, \ldots, n)$. Let p $(0 < p < 1)$ denote

a fraction such that $F(y) = p$ and assume that there exists a unique solution for y such that $f(y) \neq 0$. This solution is called the quantile of order p and is denoted by ξ_p; that is,

$$P(Y \leq \xi_p) = F(\xi_p) = p \ .$$

In an analogous manner, the p^{th} quantile of the set of n observations is defined as the value $\hat{\xi}_p$ such that the proportion of observations $\leq \hat{\xi}_p$ is $\geq p$ and the proportion of observations $\geq \hat{\xi}_p$ is $\geq 1-p$. Thus, we can define the function $\hat{\xi}_p$ in terms of the ordered observations as

$$\hat{\xi}_p \equiv \hat{\xi}_p(X_1, X_2, \ldots, X_n) = \begin{cases} X_{[np]+1} & \text{if } np \notin I \\ \\ X_{np} & \text{if } np \in I \end{cases}$$

where I is the set of integers and [np] denotes the greatest integer not exceeding np.

We shall now state several theorems that play an important role in the formulation of our estimator, $f_n(y)$, of the true density f(y). These theorems may be found in Kendall and Stuart (1969) and Rao (1965). It is well-known that the probability element of the r^{th} order statistic from a sample of size n is given by

$$g_r(x)dx = \frac{n!}{(r-1)!(n-r)!}[F(x)]^{r-1}[1-F(x)]^{n-r}f(x)dx \ . \tag{2.1.1}$$

In addition, by letting r and n tend to infinity with their ratio, $\frac{r}{n} = p$, held constant, the asymptotic distribution of $\hat{\xi}_p$ is normal with mean ξ_p and variance $p(1-p)/n[f(\xi_p)]^2$; abbreviated $N(\xi_p,\sigma_p^2)$.

In the following section, we show how these theorems are applied to form our estimator.

2.2 A Class of Estimators Of A Density Function

Consider the estimator given in equation (2.2.1), which is essentially a weighted average of the sample distribution function,

$$\begin{aligned} f_n(y) &= \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{y-x}{h}\right) dF_n(x) \\ &= \frac{1}{nh} \sum_{r=1}^{n} K\left(\frac{y-X_r}{h}\right) \ . \end{aligned} \tag{2.2.1}$$

Instead of using the estimator above (Parzen's estimator), let us consider a kernel $K(\cdot)$ which is a function of the rank of each of the order statistics. The estimator we propose is defined by

$$f_n(y) = \frac{1}{nh} \sum_{r=1}^{n} N_r\left(\frac{y-X_r}{h}\right) \qquad (2.2.2)$$

where

$$N_r(y) = \frac{\sqrt{n}\, d(\xi_p)}{\sqrt{2\pi\, p(1-p)}} \exp\left\{-\frac{1}{2}\frac{nd^2(\xi_p)}{p(1-p)}\, y^2\right\} \quad -\infty<y<\infty \ , \qquad (2.2.3)$$

$d(\cdot)$ is some density function, and ξ_p is the p^{th} quantile of $d(\cdot)$. We will refer to the estimator given by equation (2.2.2) as the normal rank kernel estimator. In order to prevent division by zero when $r = n$, we will redefine the quantity p to be $r/(n+1)$. It is evident that estimates of this form are themselves density functions.

The function $d(\cdot)$ is an initial guess for the density from which the observations were generated. Because we choose a particular algebraic representation in order to ascertain the corresponding probability densities $N_r(\cdot)$, this does not imply that the sample was actually produced by such a population. Obviously, if we knew the parent distribution, there would be no need for estimation. Thus, our assumption for $d(\cdot)$ is analogous to Parzen's assumption of a particular kernel $K(\cdot)$.

The rationale for our estimator is given by an interesting result of order statistics stated in Theorem 2.2.1. By ranking the observations and using the limiting process mentioned previously, we obtain a kernel which varies

according to the location of each observation within the entire sample. In forming our rank kernel estimator, we have replaced the distribution of the r^{th} order statistic, on the left hand side of the result stated in Theorem 2.2.1, by the asymptotic normal distribution.

So that notation is not confused, we restate that

$f(y)$ is the true density function,

$g_r(y)$ is the true density function of the r^{th} order statistic, and

$d(y)$ is an initial estimate of $f(y)$.

Theorem 2.2.1 Given the densities, $g_r(\cdot)$ $r = 1,\ldots,n$, of the r^{th} order statistic then

$$\frac{1}{n} \sum_{r=1}^{n} g_r(y) = f(y) \ .$$

The proof follows from the following equation:

$$\frac{1}{n} \sum_{r=1}^{n} \frac{n!}{(r-1)!(n-r)!} [F(y)]^{r-1}[1-F(y)]^{n-r} f(y)$$

$$= f(y) \sum_{r=1}^{n} \binom{n-1}{r-1} p^{r-1} q^{n-r} \ ,$$

where $p = F(y)$ and $q = 1-F(y)$. The above result is an application of the binomial theorem.

One might consider using $h_r(\cdot)$ as a kernel where $h_r(\cdot)$ is the density of the r^{th} order statistic obtained from the

assumed form for $d(\cdot)$. However, to use a kernel of this form would indeed require laborious computations.

2.3 Asymptotic Properties of the Estimator

If the quantity $p = \frac{r}{n+1}$ of equations (2.2.2) and (2.2.3) is replaced by $F(x_r)$, where F is the true cumulative distribution function, one should expect the estimator to be asymptotically unchanged up to a given order of magnitude depending on h. When this is done, the kernels (2.2.3) depend solely on the values of x_j and not on the ranks. For fixed y, this estimator is of the form of equation (1.1.5) given by Parzen (1962) where $K(x_r)$ is written

$$K\left(y - \frac{h(y-x_r)}{h}\right) .$$

These kernels, however, do not satisfy the symmetry requirement. It is clear that, unless the initial guess $d(\cdot)$ is close to $f(\cdot)$, the estimator given by (2.2.2) and (2.2.3) has no advantage over the less complicated form of Rosenblatt and Parzen. In the next chapter we give a procedure which replaces $d(\cdot)$ by the estimate $\hat{f}(\cdot)$. Hendrickson (1973) has given asymptotic results when $d(\cdot)$ is replaced by $f(\cdot)$. We discuss these results without proof.

Sample Percentiles

The estimator given by (2.2.2) and (2.2.3) has the same variance (within $o(\frac{1}{nh})$) and bias (within $o(\frac{h^2}{n}) = o(\frac{1}{m})$) as that obtained by using m sample percentiles provided:

i) the sample percentiles satisfy $F(x_r) = \frac{r-\frac{1}{2}}{m}$,

ii) p is replaced by $\frac{r-\frac{1}{2}}{m}$,

iii) n is replaced by m and h is replaced by 1, and

iv) $m = \frac{n}{h^2}$.

The choice $p = F(x_r) = \frac{r-\frac{1}{2}}{m}$ is essential for the bias of the two procedures to be within $o(\frac{1}{m})$. The procedure is considerably improved by using $\frac{r-\frac{1}{2}}{m}$ rather than $\frac{r}{m+1}$. Thus, although the use of $\frac{r}{m+1}$ as done in the next chapter is instructive, better results should be expected by using $\frac{r-\frac{1}{2}}{m}$.

Bias

The mean of $f_n(y)$ when $d(\cdot) = f(\cdot)$ is

$$E(f_n(y)) = f(y)(1 - \frac{h^2}{2n} + \frac{h^2\sigma''(y)}{n}) + o(\frac{h^2}{n}) \qquad (2.3.1)$$

where $\sigma(y)$ is defined by

$$\sigma(y) = \frac{\sqrt{F(y)(1-F(y))}}{f(y)}. \qquad (2.3.2)$$

Hendrickson (1973) gives an adjustment to the estimator which eliminates the value $\frac{h^2}{2n}$. Usually, $\sigma(y)$ has little curvature.

Variance

The variance of $f_n(y)$ is

$$\mathrm{Var}(f_n(y)) = \frac{f(y)}{2\sqrt{\pi n}} + o(\frac{1}{\sqrt{n}} + \frac{h^2}{n}) \qquad (2.3.3)$$

This is similar to Parzen's result (1962) with $\frac{h^2}{n}$ equal to his h^2.

2.4 Determination of An h

In order to employ equation (2.2.2) we must first choose a value for h. Obviously, the choice of h will influence the resulting estimate. However, instead of indiscriminately selecting a value for h, an optimal h is usually determined by minimizing the mean square error. Unfortunately, as illustrated by the principles applied by Rosenblatt and by Parzen, a different value for h is required for each value of y at which the density f(y) is estimated.

Suppose one has a sample of size n. As stated in the previous section, one can choose m sample percentiles by the equation

$$F(x_i) = \frac{i-\frac{1}{2}}{m} ,$$

and use these m points to obtain the estimator with a value of h = 1. Alternatively, one can employ all of the n data and use a value of $h^2 = n/m$. In this section we discuss the case when the x_i are chosen by

$$\hat{F}(x_i) = \frac{i}{m+1} . \qquad (2.4.1)$$

One cannot obtain the same estimator by using all the data with a fixed spreading coefficient because the x_i are not centrally located in m equal frequency intervals. In this section we consider using points defined by (2.4.1) and choosing a spreading coefficient h* which minimizes the bias of the estimator.

The bias of the estimator can be asymptotically obtained by replacing (2.4.1) with the true percentiles satisfying

$$F(x_i) = \frac{i}{n+1} . \qquad (2.4.2)$$

These points are not random, but represent the mean. Let $f_n(y)$ be the expected density defined by the estimator given by the points (2.4.2). Since the bias depends on y, we will choose h* by minimizing the mean integrated square error (M.I.S.E.),

$$\int_{-\infty}^{\infty} E[f_n(y) - f(y)]^2 w(y)dy \ . \qquad (2.4.3)$$

The weight function $w(\cdot)$ is usually taken to be identically one or $f(\cdot)$.

By using the M.I.S.E., only a single value of h is required for all values of y. This measure is difficult to obtain numerically. We will consider a slightly different criterion to obtain the value of h to use in equation (2.2.2). The criterion we use is to minimize the average square error calculated at the observations,

$$\frac{1}{n} \sum_{i=1}^{n} [f_n(x_i) - f(x_i)]^2 \ . \qquad (2.4.4)$$

The value of h for which this minimum is attained, denoted h*, is determined only when the true density is known. Equation (2.4.4) is a Riemann sum for (2.4.3) provided the x_i are midpoints of the n equal frequency intervals corresponding to the density $w(\cdot)$. Thus there is little difference between (2.4.3) and (2.4.4) for arbitrary x_i, provided $w(\cdot)$ is appropriately chosen. Equation (2.4.4) measures the performance of $f_n(y)$.

For a given density f(y) we generate the true percentiles by (2.4.2). Using these data, $X_1, X_2, \ldots, X_n$, the estimate $f_n(y)$ is formed and then the average square error

(A.S.E.) is calculated for various values of h. From the tabulation of h versus A.S.E., the value of h* corresponding to the minimum average square error is ascertained. In our simulation study, the value of h* yielding the minimum A.S.E. has been determined to four decimal places since very little decrease in the average square error resulted from more significant digits.

To test the performance of the estimator presented in this chapter and also the estimator introduced in Chapter III, we have chosen to simulate density functions which are representative of a broad class of densities. We classify these densities as light-tailed, medium-tailed, and heavy-tailed. The light-tailed distributions are represented by a uniform density; the medium-tailed by a normal density; and the heavy-tailed by a cauchy density function. In addition, we simulate a weibull density and an exponential density since they are often used reliability models and because they are also unsymmetric. These five density functions are abbreviated in the tables as UNIF, NORM, CAUC, WEIB, and EXP respectively. The specific values of the parameters of each density used in the monte carlo studies are given in Table V.

Besides using the true density f(y) in the formula for calculating the variance, we have chosen, for the sake of comparison, to replace f(x) by the density

$$d(x) = \frac{1}{x_n - x_1} .$$

We refer to this density as the underlying density and label this as type I when the uniform density is used, and type II when the true density is used. The sample sizes are n = 10, 20, 40, 60, 80, 100.

For a given true density, underlying density, and sample size we have used the true percentiles to determine the value of h* yielding minimum A.S.E. This value of h* was then used for estimating the density of twenty random samples generated from the true density. For each sample, the average square error was computed and then the average of the average square error for the twenty experiments was also calculated. The same random sample was used in computing the estimate for the two underlying densities so that their individual estimates can be compared both separately and collectively.

Tables I and II contain the values of h* and the corresponding minimum average square error for each of the six sample sizes for the five simulated densities. Tables III and IV contain the average of the average square errors for the twenty random samples and also the median of the average square errors for the twenty experiments. The median was included since it will be less affected than the average by an occasional large error.

Tables I and II indicate that as the sample size increases the average square error continually decreases for each of the five densities. Furthermore, for the type I underlying density, the smoothness of the estimated curves using h* for the true percentiles tends to decrease at the tails of the estimate with the appearance of more noticeable bumps as n increases. The estimated curves using the type II underlying density for the true percentiles are much smoother in the sense that the bumps have been entirely eliminated. In addition, the type II density yielded estimated curves which are very hard to visually distinguish between one another for the various sample sizes.[1] This is in contrast with the type I estimates where the tails become much more bumpy as n increases.

For both types of underlying density, values of h less than that value (h*) yielding minimum A.S.E. produce curves which become extremely oscillatory as h decreases while values of h greater than that value yielding minimum A.S.E. produce curves which become smoother as h increases. However, as the value of h increases, the estimated density begins to

---

[1]The exception is the rate at which the exponential density estimates reach their maximum for different n. Also, for the uniform density function, the estimates are steeper (more vertical) at the extremes as n increases.

assume the characteristics of a normal density function.[2] Aesthetically, the "best" estimates using the type I density are for these values of h since even the slightest protuberance (in the estimated curve using h*) has been smoothed out. At the opposite extreme, as the value of h decreases, the estimate is no longer a single curve but a group of n disjointed curves.

For a given sample size and underlying density, it is difficult to distinguish between the shape of one estimated density and another for various values of h with accuracy greater than one decimal place once the oscillatory nature of small h is overcome. For accuracy greater than two decimals, it is impossible to visually differentiate between estimated densities.

Unfortunately, Tables I and II do not indicate a general trend in the value of h* that minimizes the A.S.E. as n increases. For the weibull type I estimates, the value of h* decreases for n = 10 through n = 40 and then increases, while for the weibull type II estimates the value of h* continually increases. For both the exponential and uniform type I and II estimates, the value of h* decreases for all n but less noticeably as n approaches 100. For the normal type I estimates, the value of h* decreases from n = 10

---

[2]The estimates of the exponential density do not respond to this principle as rapidly.

through n = 40 and then increases, while for the normal type II estimates the value of h* increases for all n. Finally, for the cauchy type I estimates, the value of h* decreases for all n, while for the cauchy type II estimates the value of h* increases from n = 10 through n = 60 and then decreases. The drastic change in the value of h* evident in Table II for the cauchy estimates is explained by the fact that the analysis of the plot of h versus A.S.E. for each n had two local minimums. One of these minimums was the global minimum for $n \leq 60$ while the other minimum was global for $n \geq 80$. No other type I and II estimates had more than one minimum.

Generally, for each sample size, the value of h* yielding minimum A.S.E. for the type I estimates is less than that obtained for the type II estimates[3] whereas the average square error for the type I estimates is greater than that for the type II estimates.[4] From the latter result and the fact that the estimated densities are much smoother, the type II estimates afford the "best" estimates.

Now let us turn our attention to Tables III and IV which tabulate the computations of the twenty random experiments. These tables, as did the previous two, indicate that the

---

[3]The reverse is true for the uniform type II density estimates.

[4]For the uniform density, the differences are so small that there is no practical difference and neither the type I or the type II estimates are preferable to the other.

general tendency of the average of the average square error is decreasing for each of the simulated densities as the sample size increases. The type I estimated densities for the random samples are mostly polymodal; however, the general characteristics (shape, skewness, and kurtosis) of each is that of the parent distribution. The exceptions to the latter portion of this statement occur for samples of size ten. Once again, the type II estimates are smoother in the sense that many of the bumps have been eliminated. In fact, the polymodal character has been significantly reduced and many of the estimates are unimodal or bimodal.

For the most part, the average A.S.E. for the type I estimates are greater than those for the type II estimates as is also true of the median values. Again, this indicates the preferableness of the type II estimates.

Finally, we mention several exceptions. The exponential type II densities overestimate in the neighborhood of τ moreso than the type I, which explains why the fit is worse in terms of the average of the average square error. On the other hand, the general characteristics of the type II estimates are more indicative of an exponential density. Again the difference between the uniform type I and II estimates are not large enough to warrant preference for either underlying density over the other. And, the extreme change in the value of h* mentioned previously for the cauchy type II estimates is also reflected in the average of the average square errors.

TABLE I

MINIMUM AVERAGE SQUARE ERROR
AND CORRESPONDING H

TYPE I ESTIMATES

		N=10	N=20	N=40
WEIB	H	1.1243	0.9391	0.8977
	ASE	0.141742E-02	0.892545E-03	0.522061E-03
EXP	H	0.6146	0.4998	0.4407
	ASE	0.300186E-02	0.175239E-02	0.943605E-03
UNIF	H	0.9817	0.7412	0.6460
	ASE	0.229047E-04	0.812458E-05	0.239256E-05
NORM	H	1.0597	0.9129	0.8986
	ASE	0.688932E-03	0.331298E-03	0.181539E-03
CAUC	H	0.6669	0.3626	0.2004
	ASE	0.258583E-01	0.157502E-01	0.882636E-02

		N=60	N=80	N=100
WEIB	H	0.8985	0.9063	0.9161
	ASE	0.368070E-03	0.283116E-03	0.229602E-03
EXP	H	0.4199	0.4097	0.4035
	ASE	0.637000E-03	0.476551E-03	0.378361E-03
UNIF	H	0.6182	0.6056	0.5984
	ASE	0.113126E-05	0.661248E-06	0.435762E-06
NORM	H	0.9074	0.9145	0.9213
	ASE	0.132926E-03	0.105360E-03	0.869779E-04
CAUC	H	0.1419	0.1112	0.0921
	ASE	0.605136E-02	0.456620E-02	0.364505E-02

TABLE II

MINIMUM AVERAGE SQUARE ERROR
AND CORRESPONDING H

TYPE II ESTIMATES

		N=10	N=20	N=40
WEIB	H	1.3787	1.5144	1.6101
	ASE	0.881744E-04	0.433917E-04	0.126902E-04
EXP	H	0.9975	0.6455	0.6036
	ASE	0.274658E-03	0.109777E-03	0.311002E-04
UNIF	H	0.8033	0.6706	0.6141
	ASE	0.229056E-04	0.812442E-05	0.239255E-05
NORM	H	1.4515	1.6874	1.7631
	ASE	0.101324E-04	0.156191E-04	0.971784E-05
CAUC	H	1.2665	1.6373	2.1999
	ASE	0.119330E-02	0.352280E-03	0.145900E-03

		N=60	N=80	N=100
WEIB	H	1.6521	1.6735	1.6850
	ASE	0.554327E-05	0.302633E-05	0.188236E-05
EXP	H	0.5922	0.5856	0.5822
	ASE	0.145416E-04	0.845208E-05	0.550058E-05
UNIF	H	0.5983	0.5907	0.5863
	ASE	0.113124E-05	0.661245E-06	0.435761E-06
NORM	H	1.7808	1.8120	1.8379
	ASE	0.503096E-05	0.293306E-05	0.188378E-05
CAUC	H	2.6453	0.1962	0.1848
	ASE	0.112379E-03	0.755122E-04	0.475246E-04

TABLE III

AVERAGE OF ASE AND
MEDIAN VALUES USING H

TYPE I ESTIMATES

		N=10	N=20	N=40
WEIB	AVG	0.275586E-01	0.202168E-01	0.145025E-01
	MED	0.198951E-01	0.127365E-01	0.959039E-02
EXP	AVG	0.240031E-01	0.111832E-01	0.708801E-02
	MED	0.820527E-02	0.931156E-02	0.580388E-02
UNIF	AVG	0.297399E-02	0.268054E-02	0.184681E-02
	MED	0.163664E-02	0.187312E-02	0.165217E-02
NORM	AVG	0.872689E-02	0.535603E-02	0.329946E-02
	MED	0.573816E-02	0.429983E-02	0.300819E-02
CAUC	AVG	0.108234E 00	0.875186E-01	0.499589E-01
	MED	0.887091E-01	0.818137E-01	0.394537E-01

		N=60	N=80	N=100
WEIB	AVG	0.156243E-01	0.765269E-02	0.564703E-02
	MED	0.112789E-01	0.762728E-02	0.538121E-02
EXP	AVG	0.628158E-02	0.459051E-02	0.435268E-02
	MED	0.458456E-02	0.348470E-02	0.377006E-02
UNIF	AVG	0.184334E-02	0.161870E-02	0.167416E-02
	MED	0.176123E-02	0.134714E-02	0.164337E-02
NORM	AVG	0.275127E-02	0.298382E-02	0.186294E-02
	MED	0.214111E-02	0.237659E-02	0.173643E-02
CAUC	AVG	0.407544E-01	0.330077E-01	0.303375E-01
	MED	0.252487E-01	0.241719E-01	0.244362E-01

TABLE IV

AVERAGE OF ASE AND
MEDIAN VALUES USING H

TYPE II ESTIMATES

		N=10	N=20	N=40
WEIB	AVG	0.162104E-01	0.100942E-01	0.830628E-02
	MED	0.756946E-02	0.557716E-02	0.402304E-02
EXP	AVG	0.566446E-02	0.194349E-01	0.204896E-01
	MED	0.361346E-02	0.126328E-01	0.101405E-01
UNIF	AVG	0.228408E-02	0.251754E-02	0.184237E-02
	MED	0.132300E-02	0.180936E-02	0.167159E-02
NORM	AVG	0.473567E-02	0.240921E-02	0.167529E-02
	MED	0.284773E-02	0.197917E-02	0.137956E-02
CAUC	AVG	0.534145E-01	0.272772E-01	0.172913E-01
	MED	0.492499E-01	0.115703E-01	0.167358E-01

		N=60	N=80	N=100
WEIB	AVG	0.763084E-02	0.466218E-02	0.396903E-02
	MED	0.552627E-02	0.382822E-02	0.372052E-02
EXP	AVG	0.164149E-01	0.136696E-01	0.125414E-01
	MED	0.129085E-01	0.112653E-01	0.113174E-01
UNIF	AVG	0.180796E-02	0.161113E-02	0.167376E-02
	MED	0.168230E-02	0.135095E-02	0.163199E-02
NORM	AVG	0.141103E-02	0.142263E-02	0.103139E-02
	MED	0.820252E-03	0.107666E-02	0.781898E-03
CAUC	AVG	0.115773E-01	0.186757E 00	0.184296E 00
	MED	0.701361E-02	0.153693E 00	0.164644E 00

TABLE V

SIMULATED DENSITIES

WEIBULL $f(y) = \frac{\rho}{\theta} (y-\tau)^{\rho-1} \exp\left[\frac{-(y-\tau)^{\rho}}{\theta}\right]$ $y \geq \tau$

$\rho = 2$ $\theta = 3$ $\tau = 1$

EXPONENTIAL $f(y) = \frac{1}{\theta} \exp\left[\frac{-(y-\tau)}{\theta}\right]$ $y \geq \tau$

$\theta = 2$ $\tau = 1$

UNIFORM $f(y) = \frac{1}{\beta-\alpha}$ $\alpha \leq y \leq \beta$

$\beta = 10$ $\alpha = 0$

NORMAL $f(y) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right]$ $-\infty < y < \infty$

$\sigma = \sqrt{2}$ $\mu = 10$

CAUCHY $f(y) = \frac{1}{\pi} \frac{\alpha}{\alpha^2 + (y-\mu)^2}$ $-\infty < y < \infty$

$\alpha = .3$ $\mu = 5$

CHAPTER III

AN ITERATIVE ESTIMATOR OF A DENSITY FUNCTION

3.1 The Iterative Estimator

The estimator in the preceeding chapter had the generic form

$$f_n(y) = \frac{1}{nh} \sum_{r=1}^{n} \frac{\sqrt{n}\, d(\xi_p)}{\sqrt{2\pi p(1-p)}} \exp\left\{ \frac{-(y-x_r)^2 n d^2(\xi_p)}{2p(1-p)h^2} \right\}$$

whereas the estimator we propose in this chapter is given by

$$f_n^m(y) = \frac{1}{n} \sum_{r=1}^{n} N_r(y-x_r) \qquad m = 1,2,\ldots \qquad (3.1.1)$$

where

$$N_r(z) = \frac{\sqrt{n}\, f_n^{m-1}(x_r)}{\sqrt{2\pi p(1-p)}} \exp\left\{ \frac{-z^2 [f_n^{m-1}(x_r)]^2 n}{2p(1-p)} \right\} .$$

The superscript m denotes the iteration count.

The estimator in equation (3.1.1) uses the successive substitution method of iteration. Thus, we call this estimator the iterative estimator. Equation (3.1.1) uses a recursive relationship instead of an underlying density $d(\xi_p)$. A sequence of approximations $f_n^1, f_n^2, \ldots, f_n^m$ is computed from a starting value f_n^o by means of the following procedure: choose $f_n^o(x_j)$ for all j (j = 1,...,n) and hence

define $f_n^1(y)$; using $f_n^1(x_j)$ for all j define $f_n^2(y)$; and continue with this process until, for some m, $f_n^{m-1}(x_j)$ is within a given ε of $f_n^m(x_j)$ for all j. This gives the estimate of the true density function as

$$f_n(y) = \frac{1}{n} \sum_{r=1}^{n} \frac{\sqrt{n}\, f_n^m(x_r)}{\sqrt{2\pi p(1-p)}} \exp\left\{ \frac{-(y-x_r)^2 [f_n^m(x_r)]^2 n}{2p(1-p)} \right\} .$$

By means of computer simulations, we have found that the choice of $f_n^o(x_r)$ for each r (r = 1,2,...,n) to begin the iteration procedure is not critical as long as it is positive. Intuition leads one to suspect that too small an initial guess for $f_n^o(x_r)$ for each r will serve to increase the value of $f_n^1(x_r)$. Likewise, too large an initial guess for $f_n^o(x_r)$ for each r will serve to decrease the value of $f_n^1(x_r)$. The same effect will occur on each subsequent iteration until convergence is obtained. For convenience, the reciprocal of the sample range was used for the starting values,

$$f_n^o(x_r) = \frac{1}{x_n - x_1} \qquad r = 1,2,\ldots,n .$$

The iteration procedure was terminated for the data in Appendix A when for all r = 1,2,...,n

$$|f_n^{m-1}(x_r) - f_n^m(x_r)| \leq .001 .$$

The iterative estimator converged for all of the computer simulations conducted. Furthermore, whenever a variety of different starting values (some of which were absurd) for $f_n^o(x_r)$, r = 1,2,...,n, were tested, the resultant estimates were identical.

3.2 Comparison of the Rank Kernel Estimator and the Iterative Estimator

As we did in Section 2.4, we have performed an experimental analysis using both true percentiles and random data for the thirty combinations of five true densities and six sample sizes. Table VI contains the number of iterations required to obtain convergence and the resultant average square error for the true percentiles for each combination of true density and sample size. These values of the average square error correspond to those found in Tables I and II.

The results of each set of twenty random samples are presented in Tables VII and VIII. Again, the same random samples that were used for both the type I and type II estimates in Chapter II were used here so that the experimental outcomes would be directly comparable. Table VII contains the average of the average square errors and also the median for each set of twenty monte carlo studies. The contents of this table correspond to those found in Tables III and IV.

Finally, the average, the medium, the minimum, and the maximum number of iterations required for convergence of the random experiments are found in Table VIII.

Before presenting the analysis between the estimates of the rank kernel estimator and the estimates of the iterative estimator, let us first discuss the tabulations of the latter.

Table VI indicates, for each simulated density, that both the average square error and the number of iterations needed to obtain convergence continually decreases as the sample size increases. In addition, the smoothness of the estimated density curves is quite similar to the smoothness of the type II estimates. If you recall, the type I estimates are bumpy in the tails while the type II estimates eliminated this protuberancy. Generally, for a given sample size, it is difficult to visually differentiate between the curve of the iterative estimate and that of the type II estimate using h* for the true percentiles.

The tabulations in Table VII also indicate the general tendency of the average of the average square errors to decrease with increasing sample size. The median values occassionally behave erratically, but the trend is to decrease as the sample size increases. Also, the median values are consistently less than their corresponding averages, as expected due to the occassional large error, but the magnitude of the differences decreases as n increases.

The estimated densities for the random samples are mostly polymodal; however, the general characteristics of the estimates are indicative of their parent population. The iterative estimates are inclined to accentuate oscillations where a concentration of data occurs and to de-emphasize them where data is sparse such as outlying observations. Thus, the densities are smoother in the tails than the type I estimates of Chapter II but have more pronounced fluctuations at the interior of the estimates.

Upon comparison of the average square errors in Tables I, II, and VI, we find that the iterative estimates are preferrable to the type I estimates while the type II estimates are preferrable to the iterative estimates. The only exception is that both the type I and II uniform estimates are preferred to the iterative uniform estimates. A similar comparison of Tables III, IV, and VII reveals that the iterative procedure yields the largest average A.S.E. and median values. Consequently, both types of the rank kernel estimate are favored to the iterative estimator.

Therefore, in our attempt to eliminate, by using the iterative estimator, the necessity to determine that value of h yielding minimum A.S.E. (h*) used in the rank kernel estimator, we have sacrificed precision in terms of the average square error. On the other hand, the iterative estimates are generally smoother than the type I estimates

but not as smooth as the type II estimates.

3.3 Estimation Using Sample Percentiles

The results in this chapter were obtained using the true percentiles. When random data are given, the density estimator is obtained by using sample percentiles. By using a subset of the data, rather than all the data as done in Rosenblatt's kernel method, a significant amount of computer time and, consequently, expenses is conserved. This saving is especially realized for very large samples.

Two examples are given in Chapter V which illustrate the employment of the sample percentiles and also how these estimation procedures may be applied to grouped data.

As mentioned earlier, the results we have obtained apply to the case when the k sample percentiles are defined by the relationship

$$\hat{F}(x_r) = \frac{r}{k+1} \qquad r = 1,2,\ldots,k \ . \tag{3.3.1}$$

We refer the reader to Hendrickson (1972) to similar procedures where the estimator is altered by considering other possible percentiles defined similarly to (3.3.1). These procedures alter the bias, discussed in the previous section, by an order of magnitude equal to $\frac{1}{k}$. These results are extended by Hendrickson (1973) to the multivariate densities.

TABLE VI

NUMBER OF ITERATIONS AND
AVERAGE SQUARE ERROR

TRUE PERCENTILES

		N=10	N=20	N=40
WEIB	IT	13	10	8
	ASE	0.200009E-02	0.425380E-03	0.940092E-04
EXP	IT	10	9	8
	ASE	0.102828E-02	0.529927E-03	0.260230E-03
UNIF	IT	5	5	5
	ASE	0.104284E-03	0.461289E-04	0.212541E-04
NORM	IT	11	7	6
	ASE	0.770359E-03	0.141040E-03	0.297105E-04
CAUC	IT	11	12	12
	ASE	0.121435E-01	0.167850E-01	0.596293E-02

		N=60	N=80	N=100
WEIB	IT	8	7	7
	ASE	0.403479E-04	0.213145E-04	0.134089E-04
EXP	IT	8	7	7
	ASE	0.171918E-03	0.129332E-03	0.103042E-03
UNIF	IT	4	4	4
	ASE	0.129761E-04	0.960305E-05	0.761944E-05
NORM	IT	6	6	6
	ASE	0.124006E-04	0.673084E-05	0.422606E-05
CAUC	IT	9	7	6
	ASE	0.224379E-02	0.109482E-02	0.623435E-03

TABLE VII

AVERAGE OF ASE AND
MEDIAN VALUES

RANDOM DATA

		N=10	N=20	N=40
WEIB	AVG	0.259405E 00	0.284157E 00	0.105352E 00
	MED	0.809269E-01	0.540018E-01	0.866485E-01
EXP	AVG	0.160603E 00	0.698248E-01	0.930493E-01
	MED	0.447671E-01	0.308866E-01	0.274153E-01
UNIF	AVG	0.831963E-01	0.119352E-01	0.592693E-02
	MED	0.121028E-01	0.947198E-02	0.440435E-02
NORM	AVG	0.353783E-01	0.274312E-01	0.216381E-01
	MED	0.167864E-01	0.194116E-01	0.173010E-01
CAUC	AVG	0.194606E 00	0.463530E 00	0.169439E 00
	MED	0.104118E 00	0.268228E 00	0.122605E 00

		N=60	N=80	N=100
WEIB	AVG	0.952018E-01	0.471694E-01	0.373864E-01
	MED	0.678295E-01	0.451351E-01	0.356049E-01
EXP	AVG	0.344048E-01	0.418888E-01	0.330315E-01
	MED	0.273165E-01	0.265948E-01	0.262744E-01
UNIF	AVG	0.498637E-02	0.479872E-02	0.432722E-02
	MED	0.407141E-02	0.339547E-02	0.374583E-02
NORM	AVG	0.164989E-01	0.172852E-01	0.120476E-01
	MED	0.111584E-01	0.118854E-01	0.106386E-01
CAUC	AVG	0.131672E 00	0.107304E 00	0.721403E-01
	MED	0.100954E 00	0.658640E-01	0.606866E-01

TABLE VIII

AVERAGE, MEDIAN, MINIMUM, AND
MAXIMUM NUMBER OF ITERATIONS

RANDOM DATA

		N=10	N=20	N=40	N=60	N=80	N=100
WEIB	AVG	18.05	18.00	15.50	17.05	15.40	15.25
	MED	15.00	16.00	15.00	16.00	15.00	15.00
EXP	AVG	17.30	18.20	17.15	16.05	17.90	16.70
	MED	17.00	17.50	15.00	16.00	17.00	15.00
UNIF	AVG	15.15	13.15	12.90	13.80	13.40	12.40
	MED	14.00	12.00	12.00	13.00	12.50	11.00
NORM	AVG	15.55	15.05	14.80	14.30	12.45	15.25
	MED	15.00	14.50	14.00	14.00	13.00	14.00
CAUC	AVG	18.30	19.00	17.35	17.40	15.05	16.50
	MED	18.50	18.00	16.00	17.00	15.00	16.00

		N=10	N=20	N=40	N=60	N=80	N=100
WEIB	MAX	47	39	23	27	19	21
	MIN	10	11	12	13	12	13
EXP	MAX	34	31	45	23	35	28
	MIN	10	12	12	12	13	12
UNIF	MAX	26	23	22	22	23	20
	MIN	10	9	9	9	9	9
NORM	MAX	29	20	27	20	15	33
	MIN	10	10	11	11	10	11
CAUC	MAX	27	27	32	30	21	22
	MIN	14	13	13	13	12	12

CHAPTER IV

A SEQUENTIAL SPLINE PROCEDURE FOR ESTIMATION OF A DENSITY FUNCTION

4.1 Introduction

Within the last few decades, spline functions have received considerable attention in the literature, mainly for the purpose of interpolation. Spline approximation has been discovered in several independent studies for four different criteria of optimality in the context of approximation of functions and best quadrature formulas. Splines (the word "splines" is frequently used as a synonym for spline functions) were first introduced in the middle 1940's by Schoenberg (1946) while investigating problems pertaining to the smoothing of equidistant data. The name spline function was derived from the mathematical analogy of a mechanical spline, a flexible device used by draftsmen to draw smooth curves by attaching weights at various locations on the spline in order to pass through specified points. The two free ends of the mechanical spline are straight, suggesting the terminology natural spline function (to be defined shortly). Basically, a spline function is a piecewise polynomial satisfying continuity conditions of the function and its derivatives that are less strigent then those of a polynomial. As such, they are direct generaliza-

tions of polynomials. Before giving the formal definition of a spline function, let us first present a brief historical background.

Schoenberg and Whitney (1953) obtained criteria for the existence of a particular type of spline interpolation. In 1957, Holladay demonstrated the minimum curvature property of cubic splines for interpolation; that is, minimization of the integral square measure of approximation to the second derivative. Ahlberg, Nilson, and Walsh (1962) extended this property to periodic cubic splines, and then exhibited the minimum norm property for periodic splines of odd degree (1965c). DeBoor (1963) and Schoenberg (1964a) investigated various approximation properties for nonperiodic splines of odd degree. The convergence of higher order spline approximations to derivatives was first obtained by Ahlberg, Nilson, and Walsh (1962) while Birkhoff and de Boor (1964) also considered spline approximation of derivatives.

The existence and uniqueness of bicubic spline surfaces of interpolation were demonstrated by deBoor (1962) after Birkhoff and Garabedian (1960) had extended the problem of spline approximation to two dimensions. Later, Ahlberg, Nilson, and Walsh (1965b) further extended these results to multidimensional splines. Also, Birkhoff and deBoor (1965) discuss splines of several variables.

The concept of trigonometric spline functions was

introduced by Schoenberg (1964b), which he related to a linear differential operator. The application of spline functions to linear differential operators was also investigated by Greville (1964), Ahlberg, Nilson and Walsh (1964, 1965a), and deBoor and Lynch (1966).

The preceeding articles are but a few of the proliferation of literature treating various approximations by spline functions. Several more recent papers and texts are Greville (1967, 1969), Karlin and Ziegler (1966, 1967), Reinsch (1967), and Schoenberg (1967, 1969). In addition, at least two dozen articles by such authors as deBoor, Golomb, Hall, Karlin and Karon, Meir and Sharma, and Schoenberg have appeared in the Journal of Approximation Theory since its inception in 1968. For an extensive bibliography, see Sard and Weintraub (1971). The formal definition of a spline function is presented now.

A spline function S(x) of degree m is a function characterized over the real line, R, by the following two properties:

(a) S(x) is given in each of the r+1 intervals (θ_i,θ_{i+1}) for i = 0,1,...,r by a polynomial of degree at most m where $\theta_0 = -\infty$ and $\theta_{r+1} = \infty$,

(b) S(x) and its derivatives of order 1,2,...,m-1 are continuous on R.

The sequence of real numbers $\theta_1,\theta_2,\ldots,\theta_r$ are called knots (or nodes) and satisfy the relationship $\theta_i < \theta_j$ if i < j.

A spline of odd degree $m = 2k-1$ is called a natural spline if, in addition to satisfying properties (a) and (b), it also satisfies property (c):

(c) $S(x)$ reduces to a polynomial of degree $k-1$ in each of the intervals $(-\infty,\theta_1)$ and (θ_r,∞).

It is the third degree (cubic) spline function that approximates the behavior of the mechanical spline.

Thus, a spline function is a class of functions defined by piecewise polynomial arcs of degree m in each interval (θ_i,θ_{i+1}) such that the composite function has continuous derivatives through the (m-1)st. In the special cases when $m = 0$ or $m = 1$, a spline is a step function (and condition (b) is non-functional) or a piecewise linear function (polygon), respectively. Another special case is for $r = 0$ for which a spline is a single polynomial on R.

Schoenberg and Whitney (1953) showed that any spline $S(t)$ of degree m can be represented in the form

$$S(t) = P_m(t) + \sum_{j=1}^{r} c_j (t-\theta_j)_+^m \qquad (4.1.1)$$

where $P_m(\cdot)$ denotes a polynomial of degree m. The subscript "+" is used to denote the truncated power function defined as

$$t_+^m = \begin{cases} t^m & t \geq 0 \\ 0 & t < 0 \, . \end{cases}$$

In the case of a natural spline, equation (4.1.1) becomes

$$s(t) = P_{k-1}(t) + \sum_{j=1}^{r} c_j (t-\theta_j)_+^{2k-1} . \qquad (4.1.2)$$

It is evident that s(t) satisfies property (c) in the interval $(-\infty,\theta_1)$. However, to satisfy this property in the interval (θ_r,∞), we must equate to zero the coefficient of powers greater than k-1 of t; or this implies

$$\sum_{j=1}^{r} c_j \theta_j^i = 0 \qquad i = 0,1,\ldots,k-1 .$$

Now let us turn our attention to the development of our spline estimator.

4.2 The Spline Estimator

Given a set of data $Y_1,Y_2,\ldots,Y_n$ obtained by experimentation, we want to estimate the density function f(y) from which the observations were produced. Using the concept of a spline function presented in the preceeding section, we define an estimator of the unknown density function. The estimator is denoted by p(y) and its mathematical form is given in equation (4.2.1);

$$p(y) = \exp\{S_2(y)\}$$

$$= \exp\{a_1 y^2 + by + c_1 + \sum_{i=2}^{k} c_i (y-a_i)_+^2\}. \qquad (4.2.1)$$

With slight modification in subscripts and parameters the function $S_2(\cdot)$ is recognized as a spline of degree two. We will call this estimator of the unknown probability distribution, the exponential spline estimator (or just spline estimator). Obviously $p(y) \geq 0$ for all y; however, in order that the spline estimator fulfill the requirements of a probability density function, we will require

$$\int_{-\infty}^{\infty} p(y)\, dy = 1 .$$

Using the method of maximum likelihood, we can determine estimates of the unknown parameters $\{a_i\}_1^k$, $\{c_i\}_1^k$, and b; and hence, an estimate of f(y). Therefore, in principle, the problem of determining the "best" approximation to f(y) is reduced to the simultaneous solution of 2k+1 equations.

The equation we desire to maximize is given by

$$L = \sum_{j=1}^{n} \ell n\, p(y_j) - \lambda \left[\int_{-\infty}^{\infty} p(y)dy - 1 \right] \qquad (4.2.2)$$

where λ is the Lagrangian multiplier and the term in brackets

is the restriction that the estimate integrate to one. Expanding equation (4.2.2) we obtain

$$L = a_1 \sum_{j=1}^{n} y_j^2 + b \sum_{j=1}^{n} y_j + nc_1 + \sum_{j=1}^{n} \sum_{i=2}^{k} c_i (y_j - a_i)_+^2$$
$$- \lambda \left[\int_{-\infty}^{\infty} \exp\{a_1 y^2 + by + c_1 + \sum_{i=2}^{k} c_i (y - a_i)_+^2\} dy - 1 \right]. \qquad (4.2.3)$$

Differentiating equation (4.2.3) with respect to the unknown parameters and the Lagrange multiplier, and setting the resultant equations equal to zero yields the following set of equations:

$$\sum_{j=1}^{n} y_j^2 = \lambda \int_{-\infty}^{\infty} y^2 \, p(y) \, dy \qquad (4.2.4a)$$

$$\sum_{j=1}^{n} y_j = \lambda \int_{-\infty}^{\infty} y \, p(y) \, dy \qquad (4.2.4b)$$

$$n = \lambda \int_{-\infty}^{\infty} p(y) \, dy \qquad (4.2.4c)$$

$$\sum_{j=1}^{n} (y_j - a_i)_+^2 = \lambda \int_{-\infty}^{\infty} (y - a_i)_+^2 p(y) dy \qquad (4.2.4d)$$

$$i = 2, \ldots, k$$

$$\sum_{j=1}^{n} (y_j - a_i)_+ = \lambda \int_{-\infty}^{\infty} (y - a_i)_+ p(y)dy \qquad (4.2.4e)$$

$$i = 2,\ldots,k$$

$$c_i \neq 0$$

$$\int_{-\infty}^{\infty} p(y)dy = 1 \quad . \qquad (4.2.4f)$$

From equations (4.2.4c,f) we find that $\lambda = n$. Therefore, we can reduce the set of equations (4.2.4) to the following:

$$\frac{1}{n} \sum_{j=1}^{n} y_j^2 = \hat{E}(y^2) \qquad (4.2.5a)$$

$$\bar{y} = \hat{E}(y) \qquad (4.2.5b)$$

$$\frac{1}{n} \sum_{j=1}^{n} (y_j - a_i)_+^2 = \int_{a_i}^{\infty} (y - a_i)^2 p(y)dy \qquad (4.2.5c)$$

$$i = 2,\ldots,k$$

$$\frac{1}{n} \sum_{j=1}^{n} (y_j - a_i)_+ = \int_{a_i}^{\infty} (y - a_i) p(y)dy \qquad (4.2.5d)$$

$$i = 2,\ldots,k$$

$$c_i \neq 0$$

$$\int_{-\infty}^{\infty} p(y)\, dy = 1 \quad . \tag{4.2.5e}$$

The equations (4.2.5a,b) are equivalent to the maximum likelihood equation for the variance. Let us mention here, that the parameters $\{a_i\}_2^k$ may be chosen by the experimenter, in which case the conditions given by equation (4.2.5d) are inoperative. We will say that these parameters are <u>fixed</u> if they are known a priori and that they are <u>free</u> if equation (4.2.5d) is functional.

In order that the spline estimator p(y) convey more meaning we alter the general form of equation (4.2.1) and rewrite it as follows:

$$f_n(y) \equiv p(y) = B_1 \exp\left\{ - \frac{(y-u)^2}{2s^2} - \sum_{i=2}^{k} c_i^*(y-a_i^*)_+^2 - c_1^*(y-a_2^*)_-^2 \right\}. \tag{4.2.6}$$

Now the 2k+1 unknown parameters are B_1, u, $\{c_i^*\}_1^k$, and $\{a_i^*\}_2^k$. The reader may verify that the same conditions as those given by the set of equations (4.2.5) must be satisfied. Most of the conditions are easily verified, but some difficulty is encountered in showing equations (4.2.5a and 4.2.5d for i = 2).

Thus, using the estimator in equation (4.2.6), if we set $c_i^* = 0$ for all i (i = 1,...,k) and $u = \bar{y}$, then the exponential spline estimator becomes the normal density

function with mean and variance, $\bar{y}$ and s^2 respectively. This is the initial stage of our sequential procedure for determining the estimate $p(y)$ of the unknown true density function $f(y)$; that is, we fit a normal distribution with mean $\bar{y}$ and variance s^2 to the data. Using the null hypothesis

$$H_0 : f(y) \sim N(\bar{y}, s^2)$$

we perform a goodness of fit test. If we reject this test, we proceed to fit a spline estimate with one knot. Again, we perform a goodness of fit test with the hypothesis

$$H_0 : f(y) \sim p(y) \text{ with one knot.}$$

This procedure may be repeated with as many knots as needed. As we have mentioned previously, the experimenter may choose to use a predetermined set of knots. In this case, he may want to dispense with a goodness of fit test after the addition of each knot.

Several examples employing the spline estimator are given in Chapter V while the computer program, along with a description of its usage, is presented in Appendix B.

CHAPTER V

EXAMPLES AND CONCLUSION

5.1 Introduction

Having presented three different estimators of a probability density function in the preceeding chapters, we now give several examples utilizing these estimation methods. The examples use data that are both simulated and real. These examples have been selected to illustrate some good and bad features of these estimators. In addition, one example illustrates the application of density estimation to reliability analysis.

The examples employing simulated data are given first, followed by the examples using real data. Furthermore, several of the examples have been chosen in such a manner that they are distinctively related.

5.2 Examples

Example 5.2.1 - The first example uses the exponential density with parameters $\tau = 1$ and $\theta = 2$ to illustrate a principle discussed in Chapter II. The plots of the estimated density functions using h* for sample sizes n = 10 and n = 100, and both the type I and type II underlying densities are given in Figures 1. Notice the appearance of more protuberances in the tails of the type I estimates as the sample size

increases from n = 10 to n = 100. This is illustrated in Figure 1a. On the other hand, the type II estimates given in Figures 1b and 1c are much smoother in the sense that the bumps have been entirely eliminated. The latter two estimates are shown on separate graphs because of the difficulty in distinguishing one estimate from the other. Finally, notice the change in the rate (for both type I and type II estimates) at which the exponential estimates reach their maximum as the sample size increases.

Example 5.2.2 - This example also uses the rank kernel estimator of Chapter II. The type I and type II underlying densities were used to compute estimates for a random sample of size n = 20. The observations were generated from a normal distribution with mean and variance, ten (10) and two (2), respectively. The type I estimate is plotted in Figure 2a while the type II estimate is plotted in Figure 2b. The true probability density function has been superimposed on each to facilitate comparison of the estimated and true densities.

The polymodal tendency which is characteristic of the type I estimates is evident in Figure 2a, whereas the elimination of the polymodal effect by the type II density is noted in Figure 2b. The average square errors for the two estimates are .00337 and .000693, respectively.

Example 5.2.3 - The next example uses the iterative estimation procedure of Chapter III to estimate the exponential density function discussed in example 5.2.1 for a sample of size n = 100. In addition, we have used this estimate to illustrate the application of density estimation to reliability analysis. Using the estimated density, $f_n(y)$, in conjunction with equations (1.2.1, 1.2.3) for $\hat{R}(t)$ and $\hat{z}(t)$, we have computed estimates of both the reliability and hazard functions.

Figure 3a illustrates the similarity of the smoothness of the iterative estimate to that of the type II estimate given in Figure 1c. Each iteration has been plotted to illustrate convergence of the estimation procedure. Notice that the small bumps which are present in the first iteration have been smoothed out in the final solution. Seven iterations were required to obtain convergence.

The exponential distribution implies a constant hazard rate, $1/\theta$, and an exponential reliability function. The corresponding estimated functions are given in Figures 3c and 3b, respectively. The true hazard rate function has been superimposed on the plot of the estimated hazard rate function whereas the true density function and true reliability function have not been since the estimated and true curves are very similar.

Example 5.2.4 - The iterative estimator was used to compute the estimate for the same random sample of size n = 20 from a normal probability density function that was introduced in example 5.2.2. Figure 4 illustrates the tendency to accentuate oscillations where a concentration of data occurs and to smooth the humps at the extremes. Comparing this plot with Figure 2a, we see the more pronounced fluctuations present in the iterative estimate than in the type I estimate.

The average square error is .00833. Again the true density function has been superimposed on Figure 4.

Example 5.2.5 - The final example employing simulated data uses the spline estimator of Chapter IV. The uniform density function over the domain [0,10] was used to generate a sample of size n = 39 using true percentiles. Two fixed knots were chosen at $a_2^* = 1$ and $a_3^* = 9$ and we dispensed with the goodness of fit test. The estimated density has been plotted in Figure 5.

In the next two examples, we consider the estimation of a probability density function from real data. The data were taken from Bliss (1967) in the form of grouped data. In order to employ our estimation procedures with grouped data, we have taken the number of observations occurring within each group and assumed that they are equally spaced within the interval width of the group.

Example 5.2.6 - This example is taken from Bliss' (1967) Chapter 5 on the normal distribution. The data describes the lengths, recorded to the nearest centimeter, of 578 ears of corn in an F_3 cross of Missouri dent by Tom Thumb pop. The data was grouped into a fourteen (14) cell frequency table over the domain [10.5, 24.5] with cells of width one (1) centimeter. In Figures 6a and 6b the normalized histogram has been drawn along with the estimated densities. Figures 6a and 6b plot the estimates obtained from the iterative estimator and the spline estimator, respectively. The iterative estimation procedure required nine (9) iterations to obtain convergence using only fourteen (14) percentiles. Each of these iterations has been plotted in Figure 6a. One fixed knot at $a_2^* = \bar{y} = 17.12$ was used for the spline estimate.

Example 5.2.7 - Our final example is data given in Chapter 7 of Bliss (1967) for the lengths of survival in days of 1110 mice inoculated uniformly with malaria in genetic studies of resistance. The data is presented in the form of a frequency table over the domain [3, 33] with thirty (30) cells of width one (1) day. Again, we illustrate the use of percentiles from the sample. However, we have used only twenty-eight (28) percentiles since the first and last cell of his frequency table contain no observations. Note that the normalized histograms given in the Figures 7a and 7b are bimodal. Bliss discusses this phenomenon as representing

a mixed distribution. The cause of death in the short-lived ($\leq$ 11 days) mice is attributed to severe toxemia while the cause of death in the long-lived ($\geq$ 14 days) mice is attributed to anoxic anemia.

The iterative estimation procedure required seven (7) iterations to obtain convergence; however, only the final solution has been plotted in Figure 7a. The spline procedure used one knot at the middle of each of the two histogram cells with the greatest frequencies and also one knot at the middle of the histogram cell with the least frequency; that is, $a_2^* = 5.5$, $a_3^* = 10.5$, and $a_4^* = 20.5$. The spline estimate is plotted in Figure 7b.

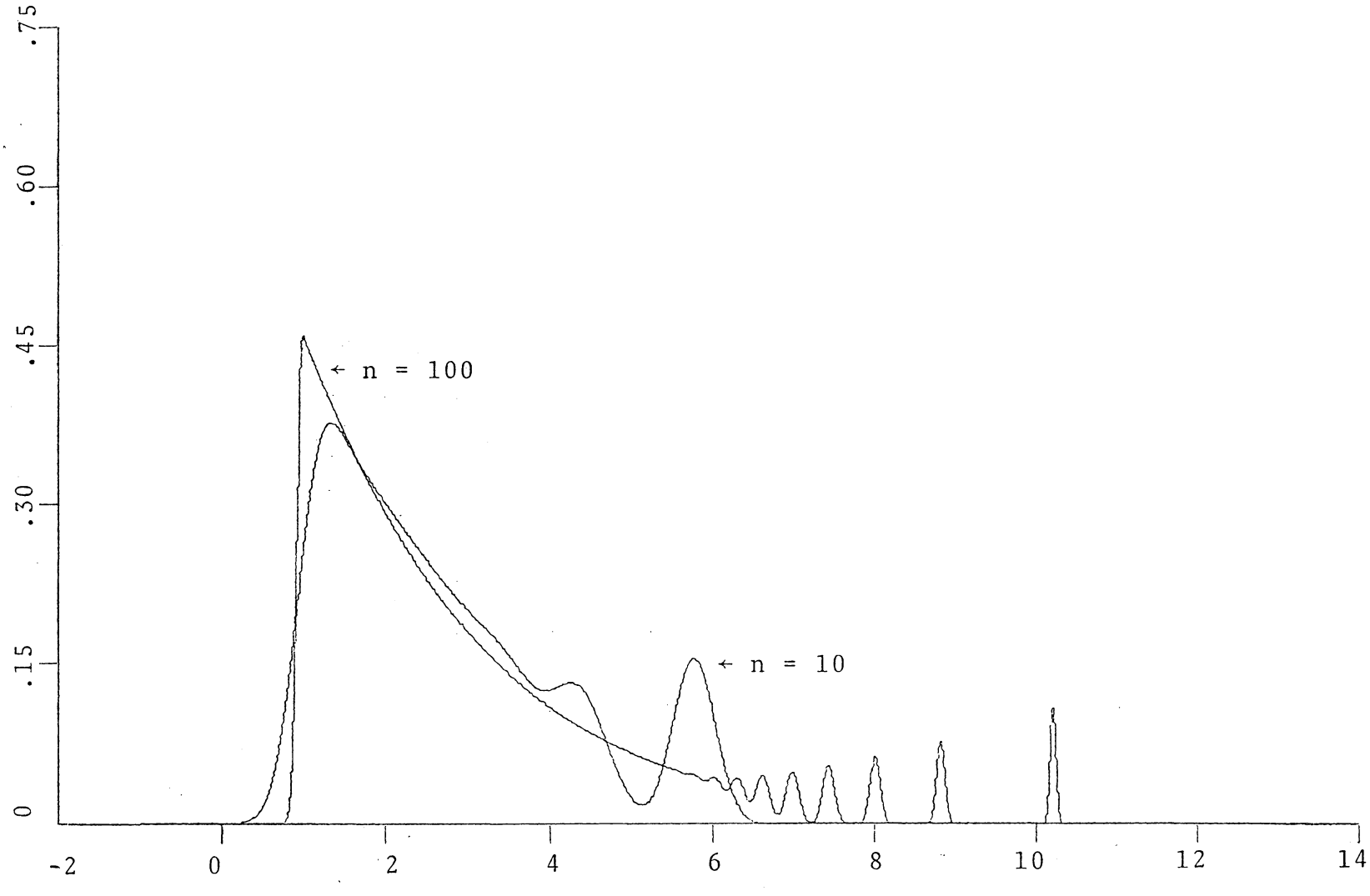

Figure 1a. Exponential Density, Type I Estimates Using True Percentiles

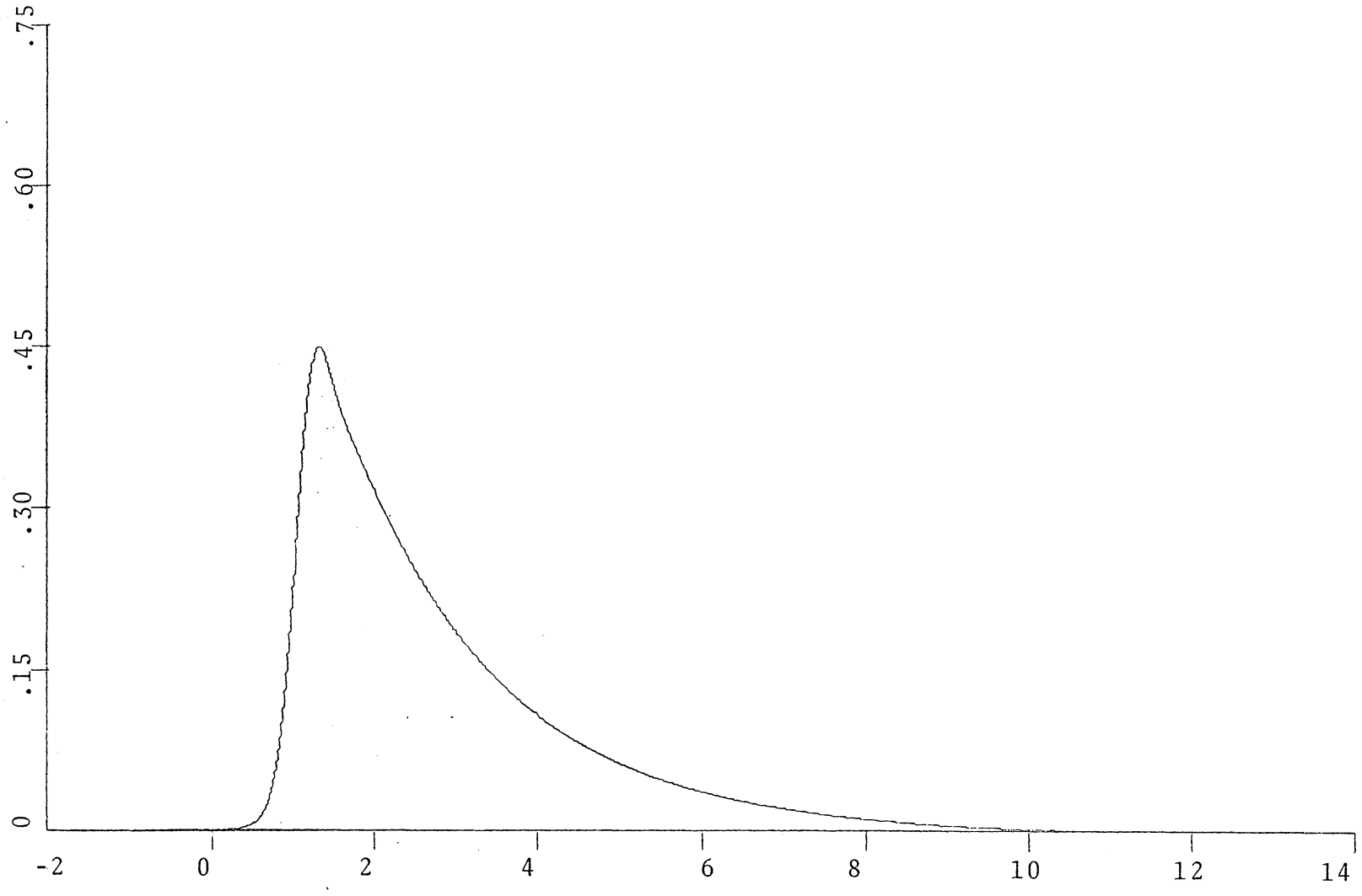

Figure 1b. Exponential Density, Type II Estimate Using True Percentiles (n = 10)

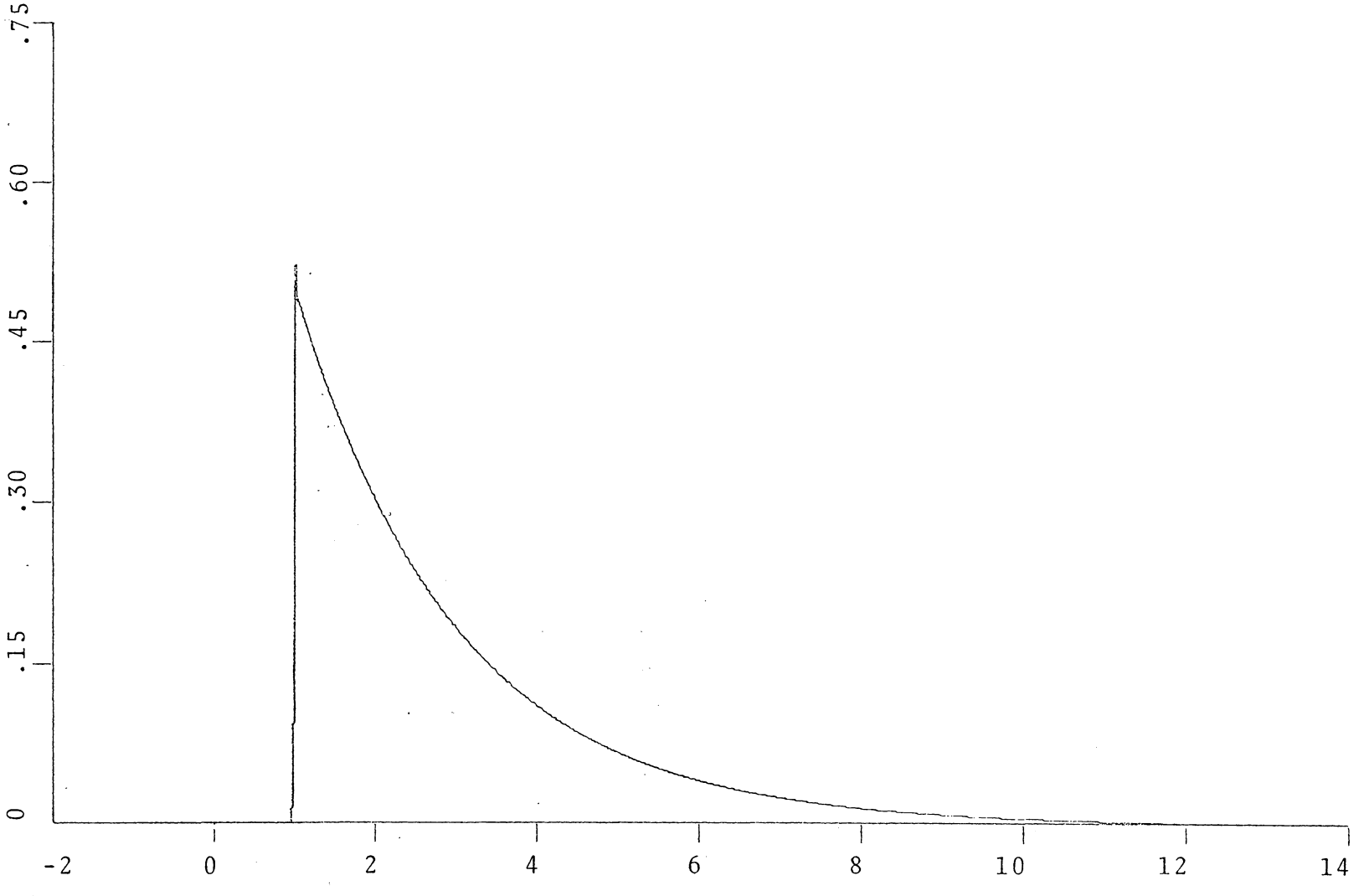

Figure 1c. Exponential Density, Type II Estimate Using True Percentiles (n = 100)

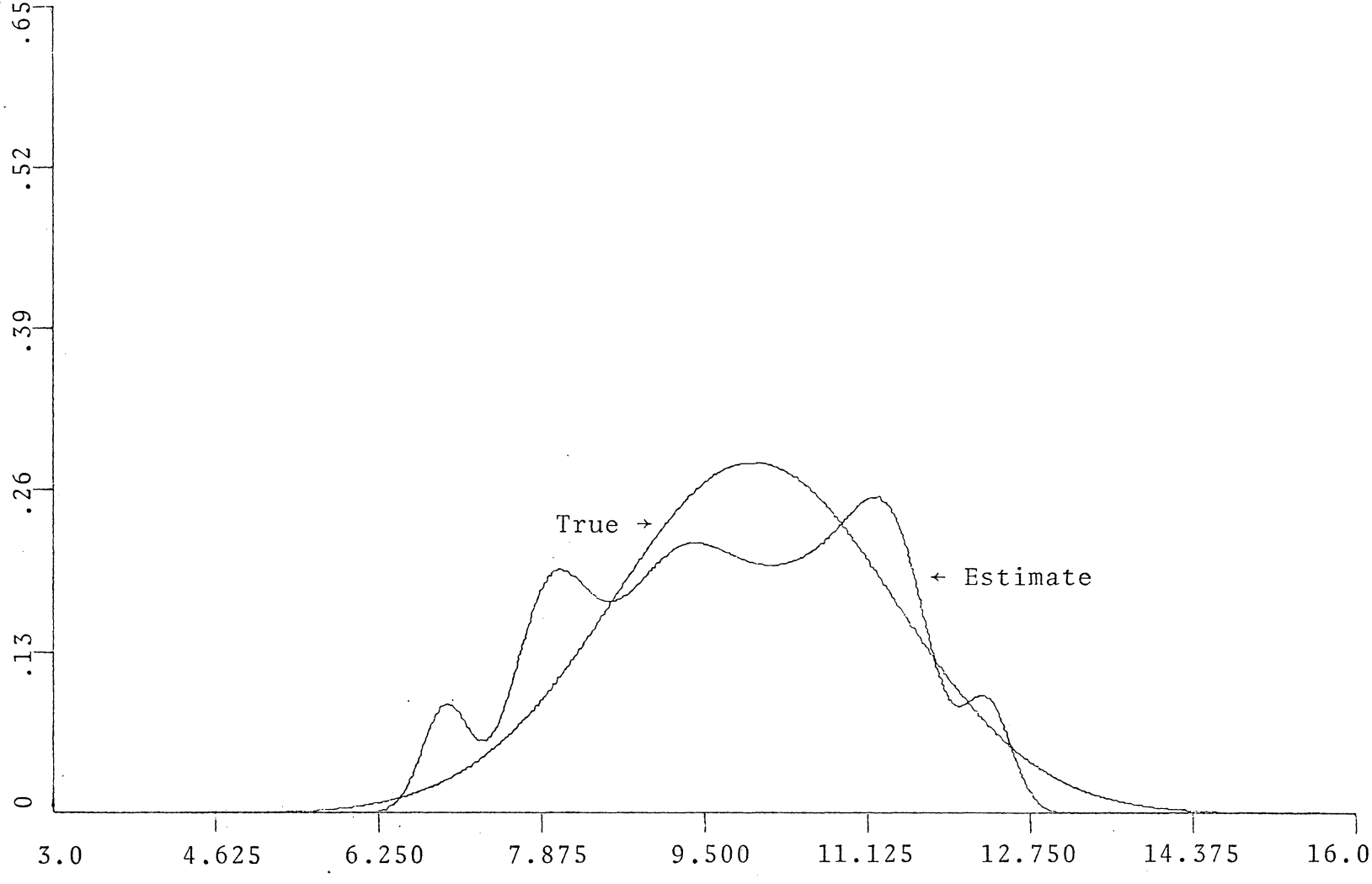

Figure 2a. Normal Density, Type I Estimate Using Random Data

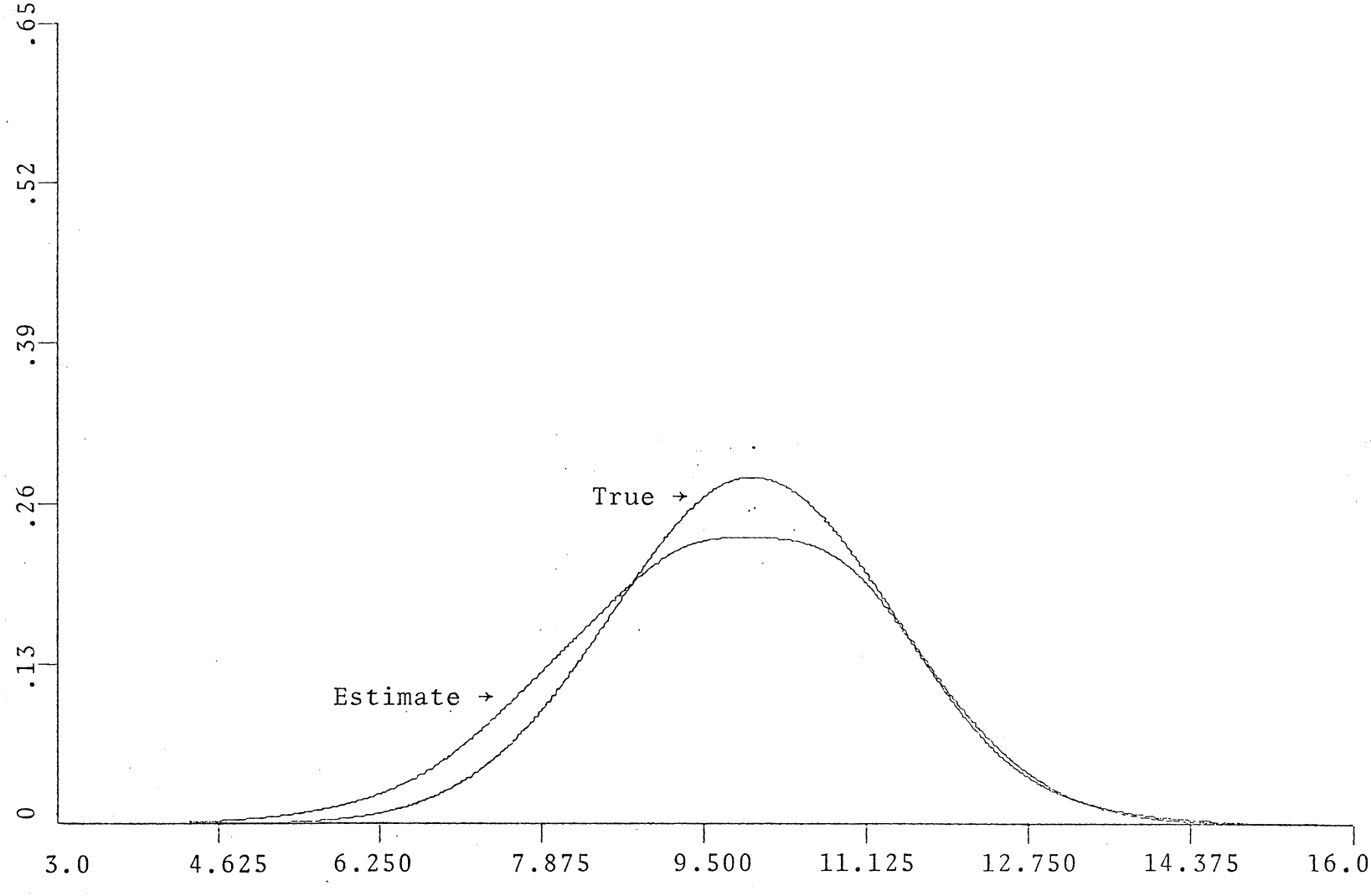

Figure 2b. Normal Density, Type II Estimate Using Random Data

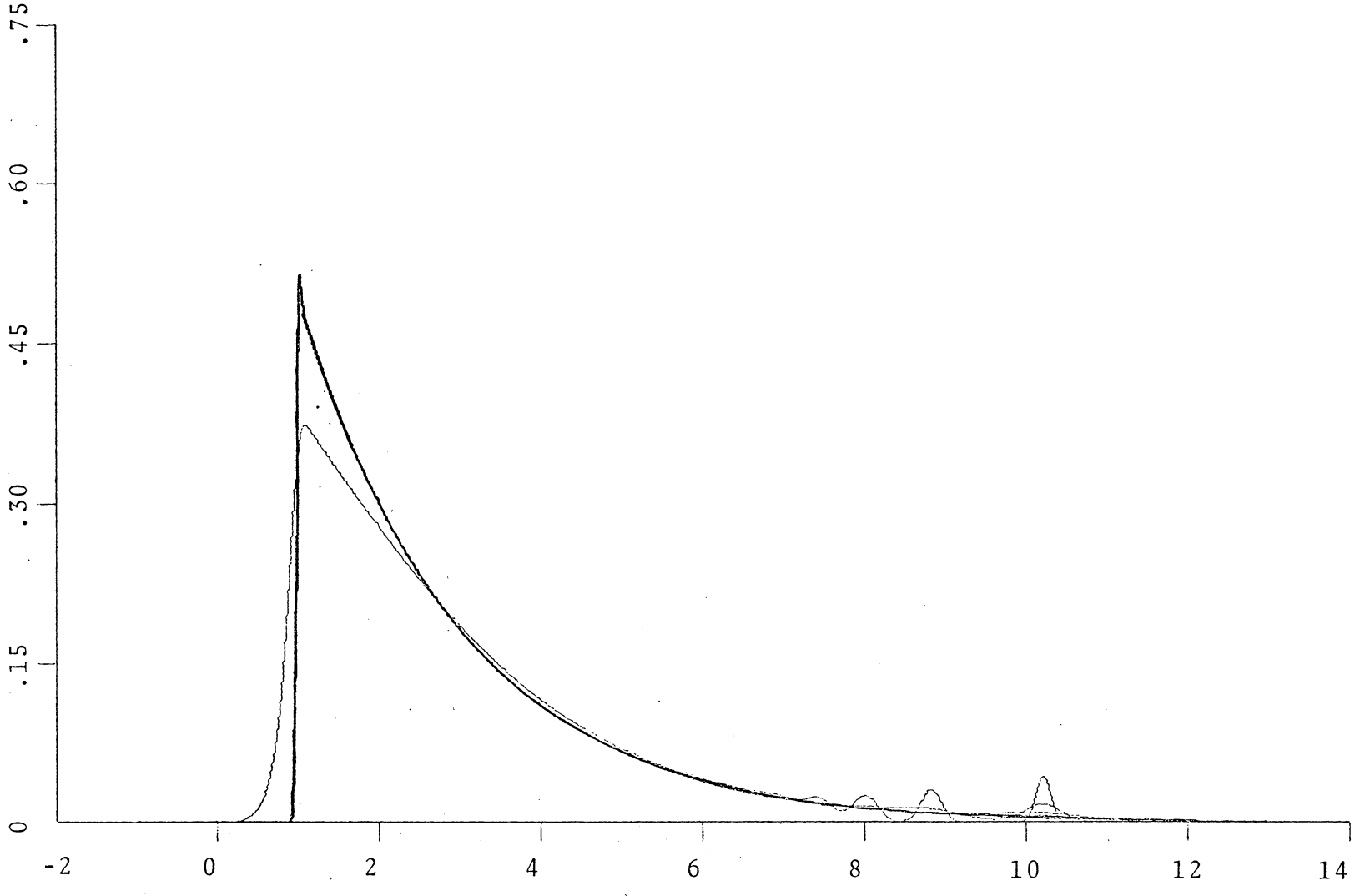

Figure 3a. Exponential Density, Iterative Estimate Using True Percentiles (n = 100)

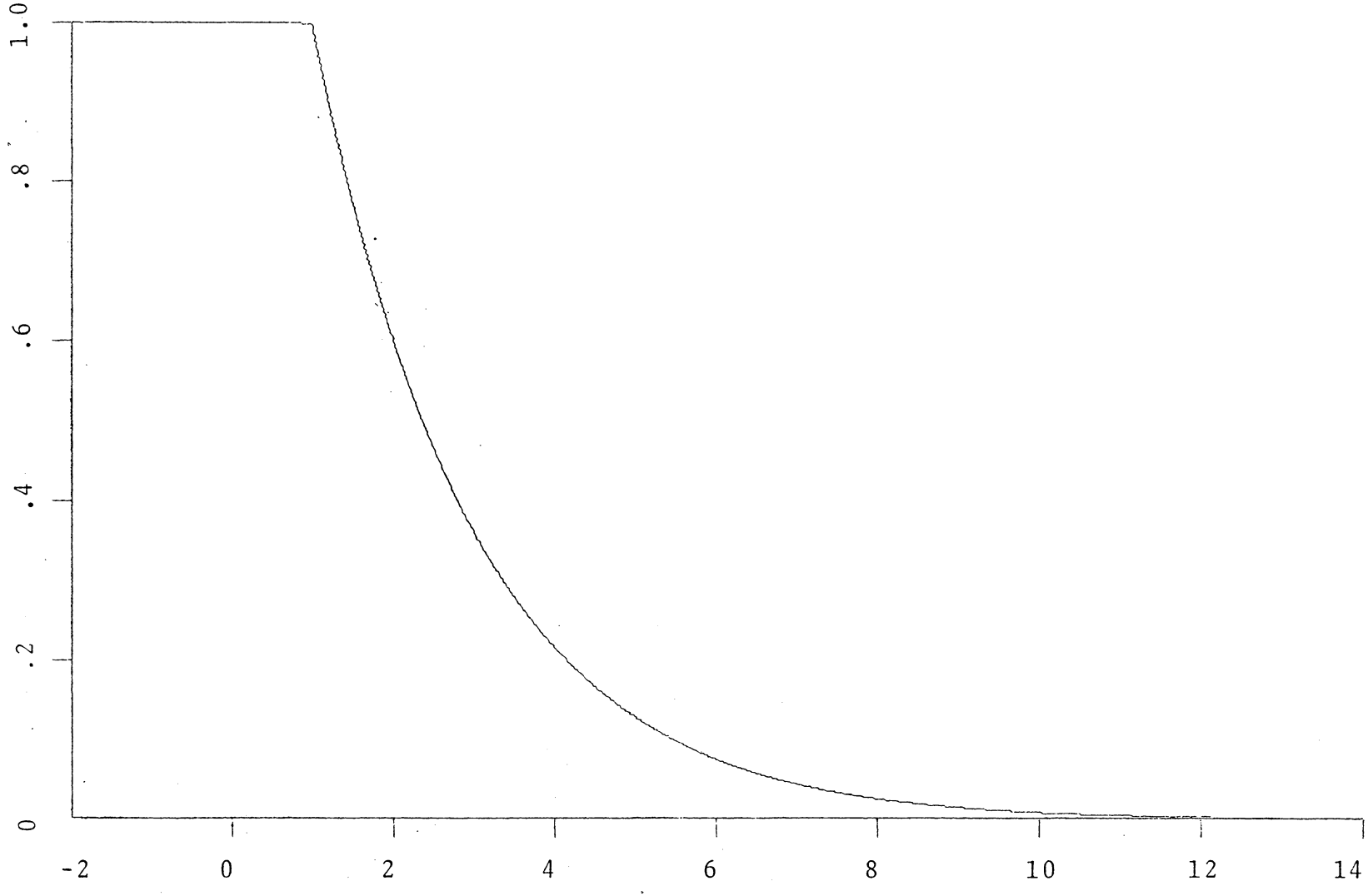

Figure 3b. Estimated Reliability Function, $\hat{R}(t)$

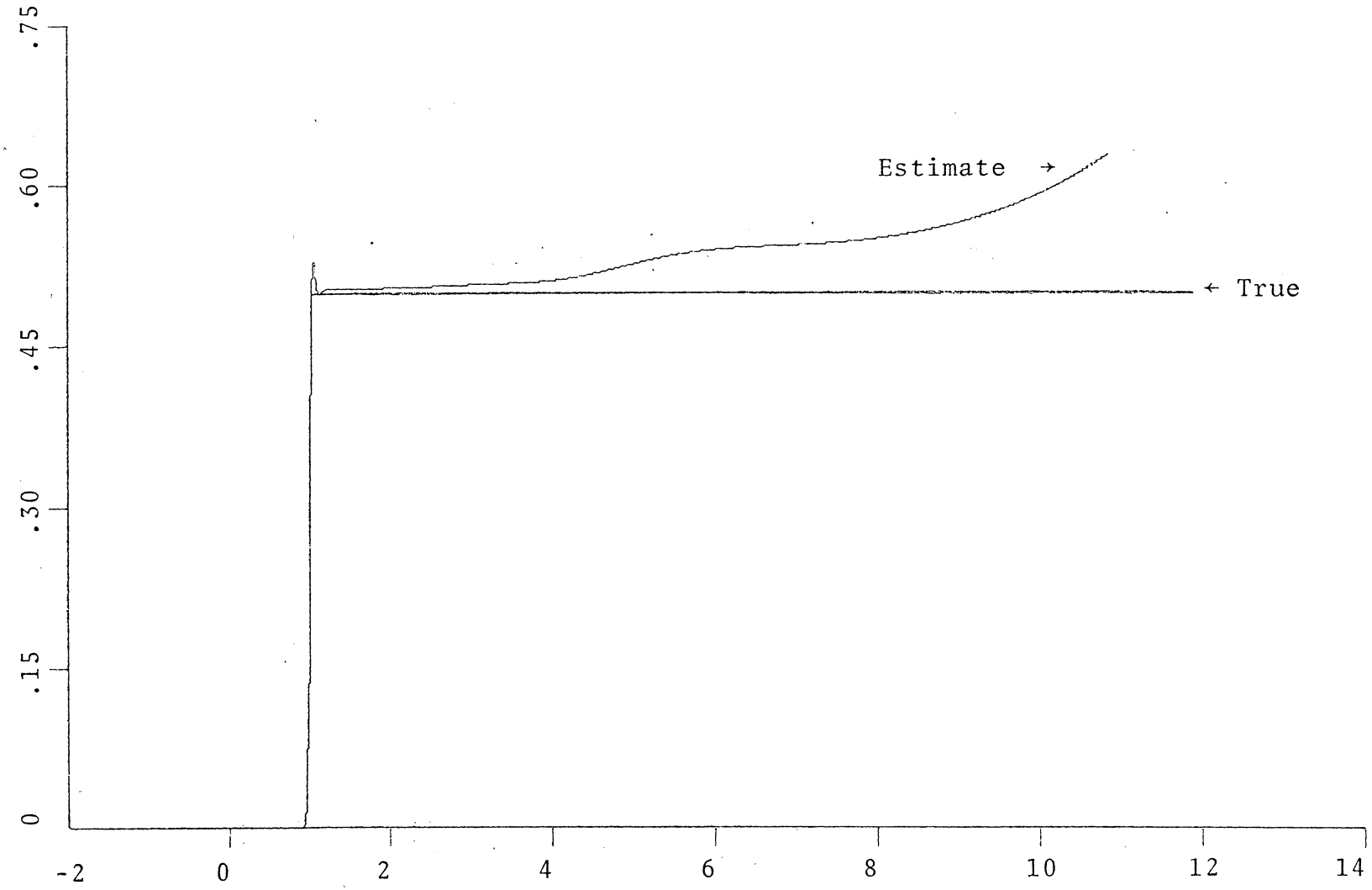

Figure 3c. Estimated Hazard Function, $\hat{z}(t)$

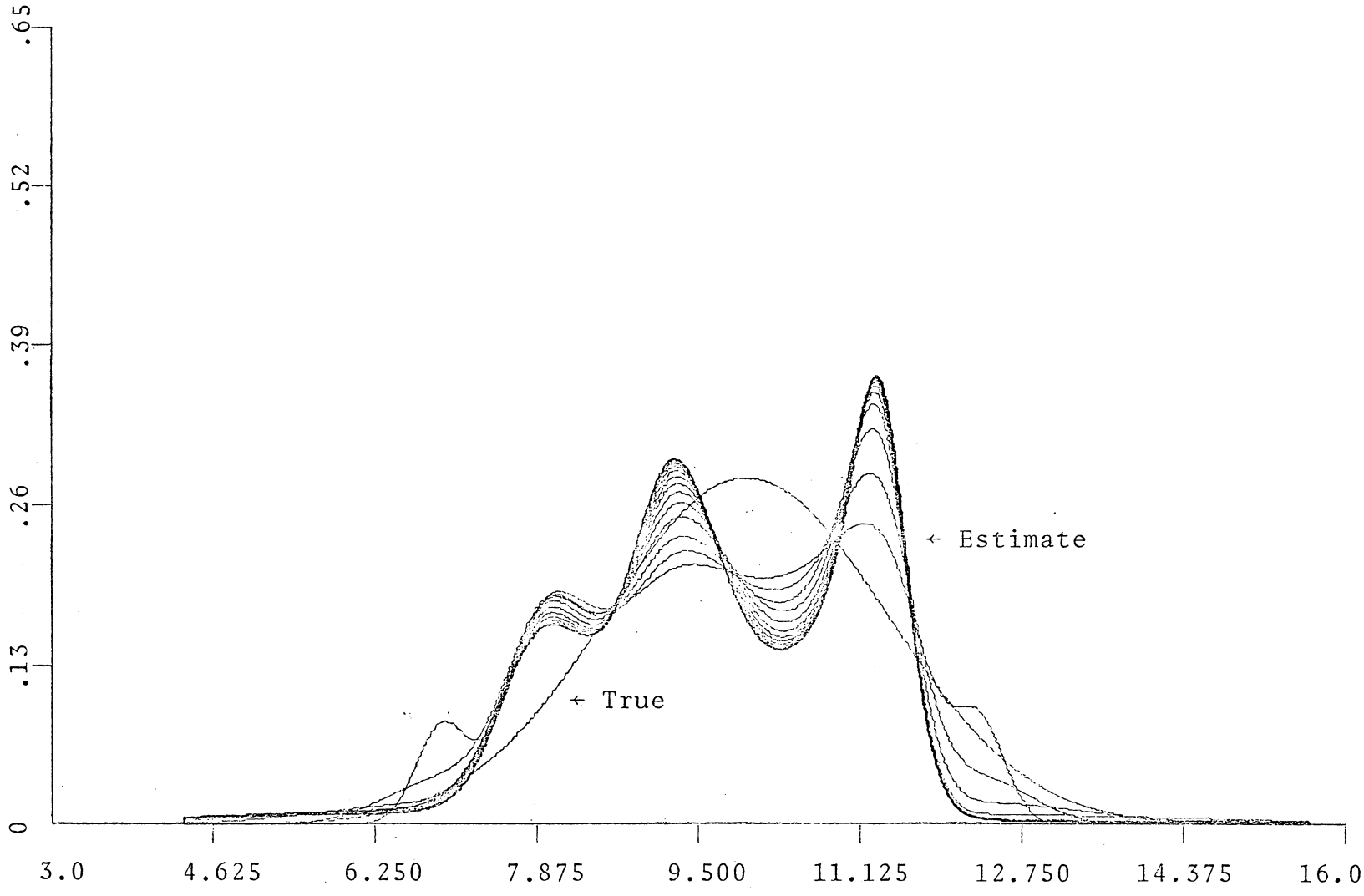

Figure 4. Normal Density, Iterative Estimate Using Random Data

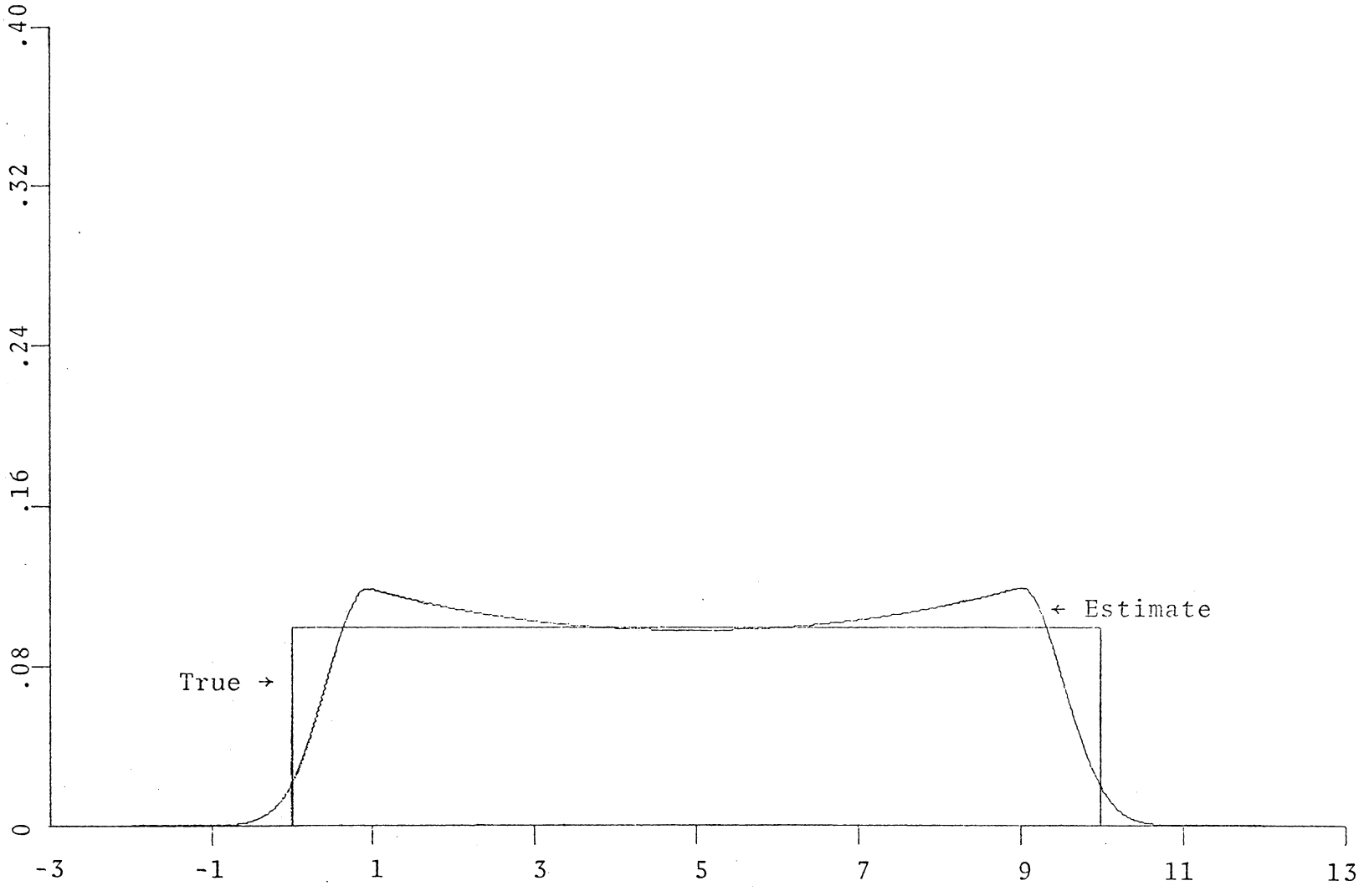

Figure 5. Uniform Density, Spline Estimate Using True Percentiles

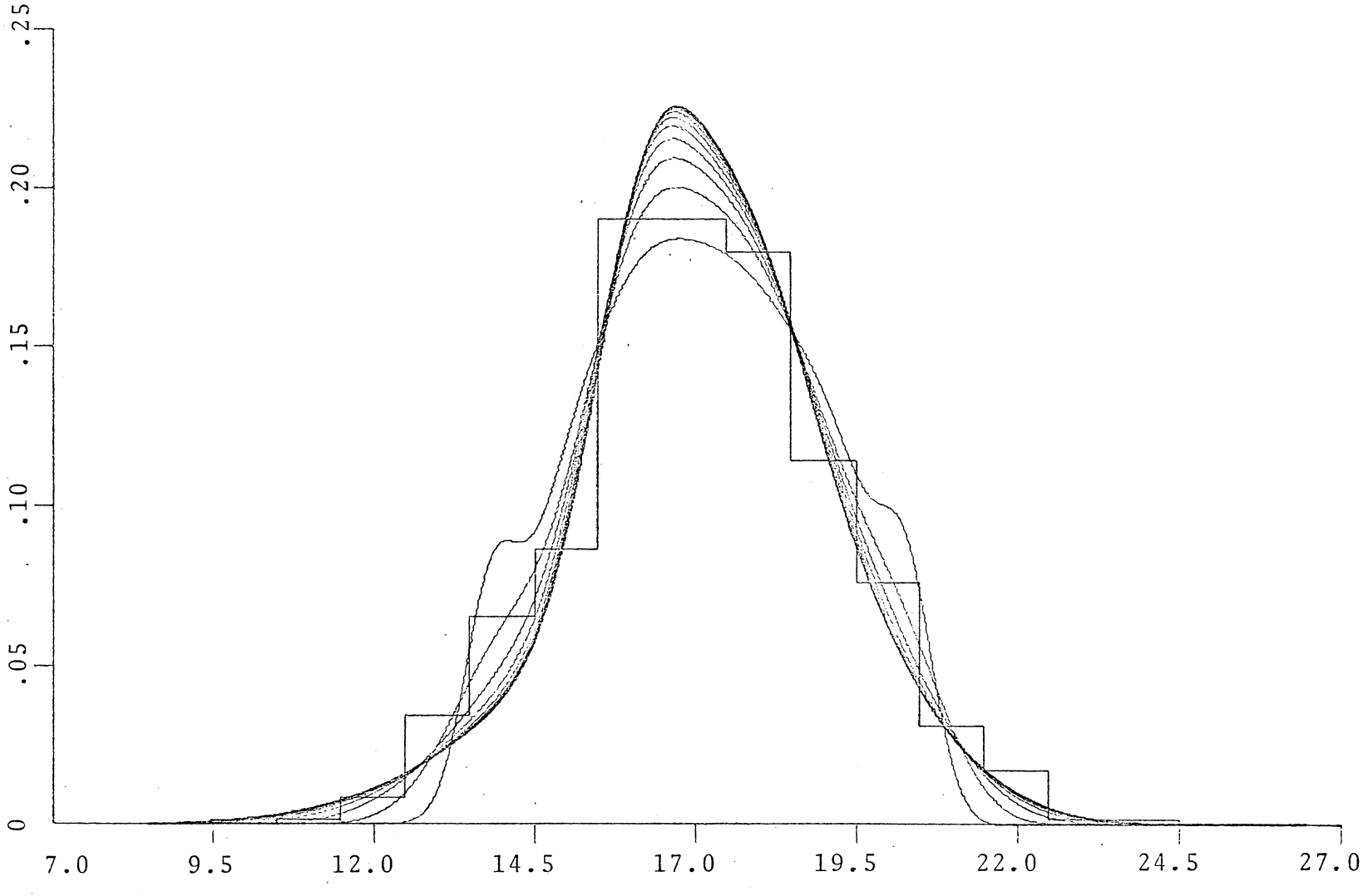

Figure 6a. Lengths of Ears of Corn, Iterative Estimate

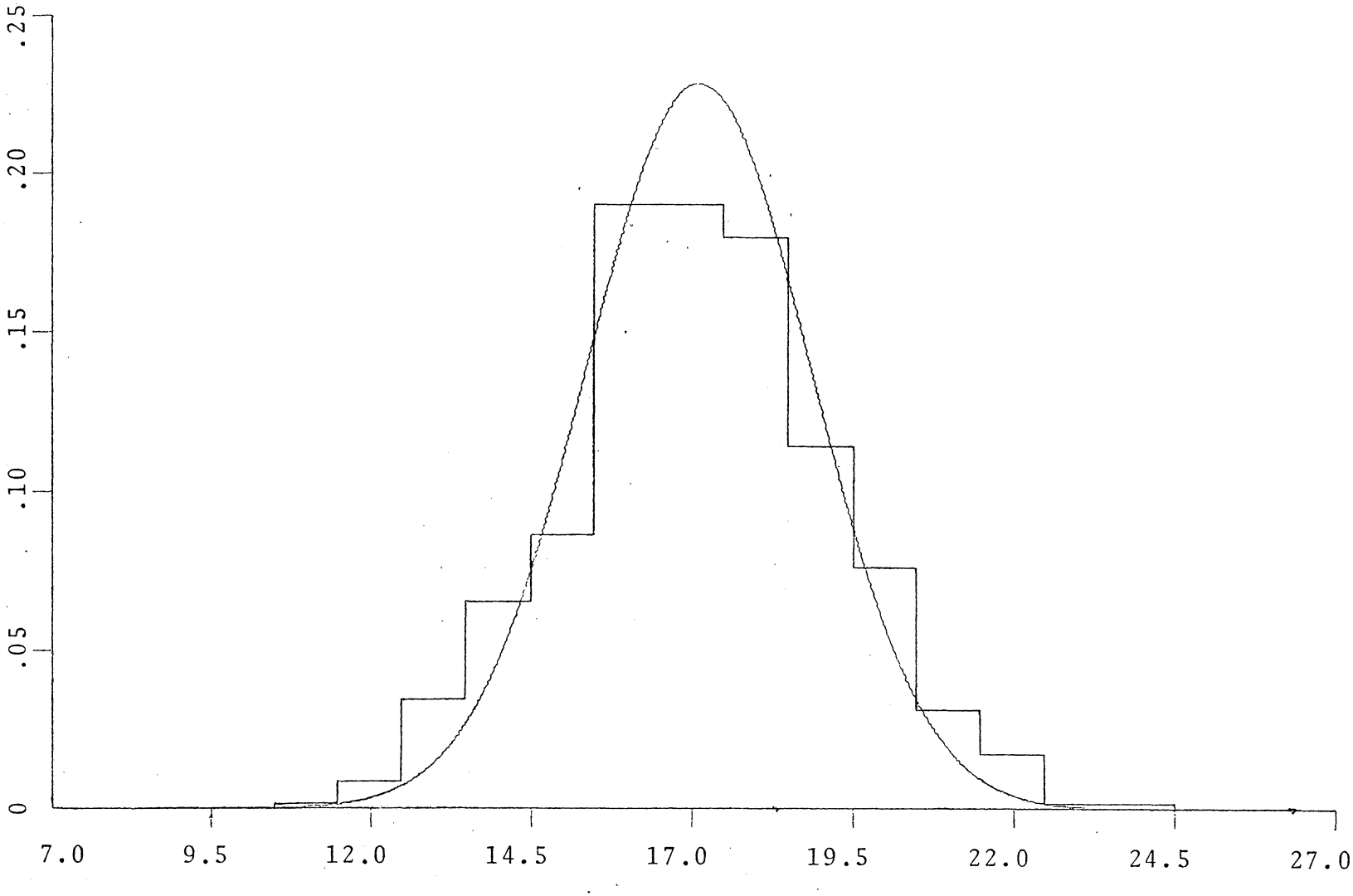

Figure 6b. Lengths of Ears of Corn, Spline Estimate

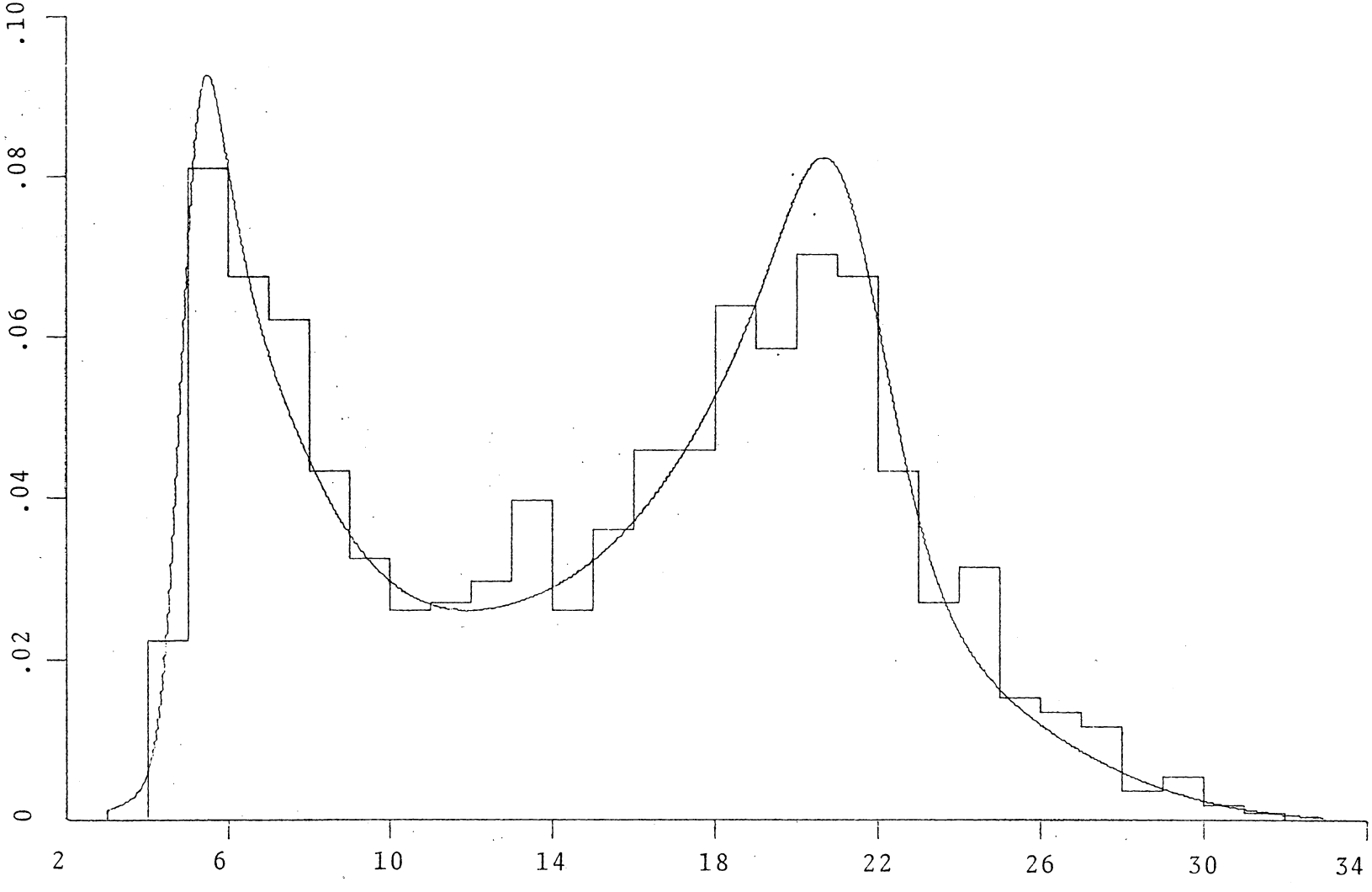

Figure 7a. Survival of Mice Inoculated With Malaria, Iterative Estimate

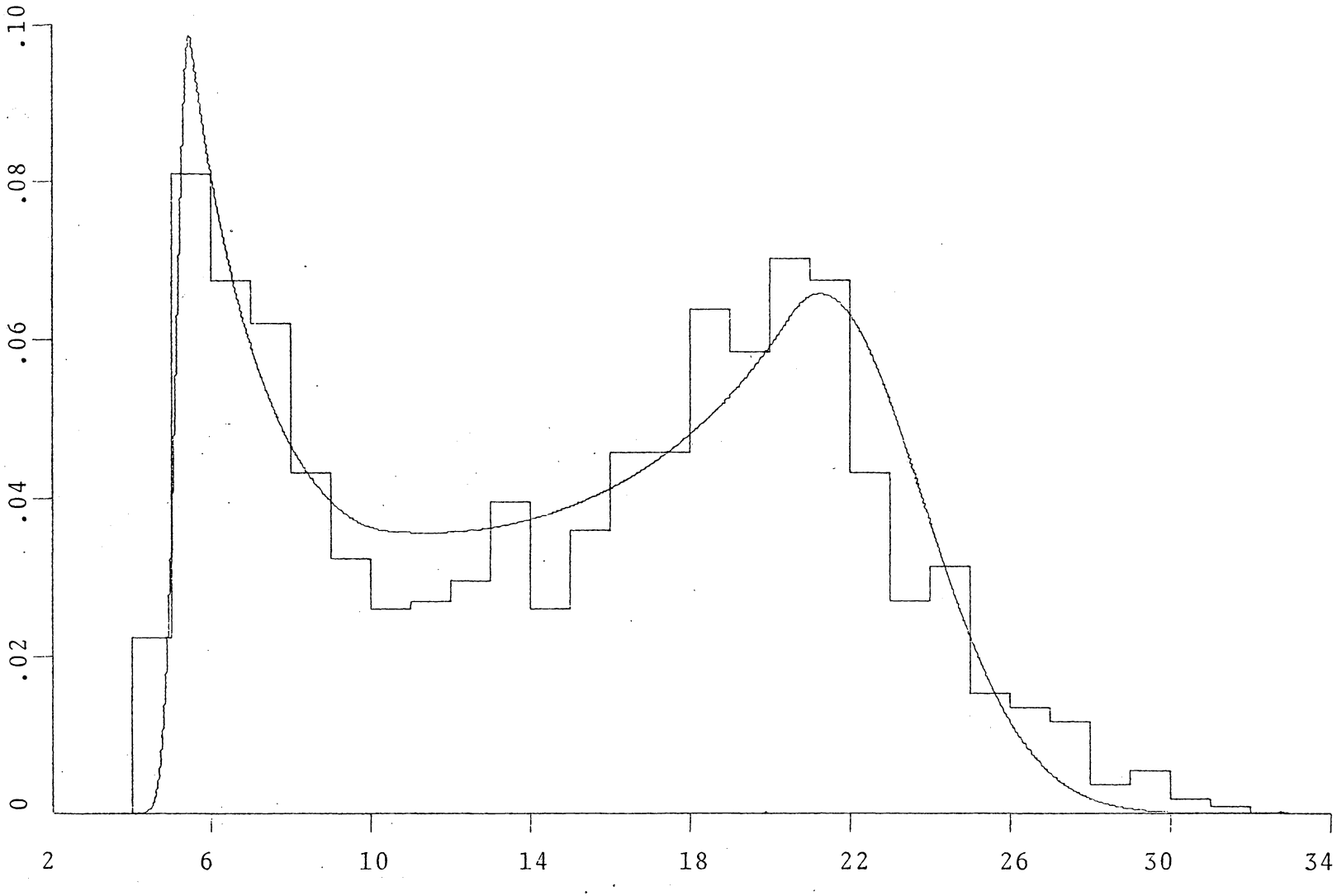

Figure 7b. Survival of Mice Inoculated With Malaria, Spline Estimate

BIBLIOGRAPHY

Ahlberg, J. H., Nilson, E. N., and Walsh, J. L. (1962). Best approximation properties of the spline fit, J. Math. Mech., 11, 225-234.

Ahlberg, J. H., Nilson, E. N., and Walsh, J. L. (1964). Fundamental properties of generalized splines, Proc. Nat. Acad. Sci. U.S.A., 52, 1412-1419.

Ahlberg, J. H., Nilson, E. N., and Walsh, J. L. (1965a). Convergence properties of generalized splines, Proc. Nat. Acad. Sci. U.S.A., 54, 344-350.

Ahlberg, J. H., Nilson, E. N., and Walsh, J. L. (1965b). Extremal, orthogonality, and convergence properties of multidimensional splines, J. Math. Anal. Appl., 11, 27-48.

Ahlberg, J. H., Nilson, E. N., and Walsh, J. L. (1965c). Best approximation and convergence properties of higher order spline approximations, J. Math. Mech., 14, 231-243.

Aizerman, M. A., Braverman, E. M., and Rozonoer, L. I. (1964a). The probability problem of pattern recognition learning and the method of potential functions, Automation and Remote Control, 25, 1175-1190 (a translation of Avtomatika i Telemekhanika).

Aizerman, M. A., Braverman, E. M., and Rozonoer, L. I. (1964b). The method of potential functions for the problem of restoring the characteristics of a function converter from randomly observed points, Automation and Remote Control, 25, 1546-1556 (a translation of Avtomatika i Telemekhanika).

Anderson, G. D. (1969). A Comparison of Methods for Estimating a Probability Density Function, Ph.D. Dissertation, University of Washington.

Bartlett, M. S. (1963). Statistical estimation of density functions, Sankhya Ser. A., 25, 245-254.

Bennett, G. K. (1970). Smooth Empirical Bayes Estimation With Application to the Weibull Distribution, Ph.D. Dissertation, Texas Tech University.

Bennett, G. K. and Martz, H. F., Jr. (1970). A Computer Analysis of Parzen's Density Estimators, unpublished manuscript, Department of Industrial Engineering, Texas Tech University.

Bennett, G. K. and Martz, H. F., Jr. (1972). A continuous empirical Bayes smoothing technique, Biometrika, 59, 361-368.

Bhattacharya, P. K. (1967). Estimation of a probability density and its derivatives, Sankhya Ser. A., 29, 373-382.

Birkhoff, G. and deBoor, C. R. (1964). Error bounds for spline interpolation, J. Math. Mech., 13, 827-835.

Birkhoff, G. and deBoor, C. R. (1965). Piecewise polynomial interpolation and approximation (in Approximation of Functions, Elsevier, New York), (editor, Garabedian, H. L.).

Birkhoff, G. and Garabedian, H. L. (1960). Smooth surface interpolation, J. Math. and Phys., 39, 258-268.

Blaydon, C. C. (1967). Approximation of distribution and density functions, Proc. IEEE., 55, 231-232.

Bliss, C. I. (1967). Statistics in Biology, Vol. I., McGraw-Hill, New York.

Boneva, L., Kendall, D. G., and Stefanov, I. (1971). Spline transformations: three new diagnostic aids for the statistical data analyst, J. Roy. Statist. Soc. Ser. B, 33, 1-70.

Cacoullos, T. (1964). Estimation of a multivariate density, Tech. Report No. 40, Department of Statistics, University of Minnesota.

Cacoullos, T. (1966). Estimation of a multivariate density, Ann. Inst. Statist. Math., 18, 179-189.

Cencov, N. N. (1962). Evaluation of an unknown distribution from observations, Soviet Math., 3, 1559-1562 (a translation of Dokl. Akad. Nauk SSSR).

Cooper, D. B. (1964). Adaptive pattern recognition and signal detection using stochastic approximation, IEEE Trans. Elec. Computers, 13, 306-307.

Craswell, K. J. (1965). Density estimation in a topological group, Ann. Math. Statist., 36, 1047-1048.

Daniels, H. E. (1962). The estimation of spectral densities, J. Roy. Statist. Soc. Ser. B, 24, 185-198.

deBoor, C. R. (1962). Bicubic spline interpolation, J. Math. and Phys., 41, 212-218.

deBoor, C. R. (1963). Best approximation properties of spline functions of odd degree, J. Math. Mech., 12, 747-749.

deBoor, C. R. and Lynch, R. E. (1966). On splines and their minimum properties, J. Math. Mech., 15, 953-969.

Elkins, T. A. (1968). Cubical and spherical estimates of multivariate probability densities, J. Amer. Statist. Assoc., 63, 1495-1513.

Epanechnikov, V. A. (1969). Nonparametric estimation of a multivariate probability density, Theor. Prob. Appl., 14, 153-158 (a translation of Teor. Verojatnost. i Primenen.).

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, 471-474.

Farrell, R. H. (1972). On the best obtainable asymptotic rates of convergence in the estimation of a density function at a point, Ann. Math. Statist., 43, 170-180.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample, Proc. Cambridge Philos. Soc., 24, 180-190.

Fix, E. and Hodges, J. L., Jr. (1951). Discriminatory analysis, nonparametric discrimination: consistency properties, Report No. 4, Project No. 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas.

Fix, E. and Hodges, J. L., Jr. (1952). Discriminatory analysis, nonparametric discrimination: small sample performances, Report No. 11, Project No. 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas.

Garabedian, H. L. (1965). (editor) Approximation of Functions, Elsevier, New York.

Gaskins, R. A. (1972). Density Estimation and Some Topics in Multivariate Analysis, Ph.D. Dissertation, Virginia Polytechnic Institute and State University.

Gessaman, M. P. (1970). A consistent nonparametric multivariate density estimator based on statistically equivalent blocks, Ann. Math. Statist., 41, 1344-1346.

Good, I. J. (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables, Ann. Math. Statist., 34, 911-934.

Good, I. J. (1971a). The probabilistic explication of information, evidence, surprise, causality, explanation, and utility (in Foundations of Statistical Inference, Holt, Rinehart and Winston, Toronto), (editors, Godambe, V. P. and Sprott, D. A.).

Good, I. J. (1971b). Nonparametric roughness penalty for probability densities, Nature Physical Science, 229, 29-30.

Good, I. J. (1971c). In discussion to article by Boneva, L., Kendall, D. G., and Stefanov, I.

Good, I. J. and Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities, Biometrika, 58, 255-277.

Good, I. J. and Gaskins, R. A. (1972). Global nonparametric estimation of probability densities, Virginia Journal of Science, 23, 171-193.

Greville, T. N. E. (1964). Interpolation by generalized spline functions, Tech. Report No. 476, Mathematics Research Center, University of Wisconsin.

Greville, T. N. E. (1967). Spline functions, interpolation and numerical quadrature (in Mathematical Methods for Digital Computers, Wiley, New York), (editors, Ralston, A. and Wilf, H. S.).

Greville, T. N. E. (1969). (editor) Theory and Applications of Spline Functions, Academic Press, New York.

Gupta, S. Das (1964). Nonparametric classification rules, Sankhya Ser. A., 26, 25-30.

Hendrickson, A. D. (1972a). Estimating densities with functions of Rosenblatt's kernel estimators, unpublished manuscript, Department of Statistics, Virginia Polytechnic Institute and State University.

Hendrickson, A. D. (1972b). Rank kernel estimation of density functions, unpublished manuscript, Department of Statistics, Virginia Polytechnic Institute and State University.

Hendrickson, A. D. (1973a). Rank kernel estimation of multivariate densities, unpublished manuscript, Department of Statistics, Virginia Polytechnic Institute and State University.

Hendrickson, A. D. (1973b). Applications of Nonparametric Density Estimation to Probability Forecasting, Proc. Third Conference on Probability and Statistics in Atmospheric Science (Boulder, Colorado, June 19-22). Boston, American Meteorological Society.

Holladay, J. C. (1957). A smoothest curve approximation, Math. Tables Aids to Computation, 11, 233-243.

Jeffreys, J. (1946). An invariant form for the prior probability in estimation problems, Proc. Roy. Soc. Ser. A, 186, 453-461.

Karlin, S. and Ziegler, Z. (1966). Chebyshevian spline functions, SIAM J. Numer. Anal., 3, 514-543.

Karlin, S. and Ziegler, Z. (1967). Chebyshevian spline functions (in Inequalities, Academic Press, New York), (editor, Shisha, O.).

Kendall, M. G. and Stuart, A. (1969). The Advanced Theory of Statistics, Griffin, London.

Kronmal, J. N. (1964). The Estimation of Probability Densities, Ph.D. Dissertation, U.C.L.A.

Kronmal, R. and Tarter, M. (1968). The estimation of probability densities and cumulatives by Fourier series methods, J. Amer. Statist. Assoc., 63, 925-952.

Leadbetter, M. R. (1963). On the nonparametric estimation of a probability density, Tech. Report No. 11, Research Triangle Institute.

Leadbetter, M. R. and Watson, C. S. (1962). On the estimation of the probability density, Tech. Report No. 3, Research Triangle Institute.

Lin, Pi-Erh (1968). Estimation of a Multivariate Density and its Partial Derivatives with Empirical Bayes Applications, Ph.D. Dissertation, Columbia University.

Loftsgaarden, D. O. and Quesenberry, C. P. (1965). A nonparametric estimate of a multivariate density function, Ann. Math. Statist., 36, 1049-1051.

Maniya, G. M. (1961). Remarks on nonparametric estimation of a bivariate probability density, Soobshch. Akad. Nauk Gruzin SSR., 27, 385-390 (in Russian).

Martz, H. F., Jr. and Hailey, M. L. (1971). A continuous nonparametric reliability estimator, AIIE Transactions, 3, 115-122.

Martz, H. F., Jr. and Krutchkoff, R. G. (1969). Empirical Bayes estimators in a multiple linear regression model, Biometrika, 56, 367-374.

Moore, D. S. and Henrichon, E. G. (1969). Uniform consistency of some estimates of a density function, Ann. Math. Statist., 40, 1499-1502.

Murthy, V. K. (1965a). Estimation of the probability density, Ann. Math. Statist., 36, 1027-1031.

Murthy, V. K. (1965b). Estimation of jumps, reliability, and hazard rate, Ann. Math. Statist., 36, 1032-1040.

Murthy, V. K. (1966). Nonparametric estimation of multivariate densities with applications (in Multivariate Analysis, Academic Press, New York),(editor, Krishnaiah, P. R.).

Nadaraya, E. A. (1963). On estimation of density functions of random variables, Soobshch. Akad. Nauk Gruzin SSR, 32, 227-280 (in Russian).

Nadaraya, E. A. (1964a). Some new estimates for distribution functions, Theor. Prob. Appl., 9, 497-500 (a translation of Teor. Verojatnost. i Primenen.).

Nadaraya, E. A. (1964b). Estimation of a bivariate probability density, Soobshch. Akad. Nauk Gruzin SSR, 36, 267-268 (in Russian).

Nadaraya, E. A. (1964c). On estimating regression, Theor. Prob. Appl., 9, 141-142 (a translation of Teor. Verojatnost. i Primenen.).

Nadaraya, E. A. (1965). On nonparametric estimates of density functions and regression curves, Theor. Prob. Appl., 10, 186-190 (a translation of Teor. Verojatnost. i Primenen.).

Nadaraya, E. A. (1970). Remarks on nonparametric estimates for density functions and regression curves, Theor. Prob. Appl., 15, 134-137 (a translation of Teor. Verojatnost. i Primenen.).

Parzen, E. (1961). Mathematical considerations in the estimation of spectra, Technometrics, 3, 167-190.

Parzen, E. (1962). On estimation of a probability density function and mode, Ann. Math. Statist., 33, 1065-1076.

Pelto, C. R. (1969). Adaptive nonparametric classification, Technometrics, 11, 775-792.

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

Priestley, M. B. (1962a). The analysis of stationary processes with mixed spectra I, J. Roy. Statist. Soc. Ser. B, 24, 215-233.

Priestley, M. B. (1962b). The analysis of stationary processes with mixed spectra II, J. Roy. Statist. Soc. Ser. B, 24, 511-529.

Rao, B. L. S. P. (1969). Estimation of a unimodal density, Sankhya Ser. A., 31, 26-36.

Rao, C. R. (1965). Linear Statistical Inference and Its Applications, Wiley, New York.

Reinsch, C. H. (1967). Smoothing by spline functions, Numer. Math., 10, 177-183.

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

Robertson, T., Cryer, J. D. and Hogg, R. V. (1968). On nonparametric estimation of distributions and their modes, unpublished manuscript, Department of Statistics, University of Iowa.

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

Rosenblatt, M. (1969). Conditional probability density and regression estimators (in Multivariate Analysis, Vol. 2, Academic Press, New York),(editor, Krishnaiah, P. R.).

Rosenblatt, M. (1970). Density estimates and Markov sequences, (in Nonparametric Techniques in Statistical Inference, University Press, Cambridge), (editor, Puri, M.).

Rosenblatt, M. (1971). Curve estimates, Ann. Math. Statist., 42, 1815-1842.

Royal Statistical Society (1957). Symposium on spectral approach to time series, J. Roy. Statist. Soc. Ser. B, 19, 1-63.

Sard, A. and Weintraub, S. (1971). A Book of Splines, Wiley, New York.

Schoenberg, I. J. (1946). Contributions to the problem of approximation of equidistant data by analytic functions. Part A - On the problem of smoothing or graduation. A first class of analytic approximation formulae, Quart. Appl. Math., 4, 45-99. Part B - On the problem of osculatory interpolation. A second class of analytic approximation formulae, ibid., 112-141.

Schoenberg, I. J. (1964a). Spline interpolation and best quadrature formulae, Bull. Amer. Math. Soc., 70, 143-148.

Schoenberg, I. J. (1964b). On trigonometric spline interpolation, J. Math. Mech., 13, 795-825.

Schoenberg, I. J. (1969). (editor) Approximations With Special Emphasis on Spline Functions, Academic Press, New York.

Schoenberg, I. J. and Whitney, A. (1953). On Polya frequency functions III: the positivity of translation determinants with an application to the interpolation problem by spline curves, Trans. Amer. Math. Soc., 74, 246-259.

Schuster, E. F. (1969). Estimation of a probability density and its derivatives, Ann. Math. Statist., 40, 1187-1195.

Schuster, E. F. (1970). Note on the uniform convergence of density estimates, Ann. Math. Statist., 41, 1347-1348.

Schuster, E. F. (1972). Joint asymptotic distribution of the estimated regression function at a finite number of distinct points, Ann. Math. Statist., 43, 84-88.

Schwartz, S. C. (1967). Estimation of probability densities by an orthogonal series, Ann. Math. Statist., 38, 1261-1265.

Schwartz, S. C. (1969). On the estimation of a Gaussian convolution probability density, SIAM J. Appl. Math., 17, 447-453.

Schwartz, R. B., Seltzer, S. M. and Stehle, F. N. (1965). Failure distribution analysis, Annals of Reliability and Maintainability, 4, 817-838.

Seheult, A. H. and Quesenberry, C. P. (1971). On unbiased estimation of density functions, Ann. Math. Statist., 42, 1434-1438.

Shinozuka, M. and Nishimura, A. (1965). On general representation of a density function, Annals of Reliability and Maintainability, 4, 897-903.

Shooman, M. L. (1968). Probabilistic Reliability: An Engineering Approach, McGraw-Hill, New York.

Specht, D. F. (1971). Series estimation of a probability density function, Technometrics, 13, 409-424.

Stoller, D. C. (1954). Univariate two-population distribution-free discrimination, J. Amer. Statist. Assoc., 49, 770-775.

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., 22, 511-519.

Tarter, M. E. and Kronmal, R. A. (1970). On multivariate density estimates based on orthogonal expansions, Ann. Math. Statist., 41, 718-722.

Taylor, J. R. and Lochner, R. H. (1965). Statistical analysis of field data, Annals of Reliability and Maintainability, 4, 905-913.

Van Ryzin, J. (1966). Bayes risk consistency of classification procedures using density estimation, Sankhya Ser. A, 28, 261-270.

Van Ryzin, J. (1969). On strong consistency of density estimation, Ann. Math. Statist., 40, 1765-1772.

Van Ryzin, J. (1970). On a histogram method of density estimation, Tech. Report No. 226, Department of Statistics, University of Wisconsin.

Wahba, G. (1971). A polynomial algorithm for density estimation, Ann. Math. Statist., 42, 1870-1886.

Watson, G. S. (1969). Density estimation by orthogonal series, Ann. Math. Statist., 40, 1496-1498.

Watson, G. S. and Leadbetter, M. R. (1963). On the estimation of the probability density I, Ann. Math. Statist., 34, 480-491.

Watson, G. S. and Leadbetter, M. R. (1964a). Hazard analysis I, Biometrika, 51, 175-184.

Watson, G. S. and Leadbetter, M. R. (1964b). Hazard analysis II, Sankhya Ser. A., 26, 101-116.

Wegman, E. J. (1969a). A note on estimating a unimodal density, Ann. Math. Statist., 40, 1661-1667.

Wegman, E. J. (1969b). Maximum likelihood histograms, Institute of Statistics Mimeo Series No. 629, University of North Carolina.

Wegman, E. J. (1969c). Nonparametric probability density estimation, Institute of Statistics Mimeo Series No. 638, University of North Carolina.

Wegman, E. J. (1970a). Maximum likelihood estimation of a unimodal density function, Ann. Math. Statist., 41, 457-471.

Wegman, E. J. (1970b). Maximum likelihood estimation of a unimodal density II, Ann. Math. Statist., 41, 2169-2174.

Wegman, E. J. (1972a). Nonparametric probability density estimation: I. A summary of available methods, Technometrics, 14, 533-546.

Wegman, E. J. (1972b). Nonparametric probability density estimation: II. A comparison of density estimation methods, J. Statist. Comput. Simul., 1, 225-245.

Weiss, L. and Wolfowitz, J. (1967). Estimation of a density at a point, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete., 7, 327-335.

Whittle, P. (1957). Curve and periodogram smoothing, J. Roy. Statist. Soc. Ser. B, 19, 38-47.

Whittle, P. (1958). On the smoothing of probability density functions, J. Roy. Statist. Soc. Ser. B, 20, 334-343.

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

Woodroofe, M. (1970). On choosing a delta sequence, Ann. Math. Statist., 41, 1665-1671.

APPENDIX A

PROGRAM K-ITER

The computer program contained in Appendix A has been entitled "K-Iter" - "K" for kernel and "Iter" for iterative. The program has a dual purpose in that by proper choice of an input parameter, the program will be executed either for the rank kernel estimator of Chapter II or for the iterative estimator of Chapter III. Several statements within the program have been coded on the far right hand side, corresponding to columns 73 - 80, with the codes ALL, W and IT. These will be explained as necessary.

The program is written in Fortran and involves several plotting subprograms (SCALE, PLOT, LINE, and AXIS). The IBM 1627 Digital Plotter (Calcomp) was used to produce the figures presented in Chapter V.

The required input data stream contains the sample size on the first card (Format #100) and the values of the program control parameters on the second card (Format #100) while subsequent, if any, data cards (Format #102) contain the sample observations. If the program is to generate the sample observations, only the first two data cards are required. The control variables and their options are described below:

IPARM = code for true distributions; equals eight (8) if the true density function is unknown (real data);

ICO(1) = code for simulated data; equals zero (0) for random data and one (1) for true percentiles;

ICO(2) = code for type of estimator; equals zero (0) if the program is to be executed for the rank kernel estimator and one (1) for the iterative estimator;

ICO(3) = code for estimated cumulative distribution function; equals zero (0) for no plot and one (1) for plot;

ICO(4) = code for sample cumulative distrubition function; equals zero (0) for no plot and one (1) for plot;

ICO(5) = code for reliability analysis; equals zero (0) for no plots and one (1) for plots;

ICO(6) = code for histogram; equals zero (0) for no plot and one (1) for plot;

ICO(7) = code for sample percentile option; equals zero (0) if estimator is to use the entire sample and one (1) if estimator is to use the sample percentiles;

NUM = number of sample percentiles (if ICO(7) = 1) and/or number of histogram cells (if ICO(6) = 1);

YS = code for plotting of estimated density function for the iterative estimator (ICO(2) = 1); equals YES if desire plot of each iteration.

An example of a data card containing the control parameters is given below:

Column	Format	Code
1-3	I3	8
4	I1	0
5	I1	1
6	I1	0
7	I1	0
8	I1	1
9	I1	1
10	I1	0
11-13	I3	10
14-16	A3	N0

In this example, the options indicate that the true density is unknown, the data is random, and the iterative estimation procedure is to be employed. Also, the plots of the reliability analysis and the ten (10) cell histogram have been requested. The entire sample is to be used and only the final solution is to be plotted.

The description of other important variables and of the IBM subroutines used in the program are given next according to the order of appearance.

IX = seed used for producing random data from IBM subroutine RANDU; must be odd number containing no more than nine digits;

SVE2 = value exceeding expected maximum of the estimated and true probability densities; contains code ALL1-1;

SVE3 = value exceeding expected maximum of the estimated and true hazard functions; contains code ALL1-2.

The last two variables are used so that both the corresponding estimated and true functions (when known) will be plotted on the same scale. Furthermore, if the iterative estimator is used, then SVE2 is used so that when plotting each iteration, they will be on the same scale. If the true functions are not known or if only the final solution is to be plotted, then these values, SVE2 and SVE3, are not critical.

N = sample size;

RHO, TAL, THETA = correspond to the parameters ρ, τ, and θ of the weibull distribution (see Table V); coded W1-2 through W1-4; the remaining variables following RXM in the COMMON statement are used as parameters of the other four densities.

The right hand code W represents the weibull distribution. The cards coded W1-1 through W1-5 define the parameters of the simulated weibull distribution, while the cards coded W2-1 and W2-2 compute the sample observations Y_i. Each of the five sets of cards containing W codes must be replaced by the corresponding sets of cards for the particular true distribution being simulated. These principles are illustrated by the computer program "E-Spline" presented in Appendix B. Note how the five sets of W codes have been replaced by the sets of U codes (for uniform density). Similar substitutions are made for any choice of a probability

density function that one desires to simulate. The card coded ALL2-1 is used to produce the data representing the true percentiles.

RANDU = IBM subroutine which generates uniform random numbers on the interval (0,1); coded ALL2-2;

X = array containing the observations;

A,B = minimum (maximum) value of abscissa for which the true density is computed; coded W3-2 (W3-3);

A1,B1 = minimum (maximum) value of abscissa for which the estimated density is computed; coded W3-4 (W3-5).

If no distribution is being simulated but real data is being used, then the only cards that need to be inserted (after removing existing W coded cards) are the two cards corresponding to the cards coded W3-4 and W3-5 unless the limits $\bar{y} \pm 5s$ are permissible. However, often these limits will be much greater than needed for support of the estimated density. Consequently, the histogram of the data may yield better limits for A1 and B1 than these calculated internally. Also, if the observations are not generated by the program but are "read in", then the loop Do 4 should be replaced by the statement

```
READ (5, 102) (X(I), I = 1,N) .
```

XMIN,DX = where X may represent Y, G, R, F, W, C, or X; used to check if corresponding plots of estimated and true functions or each iteration are on the same scale; XMIN is the minimum value of either the abscissa or ordinate and DX is the number of units per inch;

D = array corresponding to $d(\xi_p)$ or $f_n^m(x_i)$ as the case may be; card coded IT1-1 defines initial values of either $d(\xi_p)$ or $f_n^o(x_i)$;

H = corresponds to h in rank kernel estimator; equals one (1) for iterative estimator; coded IT1-2.

The cards coded IT1-1 and IT1-2 need to be changed to represent the required underlying density and value of h. For example, a weibull type II estimate would require replacement of the card coded IT1-1 by the following cards:

```
YFL = FLOAT (I)/(XN+1.)
 XX = TAL + (-THETA*ALOG(1.-YFL))**RXM
D(I) = GX(XX).
```

Similar replacements exist for other distributions.

JP = corresponds to iteration counter m in the iterative estimator; controls termination of the program if the number of iterations required to obtain convergence exceeds 250;

E = array corresponding to $f_n^m(x_i)$ used to test for convergence of the iterative estimation procedure;

AREA = area of estimated density function, estimated by rectangles;

FCT = function subprogram corresponding to equations (2.2.2, 3.1.1);

QATR = IBM subprogram which integrates a function.

We have tried to present sufficient details for use of this computer program without extreme explanations. The remaining variables and subprograms are self-explanatory or can be figured out with some thought.

```
      COMMON Y(6003),G(6003),F(6003),R(6003),
     1       D(200),X(202),E(200),
     2       CDF(202),ICO(7),YMIN,DY,IB,ID,
     3       N,PI,H,XN,XHN,SVE1,SVE2,SVE3,
     4       RHO,TAL,THETA,RXM,UM,VAR,VARS,
     5       ALPHA,SALPHA,SIG,BETA
      DATA YES/'YES'/
      IX=13795
      PI=3.141593
      SVE2=.75                                              ALL1-1
      SVE3=5.0                                              ALL1-2
      READ(5,100) N
      N1=N-1
      XN=N
      READ(5,100) IPARM,(ICO(I),I=1,7),NUM,YS
      WRITE(6,101) IPARM,ICO(1),ICO(2),YS,
     1             (ICO(I),I=3,7),NUM
C     PARAMETERS OF SIMULATED TRUE DENSITY
C     WEIBULL       (IPARM=1)                               W1-1
      RHO=2.0                                               W1-2
      TAL=1.0                                               W1-3
      THETA=3.0                                             W1-4
      RXM=1./RHO                                            W1-5
      DO 4 I=1,N
C     ICO(1)=0   RANDOM DATA
C           =1   UNIFORM CDF DATA
      IF(ICO(1)) 37,2,1
    1 YFL=FLOAT(I)/(XN+1.)                                  ALL2-1
      GO TO 3
    2 CALL RANDU(IX,IY,YFL)                                 ALL2-2
      IX=IY                                                 ALL2-3
C     WEIBULL       (IPARM=1)                               W2-1
    3 X(I)=TAL + (-THETA*ALOG(1.-YFL))**RXM                 W2-2
    4 CONTINUE
      DO 5 I=1,N1
      DO 5 J=I,N
      IF(X(I).LE.X(J)) GO TO 5
      TEMP=X(J)
      X(J)=X(I)
      X(I)=TEMP
    5 CONTINUE
      WRITE(6,103)  N, (X(I),I=1,N)
C     ICO(6)=0   DO NOT PLOT HISTOGRAM
C           =1   PLOT HISTOGRAM
      IF(ICO(6) .EQ. 1) CALL HIST(NUM)
```

```
C       ICO(7)=0   USE ENTIRE DATA
C              =1   USE SAMPLE PERCENTILES
        IF(ICO(7) .EQ. 1) CALL SAMP(NUM)
        S2=0.
        XBAR=0.
        DO 6 I=1,N
        CDF(I)=FLOAT(I)/XN
        S2= S2 + X(I)**2
    6   XBAR=XBAR + X(I)
        S2= (S2 - XBAR**2/XN)/XN
        XBAR=XBAR/ XN
        A1= XBAR - 5.*SQRT(S2)
        B1= XBAR + 5.*SQRT(S2)
        WRITE(6,104) XBAR,S2,A1,B1
C       WEIBULL       (IPARM=1)                                  W3-1
        A=TAL                                                    W3-2
        B=A + 7.                                                 W3-3
        A1=A - 1.                                                W3-4
        B1=B                                                     W3-5
        IB=(B1-A1    )*100. + 1.
        WRITE(6,105) IB
        ID=IB + 2
        DO  7 I=1,IB
        S=FLOAT(I-1)*1.E+50/1.E+52 + A1
        F(I)=S
    7   Y(I)=S
C       SCALE CONTROL FOR ABSCISSA AXIS
        Y(IB+1)=AINT(A1) - 1.
        X (N+1)=Y(IB+1)
        XLL=Y(IB+1)
        Y(IB+2)=AINT(B1) + 1.
        X (N+2)=Y(IB+2)
        CALL SCALE(Y, ID,8.0,YMIN,DY,1)
        WRITE(6,112) YMIN,DY
        CALL PLOT(2.0,3.0,-3)
        DO  8 I=1,N
    8   D(I)=1./(X(N)-X(1))                                      IT1-1
        H=1.                                                     IT1-2
        WRITE(6,110) H
        XHN=XN*H
C*****************************************
C
C       LOOP FOR CONVERGENCE OF DENSITY
C
C*****************************************
        KCT=0
```

```
      JP=1
C     ICO(2)=0   DO NOT ITERATE
C            =1   ITERATE
      IF(ICO(2) .EQ. 0) GO TO 16
    9 DO 13 I=1,N
      XX=X(I)
   13 E(I)=FCT(XX)
      DO 15 IR=1,N
      IF(ABS(D(IR)-E(IR)).GT. .001)GO TO 17
   15 CONTINUE
   16 KCT=1
      IF(ICO(2)) 19,20,19
   17 WRITE(6,106) (D(I),I=1,N)
C     YS=YES   PLOT EACH ITERATION
C       =NO    PLOT FINAL SOLUTION ONLY
      IF(YS-YES) 28,20,28
   19 WRITE(6,107) JP
      WRITE(7,108) (D(I),I=1,N)
      WRITE(6,106) (D(I),I=1,N)
   20 DO 24 I=1,IB
      XX=F(I)
   24 G(I)=FCT(XX)
      AREA=0.
      DO 26 I=1,IB
      R(I)=G(I)
   26 AREA=AREA + G(I)
      AREA=AREA*.01
C*************************************************
C
C     PLOT OF THE ESTIMATED DENSITY
C
C*************************************************
C     SCALE CONTROL FOR ORDINATE AXIS
      SVE1=0.0
      G(IB+1)=SVE1
      G(IB+2)=SVE2
      CALL SCALE(G, ID,5.0,GMIN,DG,1)
      WRITE(6,113) GMIN,DG,AREA
      IF(KCT.EQ.0) GO TO 27
      CALL AXIS( 0.0,0.0,6HX AXIS,-6,8.0,
     1           0.0,YMIN,DY)
      CALL AXIS( 0.0,0.0,8HPDF AXIS, 8,5.0,
     1           90.0,GMIN,DG)
C     ICO(6)=0   DO NOT PLOT HISTOGRAM
C            =1   PLOT HISTOGRAM
      IF(ICO(6) .EQ. 1) CALL HIST1(NUM)
```

```
   27 CALL PLOT(Y( 1 ),0.0,3)
      CALL PLOT(Y( 1 ),G( 1 ),2)
      CALL LINE(Y,G, IB,1)
      CALL PLOT(Y( IB),G( IB),3)
      CALL PLOT(Y( IB),0.0,2)
      IF(KCT.EQ.1) GO TO 30
   28 DO 29 IR=1,N
   29 D(IR)=E(IR)
      JP=JP +1
      IF(JP .LE. 250) GO TO 9
      WRITE(6,109)
      GO TO 36
   30 IF(ICO(3).EQ.0.AND.ICO(5).EQ.0)GO TO 31
      CALL ECDF(XLL)
C     ICO(5)=0   DO NOT PLOT RELIABILITY
C           =1   PLOT RELIABILITY
      IF(ICO(5) .EQ. 1) CALL REL
C     IPARM ¬=8   TRUE DENSITY KNOWN
C            =8   TRUE DENSITY UNKNOWN
   31 IF(IPARM .NE. 8) CALL TRUE(A,B,A1,B1)
C     ICO(4)=0   DO NOT PLOT SAMPLE CDF
C           =1   PLOT SAMPLE CDF
      IF(ICO(4) .EQ. 1) CALL ESCDF
      CALL PLOT(13.0,0.0,-3)
   36 CALL PLOT(0.0,0.0,-4)
   37 STOP
  100 FORMAT(I3,7I1,I3,A3)
  101 FORMAT('1IPARM=',I2,4X,'IYFL=',I2/
     1       '  ITER=',I2,6X,'YS=',A3/
     2       '   CDF=',I2,4X,'SCDF=',I2/
     3       '  IREL=',I2,4X,'HIST=',I2/
     4       '  ISAM=',I2,5X,'NUM=',I3)
  102 FORMAT(8F10.6)
  103 FORMAT('0DATA, N=',I3/
     1        (1X,5(E16.6,4X)/))
  104 FORMAT('0XBAR=',E16.6,3X,'S2=',E16.6/
     1       '   A1=',E16.6,3X,'B1=',E16.6)
  105 FORMAT('0IB= ',I5)
  106 FORMAT(///(1X,5(E16.6,4X)))
  107 FORMAT('0DISTRIBUTION CONVERGED IN ',
     1       I4,' ITERATIONS')
  108 FORMAT(5E16.6)
  109 FORMAT('0DENSITY DID NOT CONVERGE')
  110 FORMAT('0H=',E16.6)
  112 FORMAT(/' YMIN=',E16.6,3X,'DY=',E16.6/)
```

```
  113 FORMAT(' GMIN=',E16.6,3X,'DG=',E16.6,
     1       3X,'AREA=',E16.6)
      END


      FUNCTION FCT(XX)
C*****************************************************
C
C     EVALUATION OF DENSITY ESTIMATE
C     AT A POINT XX
C
C*****************************************************
      COMMON Y(6003),G(6003),F(6003),R(6003),
     1       D(200),X(202),E(200),
     2       CDF(202),ICO(7),YMIN,DY,IB,ID,
     3       N,PI,H,XN,XHN,SVE1,SVE2,SVE3,
     4       RHO,TAL,THETA,RXM,UM,VAR,VARS,
     5       ALPHA,SALPHA,SIG,BETA
      AR=0.
      DO 1 IR=1,N
      P=FLOAT(IR)/(XN + 1.)
      Q=1.-P
      D1=D(IR)
      TPQ=SQRT(XN/(P*Q))*D1
      XYZW=SQRT(XN/(P*Q*PI*2.))*D1
      Z=(XX-X(IR))*TPQ/H
      IF(ABS(Z).GT. 7.) GO TO 3
      IF(Z .EQ. 0.) GO TO 2
      ZT=Z**2/2.
      XPQR=EXP(-ZT)
      HP  =XPQR*XYZW
      GO TO 1
    2 HP=XYZW
      GO TO 1
    3 HP=0.
    1 AR=AR + HP
      FCT=AR/XHN
      RETURN
      END
```

```
      SUBROUTINE REL
C*****************************************************
C
C     PLOT OF RELIABILITY AND HAZARD
C
C*****************************************************
      COMMON Y(6003),G(6003),F(6003),R(6003),
     1       D(200),X(202),E(200),
     2       CDF(202),ICO(7),YMIN,DY,IB,ID,
     3       N,PI,H,XN,XHN,SVE1,SVE2,SVE3,
     4       RHO,TAL,THETA,RXM,UM,VAR,VARS,
     5       ALPHA,SALPHA,SIG,BETA
      R(IB+1)=1.0
      R(IB+2)=0.0
      F(IB+1)=0.0
      F(IB+2)=SVE3
      CALL PLOT(13.0,0.0,-3)
      CALL SCALE(R,ID,5.0,RMIN,DR,1)
      CALL SCALE(F,ID,5.0,FMIN,DF,1)
      WRITE(6,100) RMIN,DR,FMIN,DF
      CALL AXIS(0.0,0.0,6HX AXIS,-6,8.0,
     1          0.0,YMIN,DY)
      CALL AXIS(0.0,0.0,9HR(T) AXIS,9,5.0,
     1          90.0,RMIN,DR)
      CALL PLOT(Y(IB+1),R(IB+1),3)
      CALL PLOT(Y(1),R(IB+1),2)
      CALL LINE(Y,R,IB,1)
      CALL PLOT(13.0,0.0,-3)
      CALL AXIS(0.0,0.0,6HX AXIS,-6,8.0,
     1          0.0,YMIN,DY)
      CALL AXIS(0.0,0.0,9HZ(T) AXIS,9,5.0,
     1          90.0,FMIN,DF)
      CALL PLOT(Y(1),F(1),3)
      DO 2 I=2,IB
      IF(F(I) .EQ. 0.) GO TO 1
      CALL PLOT(Y(I),F(I),2)
      GO TO 2
    1 CALL PLOT(Y(I),F(I),3)
    2 CONTINUE
      RETURN
  100 FORMAT(/' RMIN=',E16.6,3X,'DR=',E16.6/
     1        ' FMIN=',E16.6,3X,'DF=',E16.6)
      END
```

```
      SUBROUTINE HIST(NUM)
C*****************************************************
C
C     PLOT OF HISTOGRAM
C
C*****************************************************
      DIMENSION W(23),WW(22)
      COMMON Y(6003),G(6003),F(6003),R(6003),
     1       D(200),X(202),E(200),
     2       CDF(202),ICO(7),YMIN,DY,IB,ID,
     3       N,PI,H,XN,XHN,SVE1,SVE2,SVE3,
     4       RHO,TAL,THETA,RXM,UM,VAR,VARS,
     5       ALPHA,SALPHA,SIG,BETA
      FN=NUM
      WX=(X(N)-X(1))/FN
      WY=WX/2.
      WX=(X(N)-X(1)+WX)/FN
      ST=X(1)-WY
      N1=NUM +1
      DO 1 I=1,N1
      A=I-1
      W(I)=ST + A*WX
    1 WW(I)=0.
      TN=N
      TN=1./TN
      JI=1
      DO 4 I=1,N
      DO 2 J=JI,NUM
      JJ=J
      IF(X(I) .GE. W(J) .AND.
     1   X(I) .LT. W(J+1)) GO TO 3
    2 CONTINUE
    3 WW(JJ)=WW(JJ) + TN
    4 JI=JJ
      RETURN
      ENTRY HIST1(NUM)
      IC=NUM + 2
      WW(IC)=SVE2
      W(IC)=X(N+1)
      W(IC+1)=X(N+2)
      IC1=IC+1
      CALL SCALE(W,IC1,8.0,WMIN,DW,1)
      CALL SCALE(WW,IC,5.0,CMIN,DC,1)
      WRITE(6,115) WMIN,DW,CMIN,DC
      CALL PLOT(W(1),0.0,3)
      CALL PLOT(W(1),WW(1),2)
```

```
      DO 5 I=2,N1
      CALL PLOT(W(I),WW(I-1),2)
    5 CALL PLOT(W(I),WW(I),2)
      RETURN
  115 FORMAT(/' WMIN=',E16.6,3X,'DW=',E16.6/
     1        ' CMIN=',E16.6,3X,'DC=',E16.6)
      END


      SUBROUTINE SAMP(NUM)
C*************************************************
C
C     COMPUTATION OF SAMPLE PERCENTILES
C
C*************************************************
      COMMON Y(6003),G(6003),F(6003),R(6003),
     1       D(200),X(202),E(200),
     2       CDF(202),ICO(7),YMIN,DY,IB,ID,
     3       N,PI,H,XN,XHN,SVE1,SVE2,SVE3,
     4       RHO,TAL,THETA,RXM,UM,VAR,VARS,
     5       ALPHA,SALPHA,SIG,BETA
      N1=NUM +1
      XN1=N1
      DO 2 J=1,NUM
      XK=FLOAT(J*N)/XN1
      K=XK
      YK=K
      IF(YK .EQ. XK) GO TO 1
      X(J)=X(K+1)
      GO TO 2
    1 X(J)=X(K)
    2 CONTINUE
      WRITE(6,100) NUM,(X(I),I=1,NUM)
      N=NUM
      XN=N
      RETURN
  100 FORMAT('OSAMPLE PERCENTILES, NUM=',I3/
     1       (1X,5(E16.6,4X)/))
      END
```

```
      SUBROUTINE ECDF(XLL)
C*****************************************************
C
C     PLOT OF THE ESTIMATED CDF
C
C*****************************************************
      DIMENSION AUX(10)
      COMMON Y(6003),G(6003),F(6003),R(6003),
     1       D(200),X(202),E(200),
     2       CDF(202),ICO(7),YMIN,DY,IB,ID,
     3       N,PI,H,XN,XHN,SVE1,SVE2,SVE3,
     4       RHO,TAL,THETA,RXM,UM,VAR,VARS,
     5       ALPHA,SALPHA,SIG,BETA
      EXTERNAL FCT
      TEMP=0.
      NDIM=10
      EPS=.001
      G(IB+1)=0.0
      G(IB+2)=1.0
      DO  4 I=1,IB
      XX=F(I)
      CALL QATR(XLL,XX,EPS,NDIM,FCT,ANS,
     1          IER,AUX)
      G(I)=TEMP + ANS
      PQZ=1.-G(I)
      IF(PQZ .LE. .005) GO TO 2
      F(I)=R(I)/PQZ
      GO TO 3
    2 F(I)=0.
    3 R(I)=PQZ
      TEMP=G(I)
    4 XLL=XX
      WRITE(6,111) G(IB),R(IB)
C     ICO(3)=0   DO NOT PLOT ESTIMATED CDF
C           =1   PLOT ESTIMATED CDF
      IF(ICO(3) .EQ. 0) RETURN
      CALL PLOT(13.0,0.0,-3)
      CALL SCALE( G ,ID,5.0,GMIN,DG,1)
      WRITE(6,114) GMIN,DG
      CALL AXIS(0.0,0.0,6HX AXIS,-6,8.0,
     1          0.0,YMIN,DY)
      CALL AXIS(0.0,0.0,8HCDF AXIS,8,5.0,
     1          90.0,GMIN,DG)
      CALL PLOT(Y(1),0.0,3)
      CALL PLOT(Y(1),G(1),2)
      CALL LINE(Y,G, IB,1)
```

```
      CALL PLOT(Y( IB),G( ID),3)
      CALL PLOT(Y( ID),G( ID),2)
      RETURN
  111 FORMAT('0G=',E16.6/   ' R=',E16.6)
  114 FORMAT(/' GMIN=',E16.6,3X,'DG=',E16.6/)
      END


      SUBROUTINE ESCDF
C*****************************************
C
C     PLOT OF THE SAMPLE CDF
C
C*****************************************
      COMMON Y(6003),G(6003),F(6003),R(6003),
     1       D(200),X(202),E(200),
     2       CDF(202),ICO(7),YMIN,DY,IB,ID,
     3       N,PI,H,XN,XHN,SVE1,SVE2,SVE3,
     4       RHO,TAL,THETA,RXM,UM,VAR,VARS,
     5       ALPHA,SALPHA,SIG,BETA
      CDF(N+1)=0.
      CDF(N+2)=1.
      IC=N+2
      CALL PLOT(13.0,0.0,-3)
      CALL SCALE(X ,IC,8.0,XMIN,DX,1)
      CALL SCALE(CDF,IC,5.0,CMIN,DC,1)
      WRITE(6,115) XMIN,DX,CMIN,DC
      CALL AXIS(0.0,0.0,6HX AXIS,-6,8.0,
     1          0.0,XMIN,DX)
      CALL AXIS(0.0,0.0,8HCDF AXIS,8,5.0,
     1          90.0,CMIN,DC)
      CALL PLOT(X (1),0.0,3)
      CALL PLOT(X (1),CDF(1),2)
      DO 1  I=2,N
      CALL PLOT(X (I),CDF(I-1),2)
    1 CALL PLOT(X (I),CDF(I),2)
      CALL PLOT(X (IC),CDF(IC),2)
      RETURN
  115 FORMAT(/' XMIN=',E16.6,3X,'DX=',E16.6/
     1        ' CMIN=',E16.6,3X,'DC=',E16.6)
      END
```

```
      SUBROUTINE TRUE(A,B,C,O)
C*****************************************
C
C     PLOT OF THE TRUE DENSITY AND TRUE CDF
C
C*****************************************
      COMMON Y(6003),G(6003),F(6003),R(6003),
     1       D(200),X(202),E(200),
     2       CDF(202),ICO(7),YMIN,DY,IB,ID,
     3       N,PI,H,XN,XHN,SVE1,SVE2,SVE3,
     4       RHO,TAL,THETA,RXM,UM,VAR,VARS,
     5       ALPHA,SALPHA,SIG,BETA
      IB=(B-A)*100. + 1.
      WRITE(6,100) IB
      DO  1 I=1,IB
      XX=FLOAT(I-1)*1.E+50/1.E+52 + A
      Y(I)=XX
      F(I)=FX(XX)
      G(I)=GX(XX)
    1 R(I)=1.-F(I)
      Y(IB+1)=AINT(C) - 1.
      Y(IB+2)=AINT(O) + 1.
      F(IB+1)=0.0
      F(IB+2)=1.0
      G(IB+1)=SVE1
      G(IB+2)=SVE2
      ID=IB+2
      CALL PLOT(13.0,0.0,-3)
      CALL SCALE(Y ,ID,8.0,YMIN,DY,1)
      CALL SCALE( F ,ID,5.0,FMIN,DF,1)
      CALL AXIS(0.0,0.0,6HX AXIS,-6,8.0,
     1          0.0,YMIN,DY)
      CALL AXIS(0.0,0.0,8HCDF AXIS,8,5.0,
     1          90.0,FMIN,DF)
      CALL PLOT(Y(1),0.0,3)
      CALL PLOT(Y(1),F(1),2)
      CALL LINE(Y,F, IB,1)
      CALL PLOT(Y(IB  ),F(ID),3)
      CALL PLOT(Y(ID  ),F(ID),2)
C     ICO(5)=0   DO NOT PLOT RELIABILITY
C           =1   PLOT RELIABILITY
      IF(ICO(5) .EQ. 0) GO TO 4
      DO 3 I=1,IB
      IF(R(I) .LE. .005) GO TO 2
      F(I)=G(I)/R(I)
      GO TO 3
```

```
    2 F(I)=0.
    3 CONTINUE
      CALL REL
    4 CALL PLOT(13.0,0.0,-3)
      CALL SCALE( G ,ID,5.0,GMIN,DG,1)
      WRITE(6,101) YMIN,DY,GMIN,DG,FMIN,DF
      CALL AXIS(0.0,0.0,6HX AXIS,-6,8.0,
     1          0.0,YMIN,DY)
      CALL AXIS(0.0,0.0,8HPDF AXIS,8,5.0,
     1          90.0,GMIN,DG)
      CALL PLOT(Y(1),0.0,3)
      CALL PLOT(Y(1),G(1),2)
      CALL LINE(Y,G, IB,1)
      CALL PLOT(Y( IB),G( IB),3)
      CALL PLOT(Y( IB),0.0,2)
      AVGSE=0.
      DO 5 I=1,N
      XX=X(I)
    5 AVGSE=AVGSE + (FCT(XX)-GX(XX))**2
      AVGSE=AVGSE/XN
      WRITE(6,102) AVGSE
      RETURN
  100 FORMAT('OIB= ',I5)
  101 FORMAT(/' YMIN=',E16.6,3X,'DY=',E16.6/
     1       ' GMIN=',E16.6,3X,'DG=',E16.6/
     2       ' FMIN=',E16.6,3X,'DF=',E16.6)
  102 FORMAT(//' AVG SQ ERROR= ',E16.6)
      END



      FUNCTION FX(XX)
C*****************************************
C
C     EVALUATES TRUE CDF
C
C*****************************************
      COMMON Y(6003),G(6003),F(6003),R(6003),
     1       D(200),X(202),E(200),
     2       CDF(202),ICO(7),YMIN,DY,IB,ID,
     3       N,PI,H,XN,XHN,SVE1,SVE2,SVE3,
     4       RHO,TAL,THETA,RXM,UM,VAR,VARS,
     5       ALPHA,SALPHA,SIG,BETA
C     WEIBULL       (IPARM=1)                          W4-1
      FX=1.-EXP(-(XX-TAL)**RHO/THETA)                  W4-2
```

```
      RETURN
      END


      FUNCTION GX(XX)
C*****************************************************
C
C     EVALUATES TRUE PDF
C
C*****************************************************
      COMMON Y(6003),G(6003),F(6003),R(6003),
     1       D(200),X(202),E(200),
     2       CDF(202),ICO(7),YMIN,DY,IB,ID,
     3       N,PI,H,XN,XHN,SVE1,SVE2,SVE3,
     4       RHO,TAL,THETA,RXM,UM,VAR,VARS,
     5       ALPHA,SALPHA,SIG,BETA
C     WEIBULL       (IPARM=1)                          W5-1
      GX=RHO*(XX-TAL)**(RHO-1.)*
     1    EXP(-(XX-TAL)**RHO/THETA)/THETA              W5-2
      RETURN
      END
```

APPENDIX B

PROGRAM E-SPLINE

The computer program contained in Appendix B has been entitled "E-Spline" for the exponential spline estimator of Chapter IV. Again, several statements have been coded with the codes ALL and U (for uniform density). Their meanings and necessary manipulations are identical as those described in Appendix A since the two computer programs have been structured to be as compatible as possible.

The input data stream for E-Spline has changed slightly. The sample size and the maximum number of knots to be used plus two (2) are contained on the first data card (Format #111) and the values of the program control parameters are contained on the second card (Format #100). The parameter ICO(2) is ignored by this program. Any data cards (Format #102) containing the observations follow this card (remember to replace loop DO 4 by the READ statement). The next data card (Format #102) contains the critical value of the Kolmogorov-Smirnov goodness of fit test. Subsequent data cards (Format #101 of the first subroutine NOD), if any, contain the subscript and the value of the fixed knots.

Many of the variables used in this program have been discussed in the preceeding appendix. Please refer to it for the description of the following variables; IPARM, ICO(·), NUM, YS, IX, SVE2, SVE3, N, RANDU, X, A, B, A1, B1, XMIN, DX,

QATR, and AREA. The following important variables appear only in the program given in Appendix B.

NP1 = maximum number knots to be used +2;

U1 = corresponds to parameter u in equation (4.2.6);

BB = corresponds to parameter B_1 in equation (4.2.6);

D = array corresponding to the parameters a_i^*;

C1 = array corresponding to the parameters c_i^*;

XKS = critical value of Kolmogorov-Smirnov test;

KSTEST = subroutine for the Kolmogorov-Smirnov test.

We mention in Chapter IV that the goodness of fit test may be eliminated. This can be done by removing the two cards containing the code TEST, and replacing the second card by

```
IF(I1.EQ.NP1) GO TO 23 .
```

Finally, there are two subroutines named NOD contained in the program listing. The first one is used when the knots are fixed while the second subroutine NOD is used for free knots.

Again, we mention that we have only tried to present sufficient details for use of this computer program.

```
      DIMENSION AB(15),IC1(15)
      COMMON Y(6003),G(6003),F(6003),R(6003),
     1       X(202),CDF(202),C1(15),D(16),
     2       ICO(7),YMIN,DY,IB,ID,N,PI,XN,
     3       SVE1,SVE2,SVE3,U1,S1,BB,LT,AA,
     4       EPS,NDIM,II,JX,BIN,EN,NP,KNOT,
     5       RHO,TAL,THETA,RXM,UM,VAR,VARS,
     6       ALPHA,SALPHA,SIG,BETA
      DATA YES/'YES'/
      IX=13795
      EPS=.001
      NDIM=10
      PI=3.141593
      SVE2=.40                                                  ALL1-1
      SVE3=5.0                                                  ALL1-2
      READ(5,111) N,NP1
      N1=N-1
      XN=N
      READ(5,100) IPARM,(ICO(I),I=1,7),NUM,YS
      WRITE(6,101) IPARM,ICO(1),YS,
     1             (ICO(I),I=3,7),NUM
C     PARAMETERS OF SIMULATED TRUE DENSITY
C     UNIFORM       (IPARM=6)                                   U1-1
      ALPHA=0.                                                  U1-2
      BETA=10.                                                  U1-3
      DO 4 I=1,N
C     ICO(1)=0   RANDOM DATA
C           =1   UNIFORM CDF DATA
      IF(ICO(1)) 37,2,1
    1 YFL=FLOAT(I)/(XN+1.)                                      ALL2-1
      GO TO 3
    2 CALL RANDU(IX,IY,YFL)                                     ALL2-2
      IX=IY                                                     ALL2-3
C     UNIFORM       (IPARM=6)                                   U2-1
    3 X(I)=YFL*(BETA-ALPHA) + ALPHA                             U2-2
    4 CONTINUE
      DO 5 I=1,N1
      DO 5 J=I,N
      IF(X(I).LE.X(J)) GO TO 5
      TEMP=X(J)
      X(J)=X(I)
      X(I)=TEMP
    5 CONTINUE
      WRITE(6,103)  N, (X(I),I=1,N)
C     ICO(6)=0   DO NOT PLOT HISTOGRAM
C           =1   PLOT HISTOGRAM
```

```
      IF(ICO(6) .EQ. 1) CALL HIST(NUM)
C     ICO(7)=0   USE ENTIRE DATA
C            =1   USE SAMPLE PERCENTILES
      IF(ICO(7) .EQ. 1) CALL SAMP(NUM)
      S=0.
      U=0.
      DO 6 I=1,N
      CDF(I)=FLOAT(I)/XN
      S=S + X(I)**2
    6 U=U + X(I)
      U= U/XN
      S=(S-XN*U*U)/XN
      U1=U
      S1=S
      BB=1./SQRT(2.*PI*S)
      SQ=SQRT(S)
      A1=U - 5.*SQ
      B1=U + 5.*SQ
      WRITE(6,104) U,S,A1,B1
C     UNIFORM       (IPARM=6)                                    U3-1
      A=ALPHA                                                    U3-2
      B=BETA                                                     U3-3
      A1=A -2.                                                   U3-4
      B1=B +2.                                                   U3-5
      D(1)=A1
      D(2)=B1
      C1(1)=0.
      C1(2)=0.
      IB=(B1-A1)*100. + 1.
      WRITE(6,105) IB
      ID=IB + 2
      DO  7 I=1,IB
      S=FLOAT(I-1)*1.E+50/1.E+52 + A1
      F(I)=S
    7 Y(I)=S
C     SCALE CONTROL FOR ABSCISSA AXIS
      Y(IB+1)=AINT(A1) - 1.
      X (N+1)=Y(IB+1)
      Y(IB+2)=AINT(B1) + 1.
      X (N+2)=Y(IB+2)
      CALL SCALE(Y, ID,8.0,YMIN,DY,1)
      WRITE(6,112) YMIN,DY
      CALL PLOT(2.0,3.0,-3)
C     READ CRITICAL VALUE OF K-S TEST
      READ(5,102) XKS
```

```
C       TEST NORMAL DENSITY, MEAN=U1
C       VARIANCE=S1  FOR FIT
        KNOT=0
        CALL KSTEST(DIF)
        IF(DIF .LE. XKS) GO TO 23                                    TEST
        WRITE(6,117) KNOT,DIF
C*************************************************
C
C       LOOP FOR CONVERGENCE OF SPLINE FUNCTION
C
C*************************************************
        KCT=0
        I1=3
   40   KNOT=I1-2
        NP=KNOT
        CALL NOD
        NP=NP+1
        DO  9 I=2,NP
        AB(I)=0.
        DO  9 J=1,N
        IF(X(J)-D(I)) 9,9,8
    8   AB(I)=AB(I) + (X(J)-D(I))**2/XN
    9   CONTINUE
        AB(1)=S1 + (U-D(2))**2 - AB(2)
        WRITE(6,114) (AB(I),I=1,NP)
        JN=I1-II
        DO 10 I=1,JN
   10   C1(I1-I)=0.0
C*************************************************
C
C       LOOP FOR CONVERGENCE OF U1
C
C*************************************************
        DO 20 I2=1,1500
        KCON=0
        QQ=U1
C*************************************************
C
C       LOOP FOR CONVERGENCE OF C1(IJ)'S
C
C*************************************************
        DO 15 IJ=1,NP
        NL=IJ
        IC1(NL)=0
        BIN=D(NL)
        IF(NL-1) 12,12,11
```

```
11 AA=D(NL)
   EN=D(I1)
   GO TO 13
12 AA=D(2)
   EN=AA
13 C=C1(NL)
   CALL ARE(3,SS,5,S4)
   SIG=AB(NL)
   C=C-(SIG-SS)/S4
   IF(ABS(C-C1(NL)).LE. .001) GO TO 14
   GO TO 15
14 KCON=KCON +1
   IC1(NL)=1
15 C1(NL)=C
   BIN=D(1)
   EN=D(I1)
   CALL ARE(1,AR,0,DUM)
   GLA=1./AR
   BB=BB*GLA
   CALL ARE(2,AX,4,AX2)
   U1=U1 + (U-AX)*S1/(AX2-U1*AX)
   CALL ARE(1,AR,0,DUM)
   GLA=1./AR
   BB=BB*GLA
   IF(KCON-NP) 16,18,16
16 IF(ABS(U1-Q0)  -  1.E-3) 17,17,19
17 IC2=1
   GO TO 20
18 IF(ABS(U1-Q0).LE. 1.E-3) GO TO 21
19 IC2=0
20 CONTINUE
   WRITE(6,115)
   WRITE(6,106) BB,U1,AR,AX,(C1(M),M=1,NP)
   WRITE(6,118) (IC1(M1),M1=1,NP),IC2
   GO TO 28
21 WRITE(6,110) I2
   WRITE(6,106) BB,U1,AR,AX,(C1(M),M=1,NP)
   Q=0.
   DO 22 J=1,NP
22 Q=Q + C1(J)*AB(J)*XN
   Q=Q + XN*(U-U1)**2/(2.*S1)
  1    + XN*.5 - XN*ALOG(BB)
   WRITE(6,109) Q
   CALL KSTEST(DIF)
   IF(DIF .LE. XKS) GO TO 23                    TEST
   WRITE(6,117) KNOT,DIF
```

```
C      YS=YES    PLOT EACH ITERATION
C        =NO     PLOT FINAL SOLUTION ONLY
       IF(YS-YES) 28,24,28
   23  WRITE(6,116) KNOT,XKS,DIF
       KCT=1
   24  JX=1
       JP=1
       LT=1
       DO 25 I=1,IB
       XX=F(I)
       IF(KNOT .EQ. 0) GO TO 25
       DO 41 KJ=JP,NP
       IF(XX-D(KJ)) 42,41,41
   41  JX=KJ
   42  JP=JX
   25  G(I)=FT(XX)
       AREA=0.
       DO 26 I=1,IB
       R(I)=G(I)
   26  AREA=AREA + G(I)
       AREA=AREA*.01
C*****************************************
C
C      PLOT OF THE ESTIMATED DENSITY
C
C*****************************************
C      SCALE CONTROL FOR ORDINATE AXIS
       SVE1=0.0
       G(IB+1)=SVE1
       G(IB+2)=SVE2
       CALL SCALE(G, ID,5.0,GMIN,DG,1)
       WRITE(6,113) GMIN,DG,AREA
       IF(KCT.EQ.0) GO TO 27
       CALL AXIS( 0.0,0.0,6HX AXIS,-6,8.0,
      1           0.0,YMIN,DY)
       CALL AXIS( 0.0,0.0,8HPDF AXIS, 8,5.0,
      1           90.0,GMIN,DG)
C      ICO(6)=0    DO NOT PLOT HISTOGRAM
C             =1   PLOT HISTOGRAM
       IF(ICO(6) .EQ. 1) CALL HIST1(NUM)
   27  CALL PLOT(Y( 1 ),0.0,3)
       CALL PLOT(Y( 1 ),G( 1 ),2)
       CALL LINE(Y,G, IB,1)
       CALL PLOT(Y( IB),G( IB),3)
       CALL PLOT(Y( IB),0.0,2)
       IF(KCT.EQ.1) GO TO 30
```

```
   28 I1=I1 + 1
      IF(I1 .LE. NP1) GO TO 40
      GO TO 36
   30 IF(ICO(3).EQ.0.AND.ICO(5).EQ.0)GO TO 31
      CALL ECDF
C     ICO(5)=0   DO NOT PLOT RELIABILITY
C           =1   PLOT RELIABILITY
      IF(ICO(5) .EQ. 1) CALL REL
C     IPARM ¬=8   TRUE DENSITY KNOWN
C            =8   TRUE DENSITY UNKNOWN
   31 IF(IPARM .NE. 8) CALL TRUE(A,B,A1,B1)
C     ICO(4)=0   DO NOT PLOT SAMPLE CDF
C           =1   PLOT SAMPLE CDF
      IF(ICO(4) .EQ. 1) CALL ESCDF
      CALL PLOT(13.0,0.0,-3)
   36 CALL PLOT(0.0,0.0,-4)
   37 STOP
  100 FORMAT(I3,7I1,I3,A3)
  101 FORMAT('1IPARM=',I2,4X,'IYFL=',I2/
     1                    15X,'YS=',A3/
     2       '   CDF=',I2,4X,'SCDF=',I2/
     3       '  IREL=',I2,4X,'HIST=',I2/
     4       '  ISAM=',I2,5X,'NUM=',I3)
  102 FORMAT(8F10.6)
  103 FORMAT('0DATA, N=',I3/
     1        (1X,5(E16.6,4X)/))
  104 FORMAT('0XBAR=',E16.6,3X,'S2=',E16.6/
     1       '   A1=',E16.6,3X,'B1=',E16.6)
  105 FORMAT('0IB= ',I5)
  106 FORMAT(5X,'BB=',E16.6,5X,'U1=',E16.6/
     1       5X,'AREA=',E16.6,3X,
     2       'EXPECTED VALUE=',E16.6/5X,
     3       'THE C1(J)''S ARE'/(3X,4E16.6))
  109 FORMAT(5X,'NEGATIVE LOG-LIKELIHOOD=',
     1                                E16.6)
  110 FORMAT(5X,'SPLINE HAS CONVERGED IN',
     1                  I5,' ITERATIONS')
  111 FORMAT(I3,I2)
  112 FORMAT(/' YMIN=',E16.6,3X,'DY=',E16.6/)
  113 FORMAT(' GMIN=',E16.6,3X,'DG=',E16.6,
     1       3X,'AREA=',E16.6)
  114 FORMAT(/' AB''S'/(5X,4E16.6))
  115 FORMAT(' SPLINE HAS NOT CONVERGED')
  116 FORMAT(////' ACCEPT K-S TEST, KNOT= ',
     1       I3/' XKS=',F6.3,5X,'DIF=',F6.3/)
```

```
  117 FORMAT(////' REJECT K-S TEST, KNOT= ',
     1               I3,5X,'DIF=',F6.3////)
  118 FORMAT(' IC1''S'/(1X,15I2))
      END


      SUBROUTINE KSTEST(DIF)
C*****************************************************
C
C     GOODNESS OF FIT TEST
C
C*****************************************************
      DIMENSION AUX(10)
      COMMON Y(6003),G(6003),F(6003),R(6003),
     1       X(202),CDF(202),C1(15),D(16),
     2       IC0(7),YMIN,DY,IB,ID,N,PI,XN,
     3       SVE1,SVE2,SVE3,U1,S1,BB,LT,AA,
     4       EPS,NDIM,II,JX,BIN,EN,NP,KNOT,
     5       RHO,TAL,THETA,RXM,UM,VAR,VARS,
     6       ALPHA,SALPHA,SIG,BETA
      EXTERNAL FT
      XLL=X (N+1)
      LT=1
      DIF=0.
      TEMP=0.
      JX=1
      JP=1
      DO 1 I=1,N
      XX=X(I)
      IF(KNOT .EQ. 0) GO TO 4
      DO 2 J=JP,NP
      IF(XX-D(J)) 3,2,2
    2 JX=J
    3 JP=JX
    4 CALL QATR(XLL,XX,EPS,NDIM,FT,SZ,IR,AUX)
      TEMP =TEMP + SZ
      DIF1=ABS(TEMP - CDF(I))
      IF(DIF1 .GT. DIF) DIF=DIF1
    1 XLL=XX
      RETURN
      END
```

```
      SUBROUTINE NOD
C*************************************************
C
C     DETERMINATION OF FIXED NODES
C
C*************************************************
      COMMON Y(6003),G(6003),F(6003),R(6003),
     1       X(202),CDF(202),C1(15),D(16),
     2       ICO(7),YMIN,DY,IB,ID,N,PI,XN,
     3       SVE1,SVE2,SVE3,U1,S1,BB,LT,AA,
     4       EPS,NDIM,II,JX,BIN,EN,NP,KNOT,
     5       RHO,TAL,THETA,RXM,UM,VAR,VARS,
     6       ALPHA,SALPHA,SIG,BETA
      READ(5,101) V,II
      LP=NP+2-II
      DO  1 JJ=1,LP
    1 D(NP+3-JJ)=D(NP+2-JJ)
      D(II)=V
      NS=NP+2
      WRITE(6,102) (D(J),J=1,NS)
      RETURN
  101 FORMAT(F6.2,I2)
  102  FORMAT(//' D(J)''S'/(5X,4E16.6))
      END


      SUBROUTINE NOD
C*************************************************
C
C     DETERMINATION OF FREE NODES
C
C*************************************************
      DIMENSION AL(50),SA(50)
      COMMON Y(6003),G(6003),F(6003),R(6003),
     1       X(202),CDF(202),C1(15),D(16),
     2       ICO(7),YMIN,DY,IB,ID,N,PI,XN,
     3       SVE1,SVE2,SVE3,U1,S1,BB,LT,AA,
     4       EPS,NDIM,II,JX,BIN,EN,NP,KNOT,
     5       RHO,TAL,THETA,RXM,UM,VAR,VARS,
     6       ALPHA,SALPHA,SIG,BETA
      NN=30
      XNN=NN
      H=(X(N )-X(1))/XNN
      U=X(1)-H
```

```
      DO  3 I=1,NN
      U=U+H
      W=0.
      DO  2 IP=1,N
      IF(X(IP)-U) 2,2,1
    1 W=W + (X(IP)-U)**2
    2 CONTINUE
    3 SA(I)=W/XN
      BIN=X(1)-H
      EN=D(NP+1)
      DO 4 IO=1,NN
      BIN=BIN+H
      AA=BIN
      CALL ARE(3,SY,0,DUM)
    4 AL(IO)=SY
      U=X(1)-H
      XMAX=0.
      DO 8 I=1,NN
      DIF=ABS(AL(I)-SA(I))
      U=U+H
    6 IF(DIF-XMAX) 8,7,7
    7 XMAX=DIF
      AW=U
    8 CONTINUE
      DO  9 I=1,NP
      J=I+1
      IF(D(I) .LT. AW .AND. D(J) .GT. AW)
     1                                  GO TO 10
    9  CONTINUE
   10 II=J
      LP=NP+2-J
      DO 11 JJ=1,LP
   11 D(NP+3-JJ)=D(NP+2-JJ)
      D(J)=AW
      NS=NP+2
      WRITE(6,101) AW,XMAX
      WRITE(6,102) (D(J),J=1,NS)
      RETURN
  101 FORMAT('////' AW=' ,E16.6,3X,
     1           'XMAX=',E16.6/)
  102  FORMAT(//' D(J)''S'/(5X,4E16.6))
      END
```

```
      SUBROUTINE ARE(LT1,SY,LT2,SX)
C*****************************************************
C
C     DETERMINES LIMITS OF INTEGRATION
C
C*****************************************************
      DIMENSION AUX(10)
      COMMON Y(6003),G(6003),F(6003),R(6003),
     1       X(202),CDF(202),C1(15),D(16),
     2       ICO(7),YMIN,DY,IB,ID,N,PI,XN,
     3       SVE1,SVE2,SVE3,U1,S1,BB,LT,AA,
     4       EPS,NDIM,II,JX,BIN,EN,NP,KNOT,
     5       RHO,TAL,THETA,RXM,UM,VAR,VARS,
     6       ALPHA,SALPHA,SIG,BETA
      EXTERNAL FT
      SY=0.
      SX=0.
      DO  1 J1=1,NP
      IF(D(J1+1)-BIN) 1,1,2
    2 IF(D(J1)-BIN) 4,3,3
    3 X1=D(J1)
      GO TO 5
    4 X1=BIN
    5 IF(D(J1+1)-EN) 6,6,7
    6 X2=D(J1+1)
      JX=J1
      LT=LT1
      CALL QATR(X1,X2,EPS,NDIM,FT,SZ,IR,AUX)
      SY=SY+SZ
      IF(LT2 .EQ. 0) GO TO 1
      LT=LT2
      CALL QATR(X1,X2,EPS,NDIM,FT,SZ,IR,AUX)
      SX=SX+SZ
    1 CONTINUE
      RETURN
    7 X2=EN
      JX=J1
      LT=LT1
      CALL QATR(X1,X2,EPS,NDIM,FT,SZ,IR,AUX)
      SY=SY+SZ
      IF(LT2 .EQ. 0) RETURN
      LT=LT2
      CALL QATR(X1,X2,EPS,NDIM,FT,SZ,IR,AUX)
      SX=SX+SZ
      RETURN
```

```
      END


      FUNCTION FT(XX)
C*************************************************
C
C     EVALUATION OF DENSITY ESTIMATE AND
C     OTHER FUNCTIONS AT A POINT XX
C
C*************************************************
      COMMON Y(6003),G(6003),F(6003),R(6003),
     1       X(202),CDF(202),C1(15),D(16),
     2       ICO(7),YMIN,DY,IB,ID,N,PI,XN,
     3       SVE1,SVE2,SVE3,U1,S1,BB,LT,AA,
     4       EPS,NDIM,II,JX,BIN,EN,NP,KNOT,
     5       RHO,TAL,THETA,RXM,UM,VAR,VARS,
     6       ALPHA,SALPHA,SIG,BETA
      IF(JX-1) 1,1,2
    1 W=C1(1)*(XX-D(2))**2
      GO TO 4
    2 W=0.
      DO  3 J2=2,JX
    3 W=W + C1(J2)*(XX-D(J2))**2
    4 QZ=W + (XX-U1)**2/(2.*S1)
      IF(QZ .GT. 35.) GO TO 6
      Z=BB*EXP(-QZ)
      GO TO 7
    6 Z=0.
    7 GO TO (8,9,10,11,12),LT
    8 FT=Z
      RETURN
    9 FT=Z*XX
      RETURN
   10 WW=(XX-AA)**2
      FT=Z*WW
      RETURN
   11 FT=Z*XX**2
      RETURN
   12 WW=(XX-AA)**2
      FT=Z*WW**2
      RETURN
      END
```

```
      SUBROUTINE REL
C*****************************************************
C
C     PLOT OF RELIABILITY AND HAZARD
C
C*****************************************************
      COMMON Y(6003),G(6003),F(6003),R(6003),
     1       X(202),CDF(202),C1(15),D(16),
     2       ICO(7),YMIN,DY,IB,ID,N,PI,XN,
     3       SVE1,SVE2,SVE3,U1,S1,BB,LT,AA,
     4       EPS,NDIM,II,JX,BIN,EN,NP,KNOT,
     5       RHO,TAL,THETA,RXM,UM,VAR,VARS,
     6       ALPHA,SALPHA,SIG,BETA
      R(IB+1)=1.0
      R(IB+2)=0.0
      F(IB+1)=0.0
      F(IB+2)=SVE3
      CALL PLOT(13.0,0.0,-3)
      CALL SCALE(R,ID,5.0,RMIN,DR,1)
      CALL SCALE(F,ID,5.0,FMIN,DF,1)
      WRITE(6,100) RMIN,DR,FMIN,DF
      CALL AXIS(0.0,0.0,6HX AXIS,-6,8.0,
     1          0.0,YMIN,DY)
      CALL AXIS(0.0,0.0,9HR(T) AXIS,9,5.0,
     1          90.0,RMIN,DR)
      CALL PLOT(Y(IB+1),R(IB+1),3)
      CALL PLOT(Y(1),R(IB+1),2)
      CALL LINE(Y,R,IB,1)
      CALL PLOT(13.0,0.0,-3)
      CALL AXIS(0.0,0.0,6HX AXIS,-6,8.0,
     1          0.0,YMIN,DY)
      CALL AXIS(0.0,0.0,9HZ(T) AXIS,9,5.0,
     1          90.0,FMIN,DF)
      CALL PLOT(Y(1),F(1),3)
      DO 2 I=2,IB
      IF(F(I) .EQ. 0.) GO TO 1
      CALL PLOT(Y(I),F(I),2)
      GO TO 2
    1 CALL PLOT(Y(I),F(I),3)
    2 CONTINUE
      RETURN
  100 FORMAT(/' RMIN=',E16.6,3X,'DR=',E16.6/
     1        ' FMIN=',E16.6,3X,'DF=',E16.6)
      END
```

```
      SUBROUTINE HIST(NUM)
C*************************************************
C
C     PLOT OF HISTOGRAM
C
C*************************************************
      DIMENSION W(23),WW(22)
      COMMON Y(6003),G(6003),F(6003),R(6003),
     1       X(202),CDF(202),C1(15),D(16),
     2       ICO(7),YMIN,DY,IB,ID,N,PI,XN,
     3       SVE1,SVE2,SVE3,U1,S1,BB,LT,AA,
     4       EPS,NDIM,II,JX,BIN,EN,NP,KNOT,
     5       RHO,TAL,THETA,RXM,UM,VAR,VARS,
     6       ALPHA,SALPHA,SIG,BETA
      FN=NUM
      WX=(X(N)-X(1))/FN
      WY=WX/2.
      WX=(X(N)-X(1)+WX)/FN
      ST=X(1)-WY
      N1=NUM +1
      DO 1 I=1,N1
      A=I-1
      W(I)=ST + A*WX
    1 WW(I)=0.
      TN=N
      TN=1./TN
      JI=1
      DO 4 I=1,N
      DO 2 J=JI,NUM
      JJ=J
      IF(X(I) .GE. W(J) .AND.
     1   X(I) .LT. W(J+1)) GO TO 3
    2 CONTINUE
    3 WW(JJ)=WW(JJ) + TN
    4 JI=JJ
      RETURN
      ENTRY HIST1(NUM)
      IC=NUM + 2
      WW(IC)=SVE2
      W(IC)=X(N+1)
      W(IC+1)=X(N+2)
      IC1=IC+1
      CALL SCALE(W,IC1,8.0,WMIN,DW,1)
      CALL SCALE(WW,IC,5.0,CMIN,DC,1)
      WRITE(6,115) WMIN,DW,CMIN,DC
      CALL PLOT(W(1),0.0,3)
```

```
      CALL PLOT(W(1),WW(1),2)
      DO 5 I=2,N1
      CALL PLOT(W(I),WW(I-1),2)
    5 CALL PLOT(W(I),WW(I),2)
      RETURN
  115 FORMAT(/' WMIN=',E16.6,3X,'DW=',E16.6/
     1       ' CMIN=',E16.6,3X,'DC=',E16.6)
      END


      SUBROUTINE SAMP(NUM)
C*************************************************
C
C     COMPUTATION OF SAMPLE PERCENTILES
C
C*************************************************
      COMMON Y(6003),G(6003),F(6003),R(6003),
     1       X(202),CDF(202),C1(15),D(16),
     2       ICO(7),YMIN,DY,IB,ID,N,PI,XN,
     3       SVE1,SVE2,SVE3,U1,S1,BB,LT,AA,
     4       EPS,NDIM,II,JX,BIN,EN,NP,KNOT,
     5       RHO,TAL,THETA,RXM,UM,VAR,VARS,
     6       ALPHA,SALPHA,SIG,BETA
      N1=NUM +1
      XN1=N1
      DO 2 J=1,NUM
      XK=FLOAT(J*N)/XN1
      K=XK
      YK=K
      IF(YK .EQ. XK) GO TO 1
      X(J)=X(K+1)
      GO TO 2
    1 X(J)=X(K)
    2 CONTINUE
      WRITE(6,100) NUM,(X(I),I=1,NUM)
      N=NUM
      XN=N
      RETURN
  100 FORMAT('OSAMPLE PERCENTILES, NUM=',I3/
     1       (1X,5(E16.6,4X)/))
      END
```

```
      SUBROUTINE ECDF
C*****************************************************
C
C     PLOT OF THE ESTIMATED CDF
C
C*****************************************************
      DIMENSION AUX(10)
      COMMON Y(6003),G(6003),F(6003),R(6003),
     1       X(202),CDF(202),C1(15),D(16),
     2       ICO(7),YMIN,DY,IB,ID,N,PI,XN,
     3       SVE1,SVE2,SVE3,U1,S1,BB,LT,AA,
     4       EPS,NDIM,II,JX,BIN,EN,NP,KNOT,
     5       RHO,TAL,THETA,RXM,UM,VAR,VARS,
     6       ALPHA,SALPHA,SIG,BETA
      EXTERNAL FT
      XLL=X(N+1)
      JX=1
      JP=1
      LT=1
      TEMP=0.0
      G(IB+1)=0.0
      G(IB+2)=1.0
      DO  6 I=1,IB
      XX=F(I)
      IF(KNOT .EQ. 0) GO TO 3
      DO 1 KJ=JP,NP
      IF(XX-D(KJ)) 2,1,1
    1 JX=KJ
    2 JP=JX
    3 CALL QATR(XLL,XX,EPS,NDIM,FT ,ANS,
     1          IER,AUX)
      G(I)=TEMP + ANS
      PQZ=1.-G(I)
      IF(PQZ .LE. .005) GO TO 4
      F(I)=R(I)/PQZ
      GO TO 5
    4 F(I)=0.
    5 R(I)=PQZ
      TEMP=G(I)
    6 XLL=XX
      WRITE(6,111) G(IB),R(IB)
C     ICO(3)=0   DO NOT PLOT ESTIMATED CDF
C            =1   PLOT ESTIMATED CDF
      IF(ICO(3) .EQ. 0) RETURN
      CALL PLOT(13.0,0.0,-3)
      CALL SCALE( G ,ID,5.0,GMIN,DG,1)
```

```
      WRITE(6,114) GMIN,DG
      CALL AXIS(0.0,0.0,6HX AXIS,-6,8.0,
     1          0.0,YMIN,DY)
      CALL AXIS(0.0,0.0,8HCDF AXIS,8,5.0,
     1          90.0,GMIN,DG)
      CALL PLOT(Y(1),0.0,3)
      CALL PLOT(Y(1),G(1),2)
      CALL LINE(Y,G, IB,1)
      CALL PLOT(Y( IB),G( ID),3)
      CALL PLOT(Y( ID),G( ID),2)
      RETURN
  111 FORMAT('OG=',E16.6/  ' R=',E16.6)
  114 FORMAT(/' GMIN=',E16.6,3X,'DG=',E16.6/)
      END



      SUBROUTINE ESCDF
C*****************************************************
C
C     PLOT OF THE SAMPLE CDF
C
C*****************************************************
      COMMON Y(6003),G(6003),F(6003),R(6003),
     1       X(202),CDF(202),C1(15),D(16),
     2       ICO(7),YMIN,DY,IB,ID,N,PI,XN,
     3       SVE1,SVE2,SVE3,U1,S1,BB,LT,AA,
     4       EPS,NDIM,II,JX,BIN,EN,NP,KNOT,
     5       RHO,TAL,THETA,RXM,UM,VAR,VARS,
     6       ALPHA,SALPHA,SIG,BETA
      CDF(N+1)=0.
      CDF(N+2)=1.
      IC=N+2
      CALL PLOT(13.0,0.0,-3)
      CALL SCALE(X ,IC,8.0,XMIN,DX,1)
      CALL SCALE(CDF,IC,5.0,CMIN,DC,1)
      WRITE(6,115) XMIN,DX,CMIN,DC
      CALL AXIS(0.0,0.0,6HX AXIS,-6,8.0,
     1          0.0,XMIN,DX)
      CALL AXIS(0.0,0.0,8HCDF AXIS,8,5.0,
     1          90.0,CMIN,DC)
      CALL PLOT(X (1),0.0,3)
      CALL PLOT(X (1),CDF(1),2)
      DO 1  I=2,N
      CALL PLOT(X (I),CDF(I-1),2)
    1 CALL PLOT(X (I),CDF(I),2)
```

```
      CALL PLOT(X (IC),CDF(IC),2)
      RETURN
  115 FORMAT(/' XMIN=',E16.6,3X,'DX=',E16.6/
     1        ' CMIN=',E16.6,3X,'DC=',E16.6)
      END


      SUBROUTINE TRUE(A,B,C,O)
C*****************************************
C
C     PLOT OF THE TRUE DENSITY AND TRUE CDF
C
C*****************************************
      COMMON Y(6003),G(6003),F(6003),R(6003),
     1       X(202),CDF(202),C1(15),D(16),
     2       ICO(7),YMIN,DY,IB,ID,N,PI,XN,
     3       SVE1,SVE2,SVE3,U1,S1,BB,LT,AA,
     4       EPS,NDIM,II,JX,BIN,EN,NP,KNOT,
     5       RHO,TAL,THETA,RXM,UM,VAR,VARS,
     6       ALPHA,SALPHA,SIG,BETA
      IB=(B-A)*100. + 1.
      WRITE(6,100) IB
      DO  1 I=1,IB
      XX=FLOAT(I-1)*1.E+50/1.E+52 + A
      Y(I)=XX
      F(I)=FX(XX)
      G(I)=GX(XX)
    1 R(I)=1.-F(I)
      Y(IB+1)=AINT(C) - 1.
      Y(IB+2)=AINT(O) + 1.
      F(IB+1)=0.0
      F(IB+2)=1.0
      G(IB+1)=SVE1
      G(IB+2)=SVE2
      ID=IB+2
      CALL PLOT(13.0,0.0,-3)
      CALL SCALE(Y ,ID,8.0,YMIN,DY,1)
      CALL SCALE( F ,ID,5.0,FMIN,DF,1)
      CALL AXIS(0.0,0.0,6HX AXIS,-6,8.0,
     1          0.0,YMIN,DY)
      CALL AXIS(0.0,0.0,8HCDF AXIS,8,5.0,
     1          90.0,FMIN,DF)
      CALL PLOT(Y(1),0.0,3)
      CALL PLOT(Y(1),F(1),2)
      CALL LINE(Y,F, IB,1)
```

```
      CALL PLOT(Y(IB  ),F(ID),3)
      CALL PLOT(Y(ID  ),F(ID),2)
C     ICO(5)=0   DO NOT PLOT RELIABILITY
C            =1   PLOT RELIABILITY
      IF(ICO(5) .EQ. 0) GO TO 4
      DO 3 I=1,IB
      IF(R(I) .LE. .005) GO TO 2
      F(I)=G(I)/R(I)
      GO TO 3
    2 F(I)=0.
    3 CONTINUE
      CALL REL
    4 CALL PLOT(13.0,0.0,-3)
      CALL SCALE( G ,ID,5.0,GMIN,DG,1)
      WRITE(6,101) YMIN,DY,GMIN,DG,FMIN,DF
      CALL AXIS(0.0,0.0,6HX AXIS,-6,8.0,
     1          0.0,YMIN,DY)
      CALL AXIS(0.0,0.0,8HPDF AXIS,8,5.0,
     1          90.0,GMIN,DG)
      CALL PLOT(Y(1),0.0,3)
      CALL PLOT(Y(1),G(1),2)
      CALL LINE(Y,G, IB,1)
      CALL PLOT(Y( IB),G( IB),3)
      CALL PLOT(Y( IB),0.0,2)
      AVGSE=0.
      JX=1
      JP=1
      LT=1
      DO 5 I=1,N
      XX=X(I)
      IF(KNOT .EQ. 0) GO TO 5
      DO 6 KJ=JP,NP
      IF(XX-D(KJ)) 7,6,6
    6 JX=KJ
    7 JP=JX
    5 AVGSE=AVGSE + ( FT(XX)-GX(XX))**2
      AVGSE=AVGSE/XN
      WRITE(6,102) AVGSE
      RETURN
  100 FORMAT('0IB= ',I5)
  101 FORMAT(/' YMIN=',E16.6,3X,'DY=',E16.6/
     1       ' GMIN=',E16.6,3X,'DG=',E16.6/
     2       ' FMIN=',E16.6,3X,'DF=',E16.6)
  102 FORMAT(//' AVG SQ ERROR= ',E16.6)
      END
```

```
      FUNCTION FX(XX)
C*****************************************************
C
C     EVALUATES TRUE CDF
C
C*****************************************************
      COMMON Y(6003),G(6003),F(6003),R(6003),
     1       X(202),CDF(202),C1(15),D(16),
     2       ICO(7),YMIN,DY,IB,ID,N,PI,XN,
     3       SVE1,SVE2,SVE3,U1,S1,BB,LT,AA,
     4       EPS,NDIM,II,JX,BIN,EN,NP,KNOT,
     5       RHO,TAL,THETA,RXM,UM,VAR,VARS,
     6       ALPHA,SALPHA,SIG,BETA
C     UNIFORM       (IPARM=6)                          U4-1
      FX=(XX-ALPHA)/(BETA-ALPHA)                       U4-2
      RETURN
      END


      FUNCTION GX(XX)
C*****************************************************
C
C     EVALUATES TRUE PDF
C
C*****************************************************
      COMMON Y(6003),G(6003),F(6003),R(6003),
     1       X(202),CDF(202),C1(15),D(16),
     2       ICO(7),YMIN,DY,IB,ID,N,PI,XN,
     3       SVE1,SVE2,SVE3,U1,S1,BB,LT,AA,
     4       EPS,NDIM,II,JX,BIN,EN,NP,KNOT,
     5       RHO,TAL,THETA,RXM,UM,VAR,VARS,
     6       ALPHA,SALPHA,SIG,BETA
C     UNIFORM       (IPARM=6)                          U5-1
      GX=1./(BETA - ALPHA)                             U5-2
      RETURN
      END
```

The vita has been removed from the scanned document

ESTIMATION OF A DENSITY FUNCTION WITH APPLICATIONS TO RELIABILITY

Thomas W. Jones

(ABSTRACT)

The purpose of this dissertation is to examine the problem of estimation of a univariate probability density function. Let $Y_1, Y_2, \ldots, Y_n$ be a sample of n independent observations, each distributed according to an unknown continuous density function f(y). Given this sequence of observations, how can one estimate f(y)? Chapter I presents the historical background and a literature review of existing methods for the estimation of a probability density function. In addition, a section is devoted to the application of density estimation. In particular, we consider the practical application to reliability analysis.

The estimator of the unknown density function developed in Chapter II is similar to one proposed by Rosenblatt (1956) and by Parzen (1962). However, the kernel we consider is a function of the rank of each observation. We use, as our kernel, the asymptotic distribution of the order statistics of a sample. We refer to this estimator as the normal rank kernel estimator. In order to test the performance of our estimator, we have performed an experimental analysis by monte carlo studies.

Chapter III uses the estimator of Chapter II as the foundation for the development of a recursive estimation procedure. The technique employed is the method of successive substitution in which the solution at each iteration is used to generate the next solution until convergence is achieved. Consequently, we call this estimator the iterative estimator. Again, we have performed a simulation to compare the estimator of Chapter III with that of Chapter II.

In Chapter IV, a sequential procedure is developed for estimation of a probability density function. Initially, a normal distribution with mean and variance $\bar{y}$ and s^2, respectively, is fitted to the data and a goodness of fit test is performed. This hypothesis rejected, a sequential procedure employing the concept of spline functions is used.

Several examples are given in Chapter V which illustrate the various methods of density estimation introduced in the preceeding chapters. The examples use data that are both simulated and real. Also, an example estimates both the reliability and hazard functions.

Finally, relevant computer programs (Fortran) and descriptions of their utilization appear in the appendices. The program contained in Appendix A has a dual purpose in that by proper choice of an input parameter, the program will be executed for either the rank kernel estimator or for the iterative estimator, while Appendix B contains the computer program for Chapter IV.